Analyse formelle et relationnelle de concepts pour la modélisation et l'interrogation d'une collection documentaire

Nada Mimouni, Adeline Nazarenko, Sylvie Salotti

Analyse formelle et relationnelle de concepts pour la modélisation et l'interrogation d'une collection documentaire

Nada Mimouni*, Adeline Nazarenko*
Sylvie Salotti*

*LIPN, CNRS(UMR 7030), Université Paris Nord, 99, av. Jean-Baptiste Clément
93430 Villetaneuse
Nada.Mimouni, Adeline.Nazarenko, Sylvie.Salotti@lipn.univ-paris13.fr

Résumé. Une collection documentaire est généralement représentée comme un ensemble de documents mais cette modélisation ne permet pas de rendre compte des relations intertextuelles et du contexte d'interprétation d'un document. Le modèle documentaire classique trouve ses limites dans les domaines spécialisés où les besoins d'accès à l'information correspondent à des usages spécifiques et où les documents sont liés par de nombreux types de relations. Cet article propose un modèle permettant de rendre compte de cette complexité des collections documentaires dans les outils d'accès à l'information. En se basant sur l'analyse formelle et relationnelle de concepts appliquée sur des objets documentaires ce modèle permet de représenter et d'interroger de manière unifiée les descripteurs de contenu des documents et les relations intertextuelles qu'ils entretiennent.

1 Introduction

On représente souvent les collections de documents comme des ensembles de documents mais c'est une vue très simplifiée parce que les documents sont en réalité pris dans un ensemble de relations intertextuelles qui conditionnent leur interprétation : un document très souvent ne s'interprète pas isolément mais en référence à l'ensemble des textes qu'il cite, à partir duquel il est construit et parfois même qui en dérivent.

Si le modèle documentaire classique a fait ses preuves dans la recherche d'information généraliste qui se caractérise par le volume de documents appréhendés, la diversité des requêtes des utilisateurs et la redondance de l'information, il trouve ses limites dans les domaines spécialisés comme la médecine ou les domaines réglementaires où les outils d'accès à l'information trouvent des usages professionnels et critiques. C'est en particulier le cas dans le domaine juridique où les documents sont liés les uns aux autres par des relations d'amendements, de dérivation, de transposition, de complémentation, de jurisprudence, etc. Les outils d'accès à l'information juridique doivent tenir compte de cette complexité du matériau juridique (Bourcier, 2011).

Cet article propose un modèle permettant de représenter et d'interroger de manière unifiée les descripteurs de contenu des documents et les relations intertextuelles qu'ils entretiennent. Il repose sur l'analyse formelle et relationnelle de concepts qui est appliquée sur des objets

documentaires. Cette modélisation permet de faire apparaître deux niveaux de sémantique : un niveau sémantique correspondant à la prise en compte des liens entre les documents et un niveau sémantique qui résulte de l'annotation de nos documents par des descripteurs de contenu.

Après une revue de l'état de l'art dans la section 2, nous présentons la manière dont nous proposons de modéliser les collections documentaires sur un exemple détaillé et nous montrons dans la section 4 l'intérêt de cette modélisation pour la recherche d'information.

2 État de l'art

Le modèle classique de la Recherche d'Information (RI) représente les documents comme des sacs de mots auxquels sont assignés des poids mesurant leur importance dans le texte (poids binaire, fréquence, etc.). La recherche est ensuite faite sur cet ensemble de mots pondérés. La RI sémantique, décrite dans (Baeza Yates et R., 1999; Pejtersen, 1998), enrichit la RI classique en généralisant ou en spécifiant la requête à l'aide de ressources sémantiques (thesaurus, ontologies). Mais la RI sémantique, comme la RI classique, retourne comme résultat une liste de documents indépendants sans tenir compte du graphe de documents auquel ils appartiennent (graphe des liens entre les documents).

Quand le graphe de documents est pris en compte, c'est pour améliorer le classement des documents retournés comme dans le cas des algorithmes PageRank (Brin et Page, 1998; Page et al., 1999) et HITS (Kleinberg, 1999).

Un autre ensemble de travaux met l'accent sur l'analyse des graphes de citations (Newman, 2004; Ding, 2011) dans le but d'étudier le comportement d'une communauté en interaction (catégorisation) mais non pas dans une perspective de RI comme nous proposons de le faire dans ce travail en considérant que la prise en compte des liens enrichit sémantiquement l'interprétation et la RI dans une collection de documents.

