Enrichissement automatique des scripts vidéo par méta données spatio-temporelles

Fendri Emna *, Ben-Abdallah Hanêne **, Ben Hamadou Abdelmajid ***

Laboratoire Miracl

**Institut Supérieur des Etudes Technologiques de Sfax, Tunisie, **Faculté des Sciences Economiques et de Gestion Sfax, Tunisie, ***Institut Supérieur d'Informatique et Multimédia, Sfax, Tunisie*

{Fendri.msf,Hanene}@gnet.tn, Abdelmajid.benhamadou@isims.rnu.tn

Résumé

Ce papier propose une approche d'enrichissement automatique des documents multimédias par des méta données. Cette approche se base sur l'annotation des documents multimédia par des relations spatio-temporelles enrichissant leur contenu et structure ; dans ce contexte, un document multimédia est composé d'un document XML aligné, via des points d'ancrage temporels, à une séquence vidéo. Pour les vidéos cinématographique, ces documents textuels (appelés scripts) décrivent la structure de la vidéo ainsi que leur mise en scène. Notre objectif est d'enrichir automatiquement les scripts formatés en XML des vidéos cinématographiques par des liens traduisant les relations spatiales et temporelles existant entre les différents éléments de l'arborescence représentative du document. Ces liens pourront être exploités pour optimiser la recherche des informations dans la vidéo et les réponses aux requêtes utilisateurs.

L'approche que nous proposons a) définit une nouvelle structure (DTD) pour les scripts XML qui intègre les éléments nouveaux de l'enrichissement du script ; et b) permet l'enrichissement automatique du script de vidéo d'une part par analyse du contenu des balises et son raffinement en vue d'extraire et d'annoter automatiquement des informations supplémentaires d'ordre spatial ou temporel ; et d'autre part par les relations existantes entre les nœuds/feuilles initiales ou nœuds/feuilles générés par la première analyse. Les relations spatio-temporelles obtenues à partir de la première étape sont exploitées pour optimiser la retrouvaille de l'information et améliorer aussi bien la qualité que le temps de réponse.

Mots clefs

Annotation, DTD, méta données, relations spatiales, relations temporelles, recherche vidéo

1 Introduction

Devant l'omniprésence des documents multimédias, les utilisateurs sont, de plus en plus, à la recherche d'outils et de techniques efficaces pour la retrouvaille des documents répondant à leurs besoins.

Dans le cadre de documents multimédia, les requêtes utilisateurs peuvent être de natures et de complexités très variées donnant lieu à des réponses de différents types : a) réponse textuelle pour répondre à une requête sur les informations de base, tel que le réalisateur, le titre, l'auteur, etc. ; b) réponse graphique (image ou portion de vidéo) pour répondre à une requête sur une action précise ou sur un évènement ; ou c) réponse audio pour le cas des requêtes sur des paroles/dialogues dans la vidéo.

Les systèmes de recherche d'information jusque là proposés offrent une recherche uni modale. Ils se basent soit sur le flux vidéo comme ceux présentés par Hauptmann dans [1], soit sur un document textuel (cf., [2]). Cependant, un système multi modal pourrait répondre d'une manière plus efficace aux besoins utilisateurs. En effet, il permettrait d'exprimer des requêtes dans un domaine sémantique plus vaste et d'y répondre d'une manière plus pertinente. Par exemple, pour le cas des vidéos cinématographiques, nous disposons généralement d'un document textuel (script) décrivant les dialogues, la succession des scènes et les lieux et moments des différentes scènes du film. Ainsi, le script combiné avec le flux vidéo peut être la base d'un système de recherche multi modal. Un tel système permettrait aux utilisateurs d'exprimer des requêtes couvrant des méta données [6] (e.g., le producteur, l'année de production, etc.), des phrases prononcées dans des conversations, des lieux et moments d'actions, etc.

