Classification de documents : calcul d'une distance structurelle

Karim Djemal, Chantal Soulé-Dupuy
Nathalie Vallès-Parlangeau

Université de Toulouse, IRIT, 118, route de Narbonne, F-31062 Toulouse cedex4, FRANCE
Karim.djemal@irit.fr
{Chantal.Soulé-Dupuy, Nathalie.Valles-Parlangeau}@univ-tlse1.fr

Résumé. La classification des documents numériques garantit un accès rapide et ciblé à l'information. Si nous considérons qu'un document est représenté par sa ou ses structures, définir des classes de documents revient à définir des classes de structures. Une classe structurelle représente donc des structures « proches ». Ainsi, associer la structure d'un document à sa classe structurelle revient à calculer une distance dite « structurelle ». Elle tiendra compte à la fois de l'organisation des éléments (position des nœuds, chemin), du coût d'adaptation des représentants des classes ainsi que de la représentativité des sous-graphes. Sur un corpus de documents représentant des notices de livres issus de la bibliothèque de l'université, nous discuterons de la construction de cette distance, de l'intérêt de chacun des trois paramètres utilisés.

1 Introduction

Les bases documentaires classiques retournent souvent de longues listes de documents qui doivent être parcourus afin de trouver l'information pertinente. La classification de documents offre une alternative permettant l'optimisation du processus de restitution sous forme de cluster (Soulé-Dupuy 2001).

Suivant l'objectif recherché, la classification des documents peut se faire suivant différents critères : par nature de documents (texte, video,…), par type de documents (cv, poème, lettre,…) ou encore par le contenu (thème, mots-clés,…). Nous nous intéressons ici à la classification par type de document, un type de document étant caractérisé par une structure particulière. On considère que la structure documentaire entre deux types de documents est suffisamment discriminante pour établir des classes dites « structurelles ». Les questions qui se posent à ce niveau sont : quels sont les facteurs discriminants entre des structures de documents (organisation des nœuds, types de nœuds, types de relations entre nœuds...) ? Qu'est-ce qu'une distance structurelle ? Est-ce qu'une classe doit regrouper des documents respectant une même DTD ou des documents structurellement « proches » ?

Après la présentation d'un état de l'art sur la classification de documents et la présentation du contexte de notre étude, nous abordons ces questions en proposant un calcul de distance que nous appelons « distance structurelle ». Cette distance se base sur trois critères dont nous essaierons de trouver la meilleure combinaison. Enfin, nous détaillons l'impact de ce calcul sur la classification d'un corpus de documents représentant des notices de livres issus de la bibliothèque de l'université.

2 Etat de l'art

Toute approche de classification repose sur le choix de paramètres discriminants qui permettront d'assurer la séparation entre classes. Nous pouvons distinguer, en ce qui concerne le document, deux types de paramètres : 1) paramètres structurels : le rattachement d'un document à une classe est induit par des données factuelles extraites à partir de la structure ; 2) contenu sémantique : le critère de classification porte sur la partie non structurée du document. Certes, la prise en compte du contenu dans les démarches de classification améliore les résultats en termes d'homogénéité sémantique des classes et de la pertinence des documents classés. Cependant, l'intégration du contenu requiert une phase d'indexation qui complexifie ce processus. Ainsi, plusieurs approches se limitent à des paramètres structurels pour assurer la classification de documents surtout que de nos jours les éléments de structure intègrent de la sémantique.