La méthode que nous proposons repose sur l'Analyse Formelle de Concepts (AFC) et l'Analyse Relationnelle de Concepts (ARC). L'AFC est une méthode de classification conceptuelle qui, à partir d'un jeu de données représenté sous la forme d'un tableau binaire *(objets x attributs)*, construit une hiérarchie de concepts où chaque concept représente un ensemble maximal d'objets $(extension)$ ayant en commun un ensemble maximal d'attributs $(intension)$. Dans cette hiérarchie, appelée treillis de Galois ou treillis de concepts, les concepts sont (partiellement) ordonnés selon l'inclusion ensembliste entre leurs intensions et de façon duale l'inclusion inverse entre leurs extensions. La recherche d'information a été explicitement mentionnée dans (Godin et al., 1995) comme étant l'une des applications possibles des treillis de concepts. La relation de subsumption, qui est une relation d'ordre partiel entre les concepts, permet le passage d'un concept, correspondant à une requête, à un autre plus général ou plus spécifique (Godin et al., 1995).

L'utilisation de l'AFC dans la RI a fait l'objet de plusieurs travaux. Dans (Messai et al., 2006) et (Comparot et al., 2010), les auteurs proposent des techniques de raffinement et d'expansion de requête en s'appuyant sur des ontologies de domaine, ce qui permet d'améliorer le rappel par généralisation ou par spécialisation en se basant sur la structure du treillis de Galois. Sur des données textuelles, (Carpineto et Romano, 2005) propose une méthode de recherche d'information par treillis de concepts. Dans (Messai et al., 2005), les auteurs ont utilisé les treillis de concepts pour la découverte et l'interrogation de ressources génomiques sur le web.

D'autres travaux ont mis l'accent sur la classification et la structuration des résultats fournis par les algorithmes de RI ce qui influe sur les interfaces de navigation (Nauer et Toussaint, 2008; Poshyvanyk et Marcus, 2007; Carpineto et al., 2006; Koester, 2006). L'idée principale est de créer un contexte formel à partir des résultats fournis par les moteurs de recherche sur le web, de construire le treillis correspondant à ce contexte, puis de proposer à l'utilisateur un classement des résultats tel que construit par ce treillis. Ce type d'approche est implémenté dans plusieurs systèmes opérationnels tels que CREDINO (Carpineto et al., 2006), FooCA (Koester, 2006) ou CRECHAINDO (Nauer et Toussaint, 2008). Dans son travail, E. Nauer propose de classifier les résultats de recherche sur le web pour permettre à l'utilisateur de juger la pertinence des résultats qui lui sont fournis. D. Poshyvanyk *et al.* utilisent l'AFC pour classifier les résultats de la RI suite à une requête pour localiser des concepts dans un code source. Ces travaux construisent une classification conceptuelle des documents retournés mais ne tiennent pas compte des liens entre ces documents. De plus, l'exploitation de ces documents est faite *a posteriori* sur la base des résultats qui sont calculés indépendamment.

L'approche que nous présentons ici permet d'intégrer l'intertextualité *a priori* dans le modèle documentaire grâce à l'apport de l'extension relationnelle de l'AFC (ARC). Ce modèle documentaire pourra être exploité par des outils de recherche et de navigation ce qui permettra, entre autres, de répondre à des requêtes qui portent sur les relations entre documents en tant que telles.

L'ARC, proposée par (Rouane et al., 2007), est une extension relationnelle de l'AFC permettant de prendre en compte des relations entre les objets d'un même contexte, les relations entre attributs ou éventuellement la combinaison des deux. L'ARC a été utilisée dans plusieurs domaines d'application comme la classification de services web (Azmeh et al., 2011), l'extraction de patterns d'ontologie (Rouane et al., 2010), la restructuration et la construction d'ontologie, le "refactoring" de diagrammes de cas d'usage UML (Dao et al., 2004) mais jamais, à notre connaissance, pour la RI. Nous nous plaçons dans le cadre de RI dans une collection documentaire où les documents sont inter-reliés.

Nous montrons ici comment l'AFC et l'ARC permettent de représenter une collection documentaire et les perspectives d'interrogation que cela ouvre. Nous utilisons ces techniques pour formaliser un processus de RI qui exploite à la fois le contenu sémantique des documents et leurs relations intertextuelles.

