Une stratégie pour la rétroconversion des documents

Pierre Héroux, Eric Trupin, Yves Lecourtier

Une stratégie pour la rétroconversion des documents

A Strategy for Retrospective Conversion of Documents

Pierre Héroux Éric Trupin Yves Lecoutier

Laboratoire Perception, Systèmes et Information

UFR des Sciences et Techniques
Université de Rouen
76821 Mont-Saint-Aignan Cedex
Pierre.Heroux@univ-rouen.fr

Résumé

Cet article propose une stratégie pour la rétroconversion des documents. Cette stratégie consiste en un cycle au cours duquel l'analyse et l'interprétation du document interagissent.
L'initialisation de ce cycle correspond à l'extraction d'une ébauche de la structure physique et de la structure logique du document. Ensuite, chaque itération du cycle consiste en une analyse de l'interprétation courante, la détection et la résolution des incohérences. Le cycle se termine lorsqu'il n'y a plus d'incohérence détectée.
La classification des objets composant le document est basée sur la fusion des résultats fournis par des classifieurs statistiques et structurels. Une méthode structurelle est utilisée pour effectuer la classification du document.
Les différents niveaux d'incohérence qui peuvent être rencontrés lors du cycle d'interprétation sont détaillés. Plusieurs méthodes sont proposées afin de remédier à ces incohérences. Enfin, nous présentons des perspectives en vue de l'implantation de cette stratégie au sein d'un système de traitement automatique de document.

Mots Clef

rétroconversion des documents, classification structurelle et statistique, cycle d'interprétation, résolution des incohérences.

Abstract

This paper proposes a strategy for retrospective conversion of documents. This strategy consists in a cycle. Document analysis and document understanding interact during this cycle.
The initialisation of this cycle corresponds to the extraction of the outline of the layout and logical structures of the document. Then, each iteration of the cycle consists in an analysis of the current description, the finding and solving of inconsistencies. The cycle ends when no more inconsistency occurs.
The classification of objects composing the document is based on the fusion of results provided by statistical and structural classifiers. A structural method is used to classify the whole document.
The different consistency levels which can occur during the understanding cycle are detailed. Several methods are proposed to solve inconsistencies. Finally, we present future prospects in order to implant this strategy in a document processing system.

Keywords

document retrospective conversion, structural and statistical classification, understanding cycle, inconsistencies solving.

1 Introduction

Cet article propose une stratégie pour la rétroconversion des documents d'un format papier vers un format électronique. La rétroconversion consiste à reconstruire une représentation électronique des documents selon un formalisme précis et ce à partir de l'image du document. L'intérêt porté à la rétroconversion se justifie par le fait qu'elle permet de bénéficier des avantages des documents électroniques tels que la diffusion facilitée et accélérée des documents à travers les réseaux, la possibilité de modifier, d'indexer et d'archiver les documents.
Il existe différentes vues du document [6]. En effet, un

document peut être décrit par son image, par sa structure physique, sa structure logique [3]... La structure physique représente l'aspect visuel du document. Elle est composée d'objets physiques (objets composites ou objets de base) organisés de façon hiérarchique. Chaque objet physique représente une entité graphique ou matérielle du document (volume, page, colonne, bloc...). La figure 1 représente l'aspect physique d'un document ainsi que la structure physique associée.

La structure logique représente l'organisation du document sur la base d'un découpage axé sur l'articulation de l'information véhiculée par le document. Cette structure, composée d'objets logiques, est elle aussi organisée de façon hiérarchique. La structure logique représente la façon dont le document peut être découpé en titre, sections, sous-sections, paragraphes... La figure 2 représente ce que pourrait être la structure logique du document de la figure 1.

La phase d'extraction de la structure physique d'un document à partir de son image est appelée analyse de document. L'interprétation de document consiste à reconstruire la structure logique.

Une classe de documents est un ensemble de documents qui partage une partie commune de leur structure physique et de leur structure logique. La partie commune de la structure est appelée structure générique.