Si l'on s'intéresse à une classification basée sur les paramètres structurels, on peut distinguer deux approches de construction des classes. La première approche consiste à utiliser les sous-arbres fréquents (Termier et al. 2002) ; (Costa et al. 2004) ; (Kutty et al. 2008) ; (Saleem 2008). L'objectif de leurs approches est de trouver les sous-arbres qui sont inclus dans au moins n arbres d'une collection. Bien que cette méthode de classification permette de regrouper des documents partageant des sous-arborescences communes, ces documents peuvent admettre une grande différence structurelle. De plus, le coût des calculs nécessaires à ce type d'approche augmente d'une façon exponentielle en fonction de la longueur des sous-arborescences et du nombre de documents. La deuxième approche consiste à calculer la distance entre deux arborescences. Dans la littérature, deux types de distance ont été proposés : la distance d'édition et la distance d'alignement. Shasha et Zhang (1997) définissent la distance d'édition entre deux arborescences comme étant la somme des coûts des séquences d'opérations d'édition (ajout, suppression ou modification) qui transforment une arborescence en une autre. La distance d'alignement (Romany et Bonhomme 2000) est calculée entre deux arbres isomorphes. Des nœuds « espaces » étiquetés par λ, sont donc ajoutés aux deux arbres afin de marquer les nœuds manquants ou modifiés. La distance d'alignement est la somme des coûts de chaque couple possédant des étiquettes différentes. Ces coûts ne tiennent pas compte de l'importance du sous-arbre associé au nœud modifié.

3 Le processus de classification basé sur le modèle MVDM

Afin de représenter les documents multistructurés, nous avons proposé un modèle « MVDM » (Multi View Document Model) (Djemal et al. 2008) qui intègre deux niveaux : spécifique et générique. Le niveau spécifique permet de décrire les différentes caractéristiques d'une ou plusieurs structures d'un document. Chacune de ces structures est encapsulée dans ce que nous appelons « vue ». Le niveau générique permet de regrouper, sous forme de classes (niveau générique), les documents ayant des structures similaires. Chaque structure (ou vue) générique représente une collection de structures (ou vues) spécifiques.

Le processus de classification est hiérarchique ascendant non supervisé. Le premier document inséré dans la base sert de premier représentant. Les représentants sont construits par agrégation des individus structurellement proches. La comparaison des vues est basée sur :

- la pondération des relations de chacune des deux vues à comparer ;

- l'adaptation virtuelle de vues : ajoute de nœuds virtuels à chacune des deux vues de façon à avoir deux représentations virtuellement similaires en terme de nœuds ;
- le calcul de similarité entre les deux vues doit déterminer la distance entre ces deux vues. Plus précisément, cette étape consiste à calculer un degré de similarité Sim basé sur le calcul de la distance d'alignement des différents nœuds des deux vues.

Ainsi, le rattachement d'une vue spécifique à une vue générique se fait, soit directement à l'issue de la comparaison avec les différentes vues génériques, soit suite à une adaptation de la vue générique la plus proche.

3.1 Pondération des relations

Nous avons opté pour une classification basée sur des paramètres structurels. Le premier paramètre, presque évident, concerne l'organisation des relations de la vue : organisation hiérarchique des nœuds, mais aussi ordre des nœuds sur un même niveau hiérarchique. Le second paramètre prend en compte le coût d'adaptation de la vue générique à la vue spécifique. Le troisième paramètre traduit la représentativité des chemins pour l'ensemble des structures rattachées à la classe. Ces trois paramètres permettent de calculer une pondération globale pour chaque relation. Le poids final $\mathcal{P}_f$ d'une relation ε est calculé à partir des trois pondérations $\mathcal{P}_{str}$, $\mathcal{P}_{adapt}$ et $\mathcal{P}_{rep}$.

$$\mathcal{P}_f : \mathbb{E} \rightarrow]0..1[$$
$$\varepsilon \mapsto \mathcal{P}_f(\varepsilon) = \mathcal{P}_{str}(\varepsilon) * \mathcal{P}_{adapt}(\varepsilon) * \mathcal{P}_{rep}(\varepsilon)$$

3.1.1 Pondération structurelle

La pondération structurelle consiste à attribuer des poids aux relations d'une vue générique de manière à tenir compte d'une part de la profondeur (niveau du ou des nœuds pères par rapport à la racine) et d'autre part de l'ordre (ordre des nœuds fils par rapport à leur(s) nœud(s) père(s)). Soit $\mathcal{P}_{str}$ une fonction déterminant le poids par rapport à la position approximative de chaque relation (ε) dans un graphe :