Cette approche, qui n'est pas adaptée à la RI généraliste sur le web, prend tout son sens dans le cadre d'une RI spécialisée portant sur un domaine particulier. Nous nous intéressons ici au domaine juridique où les documents sont fortement liés les uns aux autres et où ces liens jouent un rôle important dans l'interprétation que l'utilisateur fait des documents retournés par un moteur de recherche.

Même dans un domaine restreint, la complexité du calcul peut être rédhibitoire. Même si, dans les applications réelles la complexité théorique maximale n'est pas atteinte (Carpineto et Romano, 2000), la complexité du treillis, mesurée en nombre de concepts et liée à la taille des contextes formels, limite l'utilisation des treillis de concepts pour la recherche d'information. L'ARC est cependant présentée ici comme modèle de représentation de la collection documentaire, sans préjuger du modèle de calcul réel à utiliser sur un corpus de grande taille où les calculs fins pourraient n'être faits que localement.

3 Modélisation d'une collection documentaire

Une collection documentaire est un ensemble de documents d'un domaine spécifique (biologique, scientifique, médical, juridique, etc.) avec des liens entre ces documents. Nous proposons ici un modèle unifié permettant de représenter à la fois le contenu de ces documents et les liens qui existent entre eux.

3.1 Exemple de collection documentaire

Notre étude se place dans le cadre du projet Legilocal [1] dont l'objectif est de faciliter l'accès des citoyens aux documents juridiques des collectivités locales. Le corpus que nous étudions est un ensemble de documents juridiques traitant du bruit. Ces documents sont de plusieurs types : arrêtés municipaux et préfectoraux, décrets, lois, codes et ordonnances. Pour simplifier la présentation du modèle, nous ne distinguons pas ici les différents types de liens entre documents (amendements, dérivation, etc.), mais cette information peut être représentée dans le modèle. Nous illustrons la modélisation sur un ensemble de quelques arrêtés dans lesquels figurent des références à des décrets et des lois. Nous montrons comment l'AFC permet de construire un premier treillis modélisant le contenu des arrêtés, qui peut être ensuite enrichi par la prise en compte, avec l'ARC, d'informations concernant les références aux décrets.

3.2 AFC pour la modélisation du contenu textuel

Dans cette section nous montrons comment l'approche AFC est appliquée pour la formalisation du contenu de notre collection documentaire. Une définition plus détaillée de l'analyse formelle de concepts est donnée dans (Ganter et Wille, 1999).

Le contenu des documents est d'abord modélisé sous la forme d'un contexte formel qui décrit une relation binaire entre un ensemble d'objets et un ensemble d'attributs *(objet x attributs)*. Les objets correspondent aux documents. Les attributs sont des descripteurs sémantiques caractérisant le contenu de ces documents. Ces descripteurs peuvent être des simples mots-clés, mais il peut s'agir de noms de concepts issus d'une ressource sémantique suite à un processus d'annotation sémantique de documents (Kiryakov et al., 2004). Cette phase d'annotation sémantique apparaît dans certains travaux qui se basent sur l'AFC pour faire la RI dans lesquels les objets sont décrits par des attributs qui font référence à des concepts dans des structures sémantiques (e.g. l'annuaire biologique BioRegistery décrit par des concepts de MeSH et NCBI (Messai et al., 2006)). Dans ce travail, nous proposons une description sémantique de nos documents en leurs associant des descripteurs sémantiques du domaine extraits du thesaurus juridique EuroVoc [2].

La formalisation du contenu des documents est donnée par le contexte formel $\mathcal{K}_{arr} = (A, S, I)$, où A est un ensemble de documents (Arrêté préfectoral Paris, Arrêté municipal Strasbourg,...), S est un ensemble de descripteurs sémantiques du domaine (ex. nuisance sonore, bruit) et I une relation binaire entre A et S appelée incidence de $\mathcal{K}_{arr}$ et vérifiant les propriétés : $I \subseteq A \times S$ et $(a, s) \in I$ ou (aIS) où a, s sont tels que $a \in A$ et $s \in S$ signifie que

1. Legilocal est un projet FUI 2010-12. Voir http ://www.mondeca.com/fr/R-D/Projets/LegiLocal-Projet-FUI-9-Cap-digital-2010-2013.
2. http ://eurovoc.europa.eu/

le document a est caractérisé sémantiquement par le descripteur s. Des exemples de contextes formels sont donnés dans la table 1 (arrêtés) et la table 2 (décrets).