D'autre part, les objets peuvent eux aussi être regroupés en classes selon des caractéristiques communes. Une classe d'objets est décrite par un objet générique qui synthétise l'ensemble des caractéristiques communes à tous les objets de la classe.

La stratégie que nous présentons dans cet article, peut s'appliquer à la large gamme des documents qui appartiennent à des classes de documents et dont les objets peuvent eux aussi être regroupés en classe. Cette stratégie est basée sur une interaction entre les phases d'analyse et d'interprétation du document. Elle consiste en un cycle d'interprétation inspiré par le cycle perceptif décrit par Ogier dans [5]. Le cycle est initialisé par l'extraction de structures (physique et logique) primitives. Chaque itération du cycle consiste à évaluer la pertinence de la description courante, à localiser les éventuelles incohérences et à proposer une nouvelle description dans laquelle les incohérences ont été traitées. Le cycle se termine lorsqu'il n'y a plus d'incohérences rencontrée lors de l'évaluation de la pertinence de la description.

Les éléments clés de cette stratégie sont l'extraction des ébauches de la structure physique et de la structure logique du document, la classification des objets composants ces structures, la détection des incohérences dans la description du document ainsi que la façon dont ces incohérences vont être résolues.

La section 2 traite de l'extraction des premières structures qui initialise le cycle d'interprétation. Dans la section 3, nous détaillons les méthodes employées lors de la classification des objets et des structures. La section 4 présente les différents niveaux d'incohérences que nous avons isolés et qui peuvent intervenir lors du cycle d'interprétation. Nous présentons dans la section 5 les diverses méthodes employées pour proposer de nouvelles interprétations qui viendront résoudre les incohérences. Enfin, la section 6 propose des perspectives en vue de l'implantation de cette stratégie au sein d'un système de traitement automatique des documents.

2 Initialisation du cycle d'interprétation

Le but de la rétroconversion des document est de fournir une représentation des documents qui représente au moins les structures physique et logique des documents traités. Notre stratégie est basée sur un cycle qui fait interagir les phases d'analyse et d'interprétation du document. Le cycle est initialisé par une phase qui fournit les premières versions des structures physique et logique. Ces versions seront remises en cause lors de chaque itération du cycle.

Après acquisition, l'image du document subit des traitements de bas niveau. Dans un premier temps, l'image du document est binarisée puis éventuellement redressée. Une phase de segmentation extrait les différentes entités graphiques élémentaires qui seront représentées par des objets physiques de base. Ces objets physiques de base sont ensuite regroupés pour former des objets physiques composites. Ces regroupements sont effectués selon des critères de taille et de proximité. Chaque objet est ensuite étiqueté en fonction du type de son contenu (texte, graphique ou image). Cette étiquetage est réalisé à partir de critères graphiques tels que la taille ou encore la densité de pixels noirs... De nouveaux objets composites sont alors construits en regroupant les objets de même étiquette et représentant des entités graphiques adjacentes. Finalement, une première version de la structure physique est obtenue. Une description détaillée de l'extraction de la structure physique est décrite par Diana dans [1].

Un algorithme de classification structurel compare la structure physique spécifique obtenue avec les structures physiques génériques de chaque classe de documents. Cette comparaison donne une première hypothèse concernant la classe du document. Sachant qu'une classe de document est décrite non seulement par une structure physique générique, mais également par une structure logique générique, nous construisons