$\mathcal{P}_{str} : \mathbb{E} \rightarrow]0..1[$

$$\varepsilon \mapsto \mathcal{P}_{str}(\varepsilon) = \begin{cases} \frac{\beta}{N^{\alpha}}, & si\ départ(\mathcal{E}) = \upsilon_r\ ; \\ \frac{\sum_{i=1}^{k} \mathcal{P}_{str}(\mathcal{E}_i)}{k} + \frac{\beta}{N^{\alpha}}, & sinon. \end{cases}$$

Où

- υ_r est le nœud racine du graphe avec $\upsilon_r \in \mathbb{V} / anc[\upsilon_r] = \emptyset$;
- $\mathcal{E}_i \in \mathbb{E} / \forall\, i \in [1..k]$; $arrivée(\mathcal{E}_i) = départ(\mathcal{E})$;
- α : niveau moyen du nœud υ_j dans le graphe tel que $\upsilon_j = départ(\mathcal{E})$;
- β : numéro d'ordre du nœud υ_{j+1} dans l'ensemble des fils du nœud υ_j tel que $\upsilon_j = départ\ (\mathcal{E})$ et $\upsilon_{j+1} = arrivée(\mathcal{E})$;
- N : nombre maximum de fils pour un même père défini par 10^x, ∀ x ∈ [1..n] ;
- $\mathcal{P}_{str}$ permet de gérer toute vue de document avec la contrainte que $card(fils(\upsilon)) < N$ pour tout nœud υ non feuille. Par exemple, si N est égal à 100, un nœud peut avoir jusqu'à 99 fils. Nous utiliserons une valeur de N = 10.

3.1.2 Pondération d'adaptation

Deux vues peuvent être considérées comme proches sans pour autant être identiques. Ainsi, la vue générique devra être adaptée à la vue spécifique à rattacher ; les nœuds et/ou relations manquants seront intégrés à la vue générique de départ. Plus le sous-graphe concerné par la modification est important (nombre de chemins à partir de la modification) plus le coût de modification est important. Cette remarque est cruciale pour des cas où nous sommes à la limite entre l'adaptation de la vue et la création d'une nouvelle classe. Si l'adaptation coûte trop cher, alors, mieux vaudra créer une classe plutôt que de surcharger une classe existante. Aussi, il nous semble important de pondérer la ressemblance purement structurelle par le coût d'adaptation d'une vue. La pondération d'adaptation consiste à attribuer des poids aux relations d'une vue en se basant sur l'appartenance de ces relations à des chemins.

$\mathcal{P}_{adapt} : \mathbb{E} \rightarrow]0..1[$

$$\varepsilon \mapsto \mathcal{P}_{adpt}(\varepsilon) = \begin{cases} 1/m, & si\ \mathcal{E} = \mathcal{E}_f; \\ \sum_{j=1}^{t} \mathcal{P}_{adapt}(\mathcal{E}_j), & sinon. \end{cases} \quad \text{Où}$$

- m : représente le nombre total de nœuds feuilles $v_f \in \mathbb{V} / fils[v_f] = \emptyset$;
- $\mathcal{E}_f \in \mathbb{E} / v_f = arrivée(\mathcal{E}_f)$; avec v_f (un des nœuds feuilles) $\in \mathbb{V} / fils[v_f] = \emptyset$;
- $\mathcal{E}_j \in \mathbb{E} / \forall j \in [1..t]$; $arrivée(\mathcal{E}) = départ(\mathcal{E}_j)$.

3.1.3 Pondération de représentativité

Une vue générique évolue au fur et à mesure des rattachements de vues spécifiques. Certains sous-graphes représentent de nombreuses vues spécifiques alors que d'autres ne représentent que peu de vues spécifiques. L'idée ici est de favoriser les relations génériques les plus représentées au niveau spécifique.