Afin de permettre la mise en œuvre d'un système de recherche multi modal pour la vidéo, nous avons commencé par dégager la structure des scripts des vidéos cinématographiques. Cette structure représente, d'une part, les conventions de rédaction de ce type de documents ; d'autre part, elle structure leur contenu en segments portant des informations homogènes afin d'améliorer la qualité des réponses ramenées [3]. Ensuite, nous avons exploité cette structure et des travaux sur la segmentation du flux vidéo (cf., [4],[5]) pour aligner chaque script à sa vidéo à travers des points d'ancrage temporels. Une évaluation expérimentale de cette approche de recherche multi modale à base de scripts alignés nous a incité à explorer l'enrichissement des scripts pour l'optimisation aussi bien la qualité et la précision de la réponse que le temps de recherche.

Dans cet article, nous proposons une méthode d'enrichissement automatique du script

cinématographique par des balises décrivant des relations spatiales et temporelles inspirées, respectivement, des travaux de Clementini [8] et Allen [9].
Dans ce travail, nous nous sommes, aussi, inspirés des travaux de Ammous et al. [10] qui proposent une annotation manuelle des relations spatio-temporelles entre méta données dans les documents multimédia.
Pour le cas de scripts cinématographiques, notre méthode identifie ces relations soit à partir de la structure du script, soit par indicateurs spatiaux ou temporels. D'autre part, pour accélérer la recherche, nous proposons dans cet article une optimisation des relations identifiées et un algorithme de recherche approprié.
La suite de cet article est organisée en quatre sections. Dans la section 2, nous passons en revue la structure de scripts cinématographiques. Section 3 sera consacrée pour identifier les relations spatio-temporelles susceptibles d'être présentes dans ce type de scripts. La section 4 et 5 présentent, respectivement, une approche d'optimisation les relations et un algorithme de recherche accélérée. Finalement, cet article est clôturé par les travaux présentés et ceux en cours.

2 Structure du script cinématographique

Les vidéos cinématographiques sont généralement accompagnées de documents textuels structurés appelés scripts. Un script est composé d'éléments qui reflètent l'aspect narratif (exemple, dialogue, action,…) et productif (scène, plan,…) de la vidéo. Une étude statistique de scripts, réalisée dans le projet SRV [7], a pu dégager la structure générale de script vidéo cinématographique sous format de la DTD. Cette structure est illustrée dans Figure 1. Les éléments représentés en gras ont été identifiés par une approche d'indexation linguistique [3].
Notons que l'élément « durée » représente un point d'ancrage au flux vidéo.
Bien que cette structure représente les éléments structuraux du script, elle est incapable d'expliciter les relations implicites inter éléments (autre que l'aspect hiérarchique). Par exemple, la succession temporelle des scènes ou la contiguïté spatiale de scènes. En effet, Un système de recherche indexant les scripts selon la DTD de la figure 1 ne pourrait pas répondre à des requêtes telles que les scènes filmées avant l'apparition de l'actrice « Julia Roberts » ; la réponse à une telle requête suppose l'identification de la scène « *contenant* » l'actrice « Julia Roberts » puis de toutes les scènes « *Avant* » elle.
L'identification de telles relations peut améliorer les taux et la qualité de la réponse ainsi que le temps de recherche.