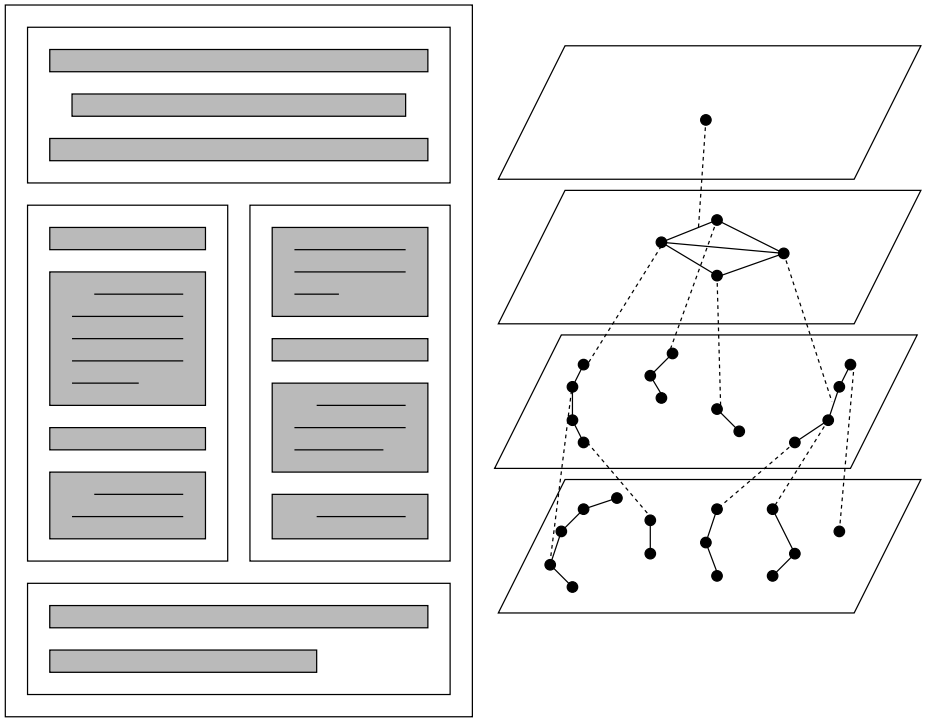

FIG. 1 – *Modélisation physique d'une page de document synthétique*

```
Document
  Titre
    Titre principal
    Sous-titre
  Première partie
    Titre
    Paragraphe
  Seconde partie
    Titre
    Paragraphe
  Troisième partie
    Titre
    Paragraphe 1
    Paragraphe 2
  Notes
    note 1
    note 2
```

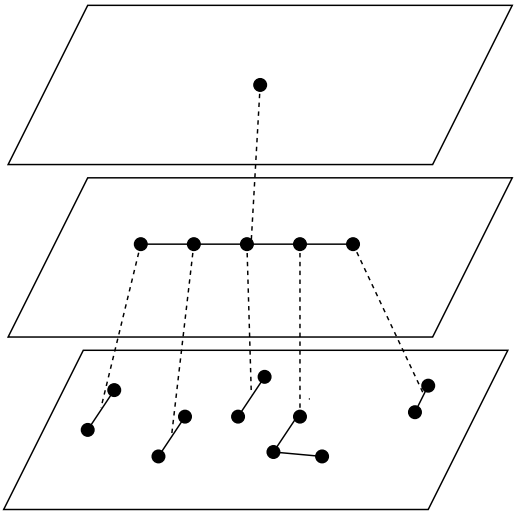

FIG. 2 – *Modélisation logique d'un document synthétique*

une première version de la structure logique en proposant une instanciation de la structure logique générique associée à l'hypothèse formulée quant à la classe du document. Cette instanciation est réalisée en associant à chaque objet physique un équivalent logique tout en observant les règles définies par la structure logique générique. Nous obtenons ainsi une première ébauche de la description du document qui initialise le cycle d'interprétation.

3 Classification

3.1 Description des objets

Dans cette section, nous décrivons les différentes méthodes utilisées dans la classification des objets, qu'ils soient physiques ou logiques, objets de base ou objets composites. Cette classification est effectuée en comparant l'objet spécifique à identifier avec l'ensemble des objets génériques représentant les classes d'objets. Chaque objet contient trois informations de nature différente :

- Un vecteur de caractéristiques numériques contient des informations propres à l'objet, indépendament de son environnement.
- Un champ contient la classe de l'objet hiérarchiquement supérieur dans la structure.
- Un graphe représente la structure de l'objet. Les nœuds du graphe représentent les classes des objets hiérarchiquement inférieurs. Les arcs du graphe représentent les relations de voisinages entre les objets fils.