	Bruit Anormalement Gênant (bag)	Nuisance Sonore (ns)	Pollution Acoustique (pa)	Sonorisation (son)	Niveau Sonore (nvs)
Arrêté Paris (AP)	x		x		
Arrêté Boulogne Billancourt (AB)	x	x		x	
Arrêté Yvelines (AY)	x	x			x
Arrêté Strasbourg (AS)			x		x

TAB. 1 – *Le contexte formel des arrêtés $\mathcal{K}_{arr}$.*

Un concept formel dans la formalisation des documents de notre collection $\mathcal{K}_{arr}$ est un ensemble de documents partageant un ensemble de descripteurs sémantiques. Un concept formel est défini comme suit.

Definition 1 (Concept formel)**.** *Soit $\mathcal{K}_{arr} = (A, S, I)$ un contexte formel. Un **concept formel** est un couple (X, Y) tel que $X \subseteq A$, $Y \subseteq S$. X et Y sont respectivement appelées $extension$ et $intension$ du concept formel (X, Y).*

Dans la théorie des treillis des relations d'ordre partiel inverse sont définies entre les extensions d'une part, et les intensions d'autre part. On parle de relations de subsumption. Notons par $\mathcal{C}$ l'ensemble des concepts formels de $\mathcal{K}_{arr}$. Soient $C_1 = (X_1, Y_1)$ et $C_2 = (X_2, Y_2)$ dans $\mathcal{C}$. C_1 est subsumé par C_2 si $X1 \subseteq X2$ où de façon duale $Y2 \subseteq Y1$ (noté par $C_1 \sqsubseteq C_2$). $(\mathcal{C}, \sqsubseteq)$ est un treillis complet appelé treillis de concepts correspondant au contexte formel $\mathcal{K}_{arr}$. On notera dans la suite $(\mathcal{C}, \sqsubseteq)$ par $\mathcal{L}(\mathcal{C})$.

La figure1 montre le treillis de concepts $\mathcal{L}(\mathcal{C})$ correspondant au contexte formel des arrêtés $\mathcal{K}_{arr}$ donné par la table 1. De la même façon, à partir du contexte formel des décrets et des lois $\mathcal{K}_{dec}$, on construit le treillis de concepts correspondant $\mathcal{L}(\mathcal{C}')$ (figure 2). Dans ces treillis, nos documents sont structurés sous forme de concepts. Un concept représente une classe de documents (l'extension) caractérisée ou décrite par un ensemble de descripteurs (l'intension).

Pour plus de clarté nous notons dans la suite ai les concepts du treillis des arrêtés et dj les concepts du treillis des décrets. Par exemple, le concept $a4$ dans le treillis des arrêtés (table1) représente l'ensemble des documents qui partagent les descripteurs $b - a - g$ (bruit anormalement gênant) et $n - s$ (nuisance sonore). Cela correspond dans notre exemple aux documents AB (arrêté de Boulogne) et AY (arrêté des Yvelines). Le lien entre les concepts $a3$ et $a4$ peut

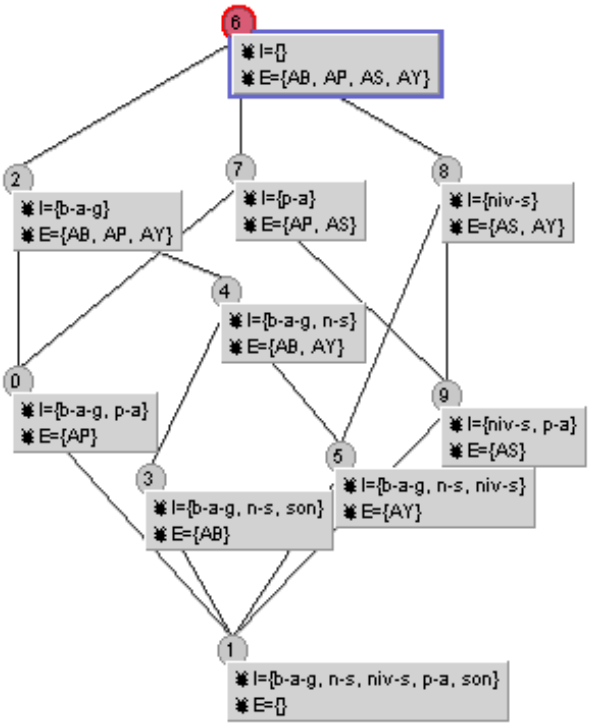

FIG. 1 – *Le treillis de concepts $\mathcal{L}(\mathcal{C})$ correspondant au contexte formel des arrêtés $\mathcal{K}_{arr}$.*

être interprété comme un lien de généralisation/spécialisation entres les classes représentées par ces concepts.