3 Identification des relations spatio-temporelles

Dans la littérature, Clementini [8] explore un nombre assez important de relations spatiales comme la disjonction, l'adjacence, l'intersection, le recouvrement, l'inclusion, l'égalité et la contiguïté. Ces relations peuvent être classées en deux types de relations temporelles : intra document ou inter documents. Allen [9] propose treize relations temporelles possibles entre deux intervalles placés sur un axe temporel.

```
<!ELEMENT script (Titre, Auteur*, Scenariste?, Producteur?, Directeur?, Ouvrage_base?, Annee?, Cast?, Introduction?, Sequence)>
<!ELEMENT Titre (#PCDATA)>
<!ELEMENT Auteur (#PCDATA)>
<!ELEMENT Scenariste (#PCDATA)>
<!ELEMENT Producteur (#PCDATA)>
<!ELEMENT Directeur (#PCDATA)>
<!ELEMENT Ouvrage_base (Oeuvre?, Ecrivain*)>
<!ELEMENT Oeuvre (#PCDATA)>
<!ELEMENT Ecrivain (#PCDATA)>
<!ELEMENT Annee (#PCDATA)>
<!ELEMENT Cast (Nomreel, Nomrole?)*>
<!ELEMENT Nomreel (#PCDATA)>
<!ELEMENT Nomrole (#PCDATA)>
<!ELEMENT Introduction (EventAction ?, ObjetAnimal ?, Objet ?, Body ?, EventNaturel ?, Lieu ?, ObjetPersonne ?, RelationTemporelle ?, RelationSpatiale ?, RelationSociale ?)>
<!ELEMENT Sequence (Scene *)>
<!ELEMENT Scene (Acetur ?, Intext ?, lieu ?, moment ?)>
<!ELEMENT Acteur (Nom?, Desc_acteur?, Dialogue?, Description?)>
<!ELEMENT Lieu (#PCDATA)>
<!ELEMENT Moment (#PCDATA)>
<!ELEMENT Duree (#PCDATA)>
<!ELEMENT Desc_dansscene ( Desc_scene?, Acteur*) >
<!ELEMENT Desc_scene (EventAction ?, ObjetAnimal ?, Objet ?, Body ?, EventNaturel ?, Lieu ?, ObjetPersonne ?, RelationTemporelle ?, RelationSpatiale ?, RelationSociale ?)>
<!ELEMENT Desc_acteur (EventAction ?, ObjetAnimal ?, Objet ?, Body ?, EventNaturel ?, Lieu ?, ObjetPersonne ?, RelationTemporelle ?, RelationSpatiale ?, RelationSociale ?)>
<!ELEMENT Dialogue (EventAction ?, ObjetAnimal ?, Objet ?, Body ?, EventNaturel ?, Lieu ?, ObjetPersonne ?, RelationTemporelle ?, RelationSpatiale ?, RelationSociale ?)>
<!ELEMENT Description (EventAction ?, ObjetAnimal ?, Objet ?, Body ?, EventNaturel ?, Lieu ?, ObjetPersonne ?, RelationTemporelle ?, RelationSpatiale ?, RelationSociale ?)>
<!ELEMENT EventAction (#PCDATA)>
<!ELEMENT ObjetAnimal (#PCDATA)>
<!ELEMENT Objet (#PCDATA)>
<!ELEMENT Body (#PCDATA)>
<!ELEMENT EventNaturel (#PCDATA)>
<!ELEMENT ObjetPersonne (#PCDATA)>
<!ELEMENT RelationTemporelle (#PCDATA)>
<!ELEMENT RelationSpatiale (#PCDATA)>
<!ELEMENT RelationSociale (#PCDATA)>
```

Figure 1 : *DTD du script vidéo cinématographique*

D'autre part, pour un document XML, nous pouvons classifier les relations entre les feuilles et/ou nouds de l'arborescence représentative de la vidéo en trois types :

a. Relations déduites de la structure : Comme montré sur Figure1, une séquence est composée de plusieurs scènes qu'on peut décomposer en segments à travers l'alignement vidéo. De ce fait, on peut décrire directement quelques relations entre « séquence » et « scène ». Par exemple, la relation « scène incluse dans séquence » est déduite automatiquement si le nœud de cette scène se trouve fils du nœud séquence.
b. Relations générées par des indicateurs spatiaux : Les indicateurs spatiaux utilisés sont exprimés dans les balises « lieu ». Les relations possibles entre deux lieux peuvent traduire une disjonction, une adjacence, l'intersection, le recouvrement, l'inclusion, l'égalité ou la contiguïté.
c. Relations générées par des indicateurs temporels : Les indicateurs temporels utilisés sont le contenu des balises « action », « Moment » et « Evt_naturel ». Les relations de base entre deux indicateurs temporels sont avant, rencontre, contient, commence, termine, égal, pendant. D'autres relations pourront être exprimées à partir des relations de base.