Ces trois informations modélisent différents points de vue de l'objet considéré. Le vecteur de caractéristiques donne des informations intrinsèques à l'objet, indépendament du contexte dans lequel il s'insère dans le document. La classe de l'objet hiérarchiquement supéreiur situe l'objet dans la structure. La structure de l'objet décrit non seulement la liste de ses constituants, mais également leur organisation, la façon dont ces constituants sont disposés les uns par rapport aux autres.
Ces différentes informations alimentent plusieurs classifieurs. Chaque classifieur présenté par la suite renvoie une liste d'hypothèses sur la classe de chaque objet. Chaque hypothèse est valuée par une mesure caractérisant la confiance dans l'hypothèse.

3.2 Synthèse des objets génériques

Les objets génériques modélisent une classe d'objets. Ils sont générés par un apprentissage supervisé. Ils synthétisent les informations de l'ensemble des objets de la classe qu'ils représentent. De ce fait, les objets génériques sont décrits par les trois informations de nature différentes présentées précédemment. Le vecteur de caractéristiques est le vecteur moyen de l'ensemble des vecteurs de caractéristiques de la base d'apprentissage. La classe de l'objet hiérarchiquement supérieur attribuée à l'objet générique est la classe de l'objet hiérarchiquement supérieur la plus fréquemment rencontrée dans la base d'apprentissage. Cette classe est valuée par la fréquence avec laquelle elle apparaît dans la base d'apprentissage de la classe apprise. Le graphe représentant la structure de l'objet générique est généré en effectuant l'intersection de l'ensemble des graphes de la base d'apprentissage de la classe.

3.3 Classification des objets

Une méthode de classification statistique évalue la distance entre le vecteur de caractéristiques de l'objet spécifique à identifier et les vecteurs de caractéristiques de chacun des objets génériques. Divers calculs de distances peuvent être mis en œuvre (distance euclidienne, distance de Hamming, distance du maximum...). Cette distance caractérise la proximité de l'objet avec chacune des classes. La confiance exprimée pour chacune des hypothèses est inversement proportionnelle à la distance calculée.
L'information concernant l'étiquette de l'objet parent est également utilisée dans la classification de l'objet. L'objet spécifique à identifier est comparé avec chaque objet générique. La confiance attribuée à l'hypothèse de la classe avec laquelle l'objet est comparé, est valuée par la fréquence attachée à l'étiquette de l'objet parent de l'objet générique, si celle-ci est égale à l'étiquette de l'objet parent de l'objet spécifique. Si tel n'est pas le cas, la confiance attribuée à l'hypothèse est nulle.
Une méthode de classification structurelle établit une autre liste d'hypothèses quant à la classe à laquelle appartient l'objet. Cette méthode effectue la comparaison de la structure de l'objet avec celle de chacun des objets génériques. Chaque hypothèse est valuée par l'inverse d'une mesure de similarité. Cette mesure de similarité est définie en déterminant le nombre de nœuds du sous-graphe commun aux deux graphes comparés [4]. Deux taux de recouvrement t_1 et t_2 sont définis comme le nombre de nœuds du sous-graphe commun rapporté au nombre de nœuds de chacun des graphes. Si l'un de ces taux vaut 1, cela signifie qu'un des deux graphes est inclus dans l'autre. Si dans ce cas de figure, l'autre taux est faible, cela signifie que le graphe inclus ne représente qu'une très faible partie du graphe incluant. En revanche, si t_1 et t_2 valent 1, alors les graphes comparés sont égaux. La mesure de similarité d peut alors être définie par :