$\mathcal{P}_{rep}$ est une fonction qui permet de calculer la représentativité d'une relation (ε).

$\mathcal{P}_{rep} : \mathbb{E} \rightarrow]0..1]$

$$\varepsilon \mapsto \mathcal{P}_{rep}(\varepsilon) = \frac{nbrRelationsExistantes}{nbrRelationsPossibles} \quad \text{Où}$$

- $nbrRelationsExistantes$: représente le nombre de représentations génériques des relations spécifiques rattachées à relation générique $\mathcal{E}$;
- $nbrRelationsPossibles$: représente le nombre total des représentants génériques des relations spécifiques qui peuvent être rattachées à la relation générique $\mathcal{E}$. Ce nombre est égal au nombre de vues spécifiques rattachées à la vue générique traitée.

3.2 Calcul de similarité

La similarité entre deux vues est définie grâce à une distance relation à relation. Soit $alignRelation$ la fonction d'alignement permettant d'associer une relation d'un graphe à une relation d'un deuxième graphe dont les nœuds de départ et d'arrivée ont les mêmes étiquettes. La distance structurelle correspond à la distance d'alignement d'une relation d'une vue avec son image dans l'autre vue. La distance d'alignement entre ε et ε', $D_n(\varepsilon, \varepsilon')$, correspond à la valeur absolue de la différence des poids de ε et ε'. Soient $\varepsilon \in \mathbb{E}$ *et* $alignRelation(\varepsilon) = \varepsilon'$; $D_n(\varepsilon, \varepsilon') = |\mathcal{P}_f(\varepsilon) - \mathcal{P}_f(\varepsilon')|$. Le degré de similarité Sim est calculé en fonction de la distance d'alignement de toutes les relations par rapport à l'ensemble des poids des deux vues.

$$sim(\mathbb{G}, \mathbb{G}') = 1 - \frac{\sum D_n(\varepsilon, \varepsilon')}{\sum \mathcal{P}_f(\varepsilon) + \sum \mathcal{P}_f(\varepsilon')}$$

4 Validation et expérimentation

La validation est basée sur un corpus de 78 notices descriptives de livres au format XML (339 Ko). Chaque document comprend en moyenne 126 nœuds (80 éléments et 46 attributs) répartis sur six niveaux max. Chaque nœud comprend au maximum 10.

Afin de montrer l'influence des trois pondérations sur les résultats de la classification, nous avons élaboré quatre tests : StrAdaptRep (trois pondérations conjointement), StrAdapt (pondération structurelle et pondération d'adaptation), AdaptRep (pondérations d'adaptation et de représentativité) et StrRep (pondération structurelle et de représentativité). Ainsi, en fixant le seuil minimal de degré de similarité à 0,8 (80% de similarité), nous calculons le nombre de classes obtenues, le nombre moyen de documents par classe ainsi que leur dispersion traduite par l'écart type (Cf. TAB. 1). Pour chaque classe de documents générée, nous calculons également la similarité moyenne entre le représentant (vue générique) et les individus (vues spécifiques), ainsi que leur écart type.

	StrAdaptRep	StrAdapt	StrRep	AdaptRep
Nombre de classes	9	10	7	12
Moyenne du nombre de documents par classe	8,666	7,8	11,1429	6,5
Ecart type du nombre de documents par classe	7,666	7,775	9,76	6,4872
Moyenne des similarités moyennes	0,8859	0,8540	0,9188	0,76
Moyenne des écarts types	0,0283	0,036	0,0207	0,0378
Pourcentage des documents mal classés	20,5%	25%	5,1%	31,5%

TAB. 1 – Statistiques sur les quatre tests effectués.

Sur TAB. 1, les résultats montrent que la prise en compte de la structure est essentielle (meilleures résultats aux trois premières colonnes que à ceux de la quatrième colonne). De plus, sur les trois premières colonnes (StrAdaptRep, StrAdapt, StrRep), on remarque que : StrRep conduit à une meilleure classification (5,1% de documents mal classés contre 20,5% et 25% pour les deux autres cas) ; dans StrRep, les structures spécifiques rattachées sont plus proches du représentant (0,91 contre 0,85 et 0,88).