Dans une perspective de RI, le treillis construit par l'AFC regroupe toutes les combinaisons possibles des attributs des documents. Ces combinaisons sont représentées par les intensions des concepts ayant comme extension tous les documents partageant ces propriétés. Pour satisfaire les critères d'une requête en terme de pertinence, la recherche consiste à identifier la classe de documents qui partage le plus d'attributs avec la requête.

3.3 ARC pour la modélisation des liens intertextuels

L'ARC, extension relationnelle de l'AFC, permet de modéliser deux types de relations : relations entre objets et relations entre attributs (propriétés). Nous nous contentons ici de faire l'étude du premier type de relation, qui exprime les relations qui existent entre nos documents. Le deuxième type de relations sera étudié dans des travaux ultérieurs.

L'approche construit à partir de contextes binaires *(objets x attributs)* et d'une relation représentée séparément dans un deuxième contexte, une Famille de Contextes Relationnels. Cette famille constitue le point de départ du processus de formation des structures conceptuelles correspondantes appelées familles de treillis relationnels. Dans notre cas, les références

	Lutte Contre le Bruit (lcb)	Tranquilité du Voisinage (tv)	Activité Bruyante (ab)	Isolation Phonique (ip)
Décret 95 (D95)	x		x	
Code Pénal (CPen)		x		x
Ordonnance 1945 (O45)		x	x	
Loi 1992 (L92)	x			x

TAB. 2 – *Le contexte formel des décrets* $\mathcal{K}_{dec}$.

entre documents sont décrites par un contexte relationnel qui définit les relations entre les objets *(objet x objet)* [3]. Dans notre exemple, nous considérons des références que les arrêtés font aux décrets et autres textes de lois. Un exemple de ces relations de référence est représenté sur la table 3.

L'approche ARC construit un unique treillis unifiant les informations provenant des contextes formels initiaux *(objets x attributs)* et du/des contexte(s) relationnel(s) *(objet x objet)*. Sur notre exemple, le treillis final résultant après enrichissement relationnel est donnée par la figure 3.

	D95	CPen	O45	L92
AP	×			
AB				×
AY		×		
AS			×	

TAB. 3 – *Relation : fait_référence*

4 Résultats et interprétation

4.1 Modèle d'une collection documentaire

Modélisée à l'aide de l'analyse formelle et relationnelle de concepts, la collection documentaire est représentée par un ensemble de classes de documents qui sont caractérisées à la fois par des descripteurs de contenus et par les relations que les documents entretiennent les uns avec les autres. Formellement, la collection documentaire est représentée par un treillis

3. Pour modéliser différents types de liens, il faut créer différents contextes relationnels.

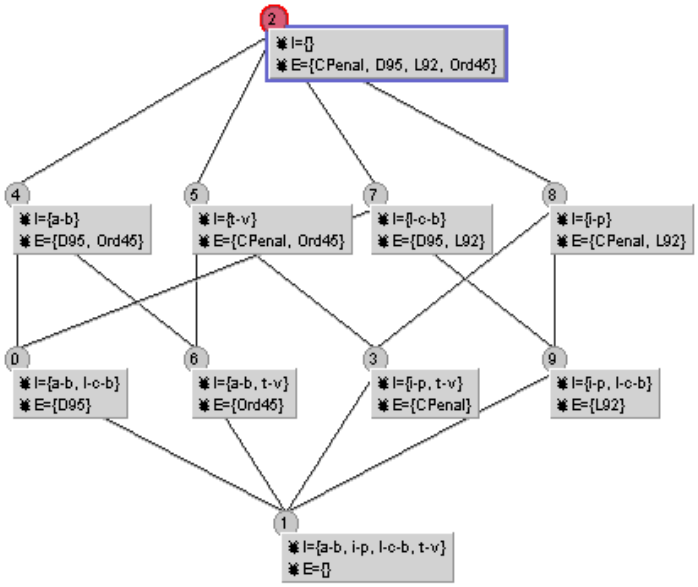

FIG. 2 – *Le treillis de concepts $\mathcal{L}(\mathcal{C}')$ correspondant au contexte formel des décrets $\mathcal{K}_{dec}$*

de concepts formels, dont les extensions sont des classes de documents et les intentions une conjonction d'attributs qui sont des descripteurs de contenu et/ou des relations vers d'autres classes de documents.