$$d = \frac{1}{t_1 \cdot t_2}.$$

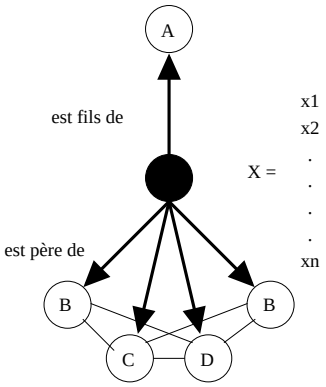

FIG. 3 – *Structure d'un objet*

La classification définitive de l'objet est établie par la fusion des résultats fournis par les trois classifieurs présentés. Plusieurs méthodes peuvent être envisagées afin d'effectuer une telle fusion. La classe définitive attribuée à l'objet peut être la classe apparaissant en meilleure position dans chacune des listes fournies par les classifieurs. Ce peut être également en multipliant les taux de confiance attribués à chaque hypothèse, les considérant alors comme des probabilités, et en choisissant l'hypothèse présentant la probabilité la plus forte. La fusion peut également être effectuée en considérant les taux de confiance attribué à chaque hypothèse non plus comme des probabilités d'appartenance, mais comme des croyances et appliquant la théorie de l'évidence...
Lors de la première itération du cycle de perception, lorsqu'aucun objet n'a encore été classifié, la seule méthode de classification statistique, ne prenant en compte que les caractéristiques intrinsèques à l'objet, est utilisée.
La coopération entre les trois classifieurs n'intervient que lors des itérations suivantes du cycle d'interprétation. La classification utilise alors non seulement les caractéristiques intrinsèques de l'objet, mais elle se réfère également au contexte hiérarchique par l'intermédiaire de l'information apportée par l'étiquette du nœud parent et le graphe des constituants de l'objet.
Finalement, lorsque tous les objets ont été classifiés, l'hypothèse retenue comme étant la classe à laquelle l'objet appartient est propagé vers l'objet dont il est un des composants et vers les objets qui le composent afin de les mettre à jour en vue de l'itération suivante du cycle d'interprétation.

3.4 Classification des structures de document

A chaque itération du cycle d'interprétation une hypothèse est émise quant à la classe d'appartenance de chaque objet du document. Les structures physiques et logiques sont alors soumises à un classifieur structurel afin de déterminer la classe du document. La classification de chacune des structures est établie en comparant la structure spécifique à identifier avec chacune des structures génériques des classes de documents connues. Les classifications des structures physique et logique proposent chacune une hypothèse quant à la classe à laquelle appartient le document. La méthode structurelle utilisée pour la classification des structures est décrites plus précisément dans [2].

4 Détection des incohérences

Après avoir proposé une hypothèse pour la classe d'appartenance de chaque objet, ainsi que des hypothèses sur la classe du document, il convient de vérifier la pertinence de la description courante du document. Cette pertinence est évaluée en parcourant les structures du document traité à la recherche d'incohérences. Nous avons isolé plusieurs niveaux de cohérence. La description d'un document sera cohérente si tous les niveaux de cohérence sont satisfaits.
Le premier niveau de cohérence est appelé cohérence intrinsèque. Un objet est dit incohérent au niveau intrinsèque si ses caractéristiques intrinsèques ne sont compatibles avec aucun des objets génériques. D'un autre côté, un objet est dit cohérent au niveau intrinsèque si les caractéristiques qui lui sont propres sont plausibles au regard des classes d'objets connues. Ce niveau de cohérence peut être établi en vérifiant que la distance entre le vecteur de caractéristiques de l'objet spécifique et le vecteur de caractéristiques de l'objet générique le plus proche est inférieur à un seuil.
Le second niveau de cohérence est appelé niveau de cohérence contextuelle de voisinage. Un objet est dit cohérent au niveau contextuel de voisinage s'il existe au moins une classe qui inclut cet objet ainsi que ces voisins dans l'organisation observée.