Si maintenant on considère la prise en compte des trois pondérations, les résultats ne sont pas si éloignés : un nombre de classes proche, une moyenne des similarités moyenne satisfaisante… Seulement, le pourcentage de documents mal classés est important (20,5%). Ceci nous amène à penser, que la prise en compte du coût d'adaptation peut être une bonne chose si toutefois nous perturbons régulièrement les classes. En effet, dans la démarche adoptée, la classification dépend de l'ordre de présentation des documents.

Une fois la classification terminée, nous avons recalculé les similarités entre documents et classes. Ceci a confirmé ce que nous disions plus haut : certains documents changent de classe, soit parce que leurs classes d'origines ont subi plusieurs adaptations, soit parce que lors de l'intégration du document la nouvelle classe n'était pas encore créée.

5 Conclusion

Nous avons présenté, dans ce papier, une distance entre structures dont l'originalité réside dans la combinaison de trois pondérations : une pondération structurelle permettant de traduire l'organisation d'une vue (hiérarchie et ordre) ; une pondération d'adaptation permettant

d'évaluer le coût de modification d'un nœud et une pondération de représentativité permettant de favoriser les relations les plus représentées.

A la suite des expérimentations menées sur un corpus de 78 documents ayant en moyenne 126 nœuds, cette « distance structurelle » repose tout d'abord sur l'organisation structurelle du graphe (hiérarchie, ordre des frères). Ensuite, la pondération par rapport à la représentativité d'un sous-arbre paraît être un facteur important (documents mieux classés, nombre de classes moindre, concentration autour du représentant de la classe). Cependant, il faut encore explorer le processus de construction des classes avant d'en tirer des conclusions définitives.

Nos prochains travaux concernent l'amélioration du processus de classification en se focalisant sur la représentativité des classes. La question qui se pose ici : quels sont les critères qui permettent d'identifier la dispersion d'une classe.

Références

Costa, G., Manco, G., Ortale, R., et Tagarelli, A. (2004). "A Tree-Based Approach to Clustering XML Documents by Structure." Lecture Notes In Computer Science, 137-148.

Djemal, K., Soule-Dupuy, C., et Valles-Parlangeau, N. (2008). "Formal modeling of multistructured documents." In Research Challenges in Information Science, RCIS 2008, Marrakech, 227-236.

Kutty, S., Tran, T., Nayak, R., et Li, Y. (2008). "Clustering XML Documents Using Closed Frequent Subtrees: A Structural Similarity Approach." Lecture Notes In Computer Science, 183-194.

Romany, L., et Bonhomme, P. (2000). "Parallel Alignment of Structured Documents." Parallel Text Processing, Kluwer Academic Publishers, Dordrecht, Boston, London, 201-218.

Saleem, K. (2008). "schema matching and integration in large scale snario." Université Montpellier II - Sciences et Techniques du Languedoc.

Shasha, D., et Zhang, K. (1997). "Approximate Tree Pattern Matching." Pattern Matching Algorithms, Oxford University Press, 341–371.

Soulé-Dupuy, C. (2001). "Bases d'informations textuelles : des modèles aux applications." Habilitation à diriger des recherches, Université Paul Sabatier.

Termier, A., Rousset, M. C., et Sebag, M. (2002). "TreeFinder: a First Step towards XML Data Mining." IEEE International Conference on Data Mining, DC, USA, 450.

Summary

The document clustering process ensures a fast and targeted access to information. If we consider that a document is represented by his or its structures, define document classes, it amounts to define structure classes. A structural class can represent "near" structures. Thus, associating the document structure to its structural class amounts to calculate a distance called "structural". It will consider both parameters: element organization (position of nodes, path) and frequency of appearance of each relation in the associated structures. We discuss here the interest of using this distance on a document corpus from the library of university.