La figure 3 montre le treillis obtenu en intégrant au treillis initial de la figure 1 l'information sur les relations que les arrêtés entretiennent avec les décrets. Par souci de lisibilité du résultat et pour faciliter l'interprétation de l'exemple, nous n'avons pris en compte que des relations entre arrêtés et décrets, ce qui correspond à la structure réelle du corpus des documents juridiques que nous traitons dans le cadre de ce travail. Il faut souligner, cependant, que cette structure n'est que partielle. Il faudrait prendre en compte d'autres types de relations, entre arrêtés ou entre décrets, par exemple [4].

Si on compare le treillis de la figure 3 avec celui de la figure 1, la plupart des concepts ont une extension inchangée mais leur intension est enrichie d'attributs relationnels. C'est le cas par exemple du concept $a4$ qui a la même extension $E = \{ABoulogne, AYvelines\}$ dans les deux treillis mais dont l'intension finale combine les descripteurs de contenu de départ ($\{b-a-g, n-s\}$) avec deux descripteurs relationnels ($\{references : c2, reference : c8\}$) qui indiquent que le nouveau concept 4 est lié à deux autres concepts formels, $d2$ et $d8$ [5].

L'introduction des relations fait aussi apparaître de nouveaux concepts. Sur l'exemple jouet présenté, c'est le cas du seul concept $n^o\,10$ qui apparaît dans le treillis de la figure 3 mais qui n'était pas dans le treillis initial. Dans ce cas, l'information relationnelle a conduit

4. Nous considérons que les relations intertextuelles ne sont pas réflexives (un document n'est pas lié à lui-même).
5. Ces deux concepts correspondent à des classes de décrets dans le treillis $\mathcal{L}(\mathcal{C}')$.

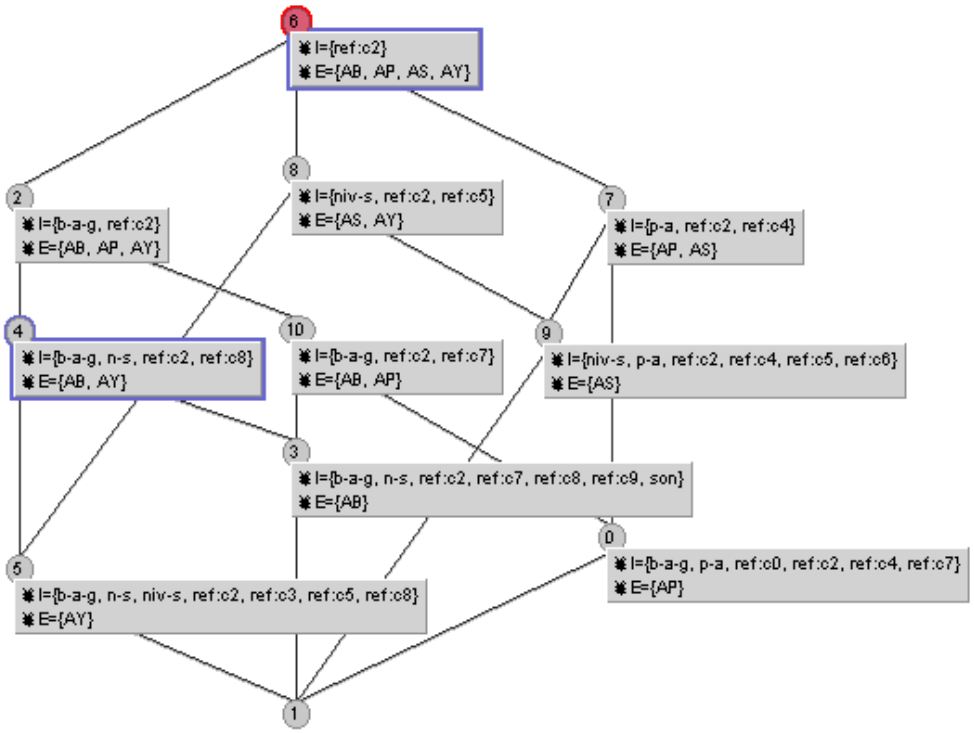

FIG. 3 – *Treillis résultant après enrichissement relationnel entre objets.*

à créer un regroupement intermédiaire ($\{ABoulogne, AParis\}$) entre ceux des concepts $a2$ ($\{ABoulogne, AParis, AYvelines\}$) et $a0$ ($\{AParis\}$) du treillis initial.