Le troisième niveau est le niveau de cohérence hiérarchique. Ce niveau traite de la possibilité qu'un objet appartenant à une classe puisse être ou pas un des composants d'un objet d'une autre classe. Par exemple, un objet « ligne de texte » peut être, un des composant d'un objet « bloc de texte ». En revanche, un objet « chapitre » ne peut pas être le composant d'un objet « alinéa ».

Enfin, nous définissons le niveau de cohérence d'interprétation. Ce niveau traite de la compatibilité ou de l'incompatibilité entre la classe attribuée à un objet logique et celle attribuée à l'objet physique le matérialisant. Par exemple, un objet physique « bloc de texte » peut correspondre à un objet logique « titre ». En revanche, un objet physique « graphique » ne peut pas avoir une objet logique correspondant de type « chapitre ». Une telle mise en correspondance entre deux niveaux d'interprétation consécutifs n'est pas toujours réalisable, un paragraphe peut, par exemple, être coupé en deux blocs de texte à la fin d'un colonne. Cependant, lorsque cette mise en correspondance est possible, le niveau de cohérence d'interprétation devra être testé. En particulier, les classifications des structures physique et logique devront retourner des classes de documents identiques.

5 Résolution des incohérences

Lorsque les incohérences de la description du document ont été détectées, elles doivent être résolues. Il convient d'analyser les raisons pour lesquelles ces incohérences sont survenues afin d'y remédier le plus efficacement possible. Diverses méthodes de résolution pourront alors être mises en œuvre.

Tout d'abord, l'incohérence peut venir du fait que la classification d'un objet ou d'une structure peut avoir retourné un résultat erroné. Un premier traitement à faible coût peut être d'évaluer la cohérence de la description non plus avec la première classe retournée par le classifieur, mais avec la seconde ou les suivantes.

Afin de rendre la classification plus fiable, de nouveaux éléments peuvent être pris en compte afin de proposer une nouvelle classe pour l'objet à identifier. Ceci se réalise au fur et à mesure des itérations du cycle d'interprétation. En effet, lors des premières classifications, nous ne disposons pas des connaissances sur le contexte de l'objet. Les classifications des objets reliés à l'objet à identifier soit par des liens de filiation (composition), soit par des liens de voisinages vont apporter une connaissance contextuelle qui viendra enrichir les connaissances intrinsèques sur l'objet.

Les incohérences de la description peuvent résulter d'autres erreurs. La façon dont les structures ont été construites ont pu conduire à des incohérences. Certains objets doivent par exemple être regroupés au sein d'un seul. D'autres doivent au contraire être éclatés en plusieurs. Les traitements utilisés lors de l'extraction ont pu ne pas être des plus appropriés. Dans ce cas, ils doivent être remis en cause. L'extraction de la structure peut alors être reconduite pour une partie seulement du document. Cette nouvelle extraction peut alors même être conduite en utilisant les méthodes qui semblent plus appropriées au type de données qu'on souhaite extraire et dont on peut avoir une idée *a priori* par la connaissance du reste du document. En effet, à chaque itération du cycle d'interprétation, on dispose d'une connaissance de plus en plus précise du document et donc du contexte des objets non encore identifiés avec certitude. La conclusion de cet article propose des perspectives qui pourraient évaluer de façon dynamique la pertinence d'utilisation des diverses méthodes en fonction du type de données traité.

Enfin, après avoir testé les diverses méthodes de remise en cause de la description du document, il se peut qu'il reste des incohérences non résolues. Ceci peut nous amener à remettre en cause, non plus les données sur lesquelles on travaille, ni les méthodes utilisées pour extraire ou classifier les objets du document, mais les connaissances dont on dispose. En effet, la classification des objets et celle des structures, tout comme l'examen de la cohérence de la description du document, font appel aux connaissances sur certaines classes d'objets et certaines classes de documents. Or, il se peut que le document présenté n'appartienne à aucune des classes connues. Il en est de même pour les objets. La classe d'un tel objet ou d'un tel document doit être apprise afin d'enrichir et d'élargir le domaine de compétence du système de traitement de documents.