L'ajout des attributs relationnels s'interprète comme l'introduction de relations entre différentes classes de documents. Dans notre exemple, la classe $a4$ des arrêtés est ainsi reliée aux classes de décrets $d2$ et $d4$. A noter qu'il y a un processus inductif à ce stade puisque la classe $\{ABoulogne, AYvelines\}$ est reliée à la classe de décrets $\{CPenal, L92\}$ alors que les seules relations intertextuelles explicites au départ étaient entre $AYvelines$ et $CPen$ et entre $ABoulogne$ et $L92$.

4.2 Interrogation

On peut considérer que le treillis initial des arrêtés représente l'ensemble des requêtes (ou combinaisons de descripteurs) qui peuvent être faites sur la collection documentaire des arrêtés et qui sont satisfiables, c'est-à-dire qui permettent de retourner des arrêtés (toutes les combinaisons de descripteurs associées à une extension non nulle). Si la requête correspond à l'intension d'un concept qui a une extension, ce sont les documents de cette extension qui sont retournés en réponse à la requête ; si la requête correspond à une intension sans extension propre, on peut proposer des spécialisations ou au contraire généraliser la requête.

Dans cette perspective de recherche d'information, on peut apprécier l'apport de l'information relationnelle et de la modélisation que nous proposons.

Il faut d'abord souligner que tous les concepts formels initiaux étant conservés dans le treillis final, toutes les requêtes satisfiables sur le premier treillis le restent sur le treillis final.

On peut répondre à davantage de requêtes puisqu'il y a plus de concepts avec une extension propre dans le treillis : l'information relationnelle affine la catégorisation de l'ensemble des documents.

Notre modélisation permet surtout de répondre à de nouvelles formes de requêtes, les requêtes relationnelles :

- On peut retrouver un ensemble de documents associés à un autre ensemble de documents, les premiers constituant en quelque sorte le contexte d'interprétation des seconds :
  Quelles sont les classes de documents qu'un auteur donné cite ou par lequel il est cité ?
  En référence à quels documents un texte doit-il être interprété ?
  Cette forme de requête s'apparente à une requête traditionnelle couplée avec une stratégie de navigation de proche en proche à partir des liens des premiers documents retournés mais s'y ajoute ici un processus inductif qui généralise les liens entre documents individuels à des classes de documents.
- On peut interroger de manière plus globale sur la catégorie de documents qui sont associés à certains textes :
  Étant donné un ensemble d'arrêtés, à quel type de décrets font-ils référence,
  au-delà de liens explicites de référence entre arrêtés et décrets ?
  Sur quoi portent les amendements apportés à un décret particulier
  ou à un ensemble de décrets ?
- On peut finalement faire porter la requête sur les catégories sémantiques ainsi mises en relation pour découvrir par exemple que les décrets portant sur l'isolation phonique (caractérisés par le descripteur $i - p$) ont donné lieu à des arrêtés sur les nuisances sonores ($n - s$, classe $a2$ dans le treillis des arrêtés, figure 1) ou que les directives sur la pollution acoustique sont transposées dans des décrets parlant de bruit.

5 Conclusion

Nous avons présenté une modélisation qui donne une représentation unifiée des descripteurs de contenus et des relations intertextuelles qui caractérisent une collection documentaire. Nous défendons en effet l'idée que la recherche d'information spécialisée doit tenir compte des relations entre documents. C'est notamment critique pour l'accès à l'information juridique. Le modèle que nous décrivons permet de tenir compte de deux types d'informations sémantiques : les descripteurs sémantiques de contenu et les relations intertextuelles. Il permet d'obtenir des réponses plus riches à des requêtes classiques portant sur le contenu des documents, en rendant compte aussi du contexte documentaire dans lequel les documents retournés doivent être interprétés. Il permet également d'exprimer des requêtes plus riches, qui portent directement sur la structure intertextuelle de la collection documentaire.