Lorsqu'une incohérence est détectée, un des principaux problèmes est de choisir lequel de ces traitements résoudra l'incohérence de façon optimale. La plupart des incohérences rencontrées dans les représentations des documents sont dues à des erreurs de classification, à des extractions et des structurations erronées en raison de l'emploi de méthodes inappropriées ou encore à des connaissances insuffisantes sur le domaine du document traité. Les différentes méthodes ne sont pas équivalentes en terme de complexité. La nouvelle classification d'un objet représente un coût moindre qu'une nouvelle extraction, qui devra finalement être suivie d'une classification. L'apprentissage d'une nouvelle classe représente un coût encore plus élevé. Notre stratégie consiste à essayer dans un premier temps les méthodes dont la complexité de mise en œuvre est la moins forte.

Le traitement appliqué à une seule incohérence dans la description peut amener à remettre en cause une grande partie de cette description. Une méthode essayée afin de résoudre une incohérence est abandonnée lorsque le nombre total d'incohérences n'a pas décrut après un certain nombre d'itérations du cycle d'interprétation. Un traitement dont la complexité est plus forte est alors essayé à son tour.

6 Conclusion

Dans cet article, nous proposons une stratégie pour la rétroconversion des documents d'un format papier vers un format électronique. Cette méthode est basée sur une cycle d'interprétation qui consiste à effectuer la classification de l'ensemble des objets physiques et logiques ainsi que des structures physique et logique du document, à analyser la cohérence de la description fournie par l'interprétation et à proposer des solution au éventuelles incohérences. Ce cycle d'interprétation fait interagir l'analyse du document et son interprétation. D'un coté, la structure logique est initialisée à partir de connaissance sur la structure physique. D'un autre coté, la structure physique n'est pas figée et des incohérences sur la structure logique peut conduire à relancer l'acquisition d'une partie du document et à ainsi remettre en cause la structure physique.

La représentation des documents décrit trois types de relations contextuelles entre les objets (relation de voisinage, relation hiérarchique, relation physique - logique). Ces différents niveaux de relation sont exploités par les diverses méthodes de classification présentées.

La stratégie proposée sera implantée dans un système de traitement automatique de document qui devra être capable de traiter une large gamme de documents et de fournir une représentation utilisable par un grand nombre d'applications.

Les travaux futurs auront pour but d'évaluer les diverses méthodes de traitement. Ceci permettrait d'envisager le choix de la méthode la plus adaptée au regard des données à traiter. Cette évaluation pourrait, par exemple, consister à déterminer si les résultat fournis par une méthode ont conduit à des incohérences ou ont par la suite été remis en cause. Ceci pourrait également permettre de déterminer le domaine d'efficacité maximale de la méthode.

Cette caractéristique pourrait donner au système la compétence d'adaptabilité, c'est-à-dire de traiter de mieux en mieux les données qu'il a coutume de traiter.

Dans cet article, nous avons exprimé le besoin d'un apprentissage incrémental de nouvelles classes afin de palier le manque de connaissances. Cet apprentissage pourrait être effectué en exploitant les similarités et les différences entre la situation inconnue et les connaissances dont dispose le système. Ceci permettrait au système d'étendre son domaine de compétence.

Enfin, afin d'améliorer les performances du système, un opérateur pourrait valider, corriger ou compléter les différentes parties de la description du document. Son intervention pourrait alors prise en compte pour un apprentissage.

Références

[1] Sébastien Diana, Éric Trupin, Frédéric Jouzel, Jacques Labiche, and Yves Lecoutier. From acquisition to modelisation of a form base to retrieve information. In *Fourth International Conference on Document Analysis and Recognition*. IAPR, 1997.

[2] Pierre Héroux, Sébastien Diana, Éric Trupin, and Yves Lecourtier. A structural classifier to automatically identify form classes. *Advances in Pattern Recognition, Lecture Notes in Computer Science*, 1451:429–439, 1998.

[3] International Standard Organization. *ISO 8613: Information Processing. Text and Office System, Office Document Architecture (ODA) and Interchange Format*, 1989.

[4] Laurent Miclet. *Méthodes structurelles pour la reconnaissances de formes*. Eyrolles, 1984.

[5] Jean-Marc Ogier, Rémy Mullot, Jacques Labiche, and Yves Lecourtier. Interprétation de document par cycles «perceptifs» de construction d'objets cohérents. application aux données cadastrales. *Traitement du Signal*, 12(6):627–637, 1995.

[6] Yves Stern. Les quatres dimensions du document. *Document numérique*, 1(1):55–60, 1997.

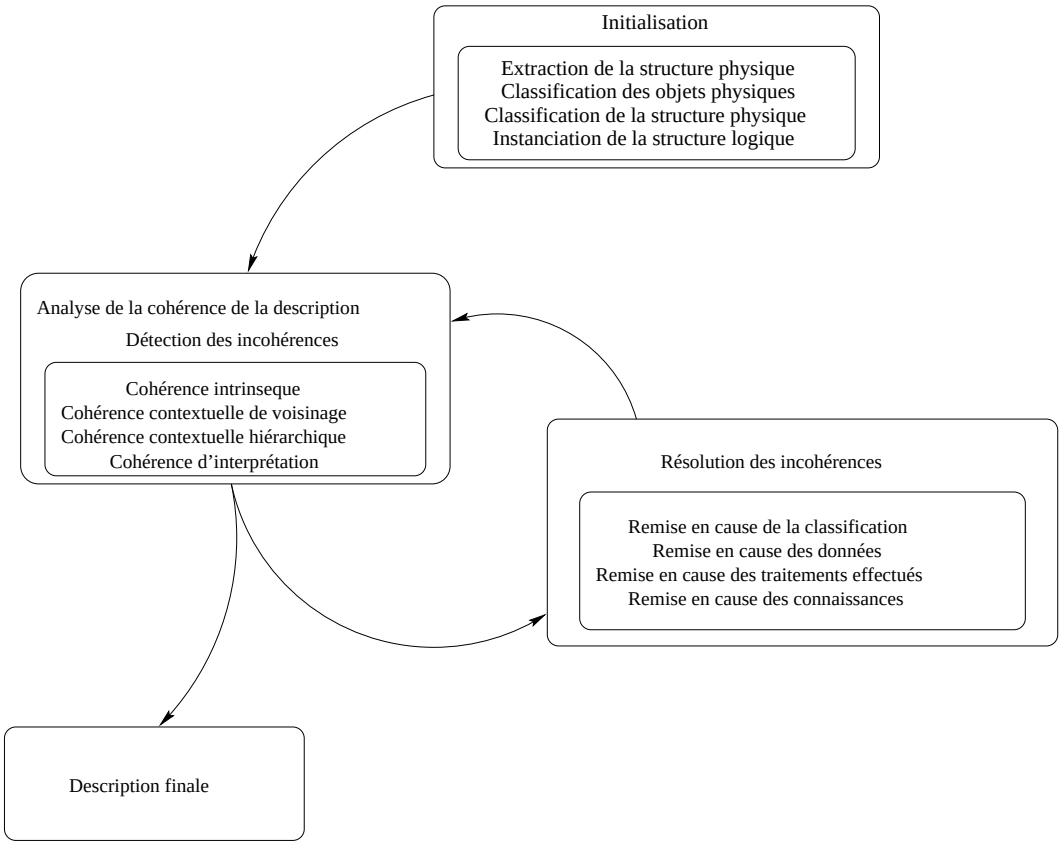

FIG. 4 – *Cycle d'interprétation du document*