Références

Azmeh, Z., M. Driss, F. Hamoui, M. Huchard, N. Moha, et C. Tibermacine (2011). Selection of composable web services driven by user requirements. *the Application and Experience Track of ICWS 2011*.

Baeza Yates, R. A. et N. B. R. (1999). *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman.

Bourcier, D. (2011). Sciences juridiques et complexité. un nouveau modèle d'analyse. *Droit et Cultures 61*(1), 37–53.

Brin, S. et L. Page (1998). The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst. 30*, 107–117.

Carpineto, C., A. D. Pietra, S. Mizzaro, et G. Romano (2006). Mobile clustering engine. In *ECIR*, pp. 155–166.

Carpineto, C. et G. Romano (2000). Order-theoretical ranking. *Journal of the American Society for Information Science 51*, 587–601.

Carpineto, C. et G. Romano (2005). Using concept lattices for text retrieval and mining. In *Formal Concept Analysis*, pp. 161–179.

Comparot, C., O. Haemmerlé, et N. Hernandez (2010). Expression de requêtes en graphes conceptuels à partir de mots-clés et de patrons. In *Journées Francophones d'Ingénierie des Connaissances (IC), Nîmes, 08/06/2010-11/06/2010*, http://www.cepadues.com/, pp. 81–92. Cépaduès Editions.

Dao, M., M. Huchard, M. R. Hacene, C. Roume, et P. Valtchev (2004). Improving generalization level in uml models iterative cross generalization in practice. In *ICCS'04: International Conference on Computational Science*, pp. 346–360.

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics 5*, 187–203.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis* (Mathematical Foundations ed.). Springer.

Godin, R., W. Mineau, et R. Missaoui (1995). Méthodes de classification conceptuelle basées sur les treillis de galois et applications. *Revue d'intelligence artificielle 9*, 105–137.

Kiryakov, A., B. Popov, D. Ognyanoff, D. Manov, et K. M. Goranov (2004). Semantic annotation, indexing, and retrieval. *Journal of Web Semantics 2*, 49–79.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM 46*, 604–632.

Koester, B. (2006). Conceptual knowledge retrieval with fooca: Improving web search engine results with contexts and concept hierarchies. In *Industrial Conference on Data Mining*, pp. 176–190.

Messai, N., M.-D. Devignes, A. Napoli, et M. Smaïl-Tabbone (2005). Querying a bioinformatic data sources registry with concept lattices. In *ICCS*, pp. 323–336.

Messai, N., M.-D. Devignes, A. Napoli, et M. Smaïl-Tabbone (2006). Treillis de concepts et ontologies pour interroger l'annuaire de sources de données biologiques bioregistry. *Ingénierie des Systèmes d'Information (ISI) 11*(1), 39–60.

Nauer, E. et Y. Toussaint (2008). Classification dynamique par treillis de concepts pour la recherche d'information sur le web. In *CORIA'08: Conférence en Recherche d'Information et Applications*, pp. 71–86.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America 101*, 5200–5205.

Page, L., S. Brin, R. Motwani, et T. Winograd (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Pejtersen, A. M. (1998). Semantic information retrieval. *Commun. ACM 41*, 90–92.

Poshyvanyk, D. et A. Marcus (2007). Combining formal concept analysis with information retrieval for concept location in source code. In *ICPC*, pp. 37–48.

Rouane, M. H., M. Huchard, A. Napoli, et P. Valtchev (2007). A proposal for combining formal concept analysis and description logics for mining relational data. In *Proceedings of the 5th international conference on Formal concept analysis, ICFCA 2007*, LNAI, pp. 51–65. Springer-Verlag.

Rouane, M. H., M. Huchard, A. Napoli, et P. Valtchev (2010). Using formal concept analysis for discovering knowledge patterns. In *CLA'10: 7th International Conference on Concept Lattices and Their Applications*, CEUR, pp. 223–234. University of Sevilla.

Summary

A collection of documents is generally represented as a set of documents but this simple representation does not take into account cross references between documents, which often defines their context of interpretation. This standard document model is less adapted for specific professional uses in specialized domains in which documents are related by many various references and the access tools need to consider this complexity. We propose a unified model based on formal and relational concept analysis applied on documentary objects that represents and queries in a unified way documents content descriptors and documents relations.