L'information spatio-temporelle existe soit dans les feuilles de l'arbre de la DTD, soit par des relations entre les nœuds. Cependant, il existe des nœuds/feuilles contenant des informations spatiales ou temporelles qui peuvent nuire dans la retrouvaille de données. Par exemple si le lieu d'une scène peut nous aider pour une requête utilisateur sur une action déroulée dans un lieu particulier ou sur le nom de l'acteur ayant joué dans un tel lieu ; l'information du lieu dans un dialogue n'est pas une information pertinente par rapport aux actions futures de recherche. Par exemple, si un acteur dit « le chat est dans la maison », l'information sur le lieu ici « dans la maison » n'est pas utile pour des recherches ultérieures, elle peut même fausser les résultats de recherche.

Dans notre approche, nous considérons, parmi les nœuds et feuilles de l'arborescence du script, ceux qui sont le plus porteurs d'informations supplémentaires, soit temporelles soit spatiales. Il s'agit de : scene, séquence, lieu_Scene relatif à une scène, lieu_DS relatif à une description scène, action_DS relative à une description scène, Moment_S relative à une scène, Objet_animal_DS relatif à une description scène, Objet_DS relatif à une description scène, évènement naturel relatif à une scène, objet personne relatif à une description scène et le nom d'un acteur.

Le tableau 1 résume les relations spatiales et temporelles qu'on peut dégager automatiquement entre les feuilles et nœuds du script ci-dessus énumérés.

Nous avons pu limiter les relations spatiales ou temporelles susceptibles d'exister entre deux nœuds/feuilles, pour chacun des nœuds/feuilles de l'arbre du script. Par exemple, une action peut être en relation « avant » soit avec une autre action, soit avec une séquence, une scène, un moment ou un évènement naturel. Le nombre de relations ainsi détectées s'élève à 117 relations.

Relations Spatiales			Relations temporelles		
Relation		Source	Relation		Source
A	Disjonction	b	H	Avant	c, a
B	Adjacence	b, a	I	Rencontre	c
C	Intersection	b	J	Contient	c, a
D	Recouvrement	b, a	K	Commence	c, a
E	Inclusion	b, a	L	Termine	c, a
F	Egalité	b	M	Egal	c
G	Contiguïté	b	N	Pendant	c

Tableau 1 *: Liste des relations spatiales et temporelles*

4 Vers l'optimisation du temps de recherche

Chaque relation générée à partir de l'étape précédente sera représentée dans l'arbre du script cinématographique par un lien entre feuille/feuille ou feuille/nœud. Le parcours du script, ainsi enrichi, permet de retrouver les réponses aux différentes requêtes portant sur des informations spatiales ou temporelles dans la vidéo.

Le graphe obtenu, après enrichissement par les relations spatiales et temporelles, est très complexe et peut donc engendrer un temps de réponse très élevé. D'autre part, la qualité du résultat de la requête (pertinence et précision) dépend du niveau de l'alignement du script avec la vidéo : si nous disposons de points d'ancrage au niveau « séquence », le fragment élémentaire contenant l'information recherchée serait une séquence. Dans notre cas, comme la DTD de la figure 1 l'indique, l'alignement se fait au niveau d'une scène par le biais de la balise « durée » [7].

Par conséquence, nous proposons d'optimiser le temps du parcours de l'arbre de deux manières. La première en dupliquant les informations spatio-temporelles des différents nœuds et feuilles vers les nœuds « séquences » et « scènes ». La deuxième en ajoutant des relations spatio-temporelles déduites de ces informations dupliquées. Figure 2 représente les relations spatio-temporelles potentielles dans un script cinématographique. Ces relations peuvent être détectées et annotées automatiquement.

Figure 3 représente comment ces relations sont intégrées dans l'élément séquence.

Notons que de toute manière, pour les scripts structurés selon la DTD de figure 3, les réponses ne peuvent être que des séquences (niveau un) ou des scènes (niveau deux). Ainsi, un algorithme de recherche sur le graphe obtenu peut se limiter à la profondeur deux.

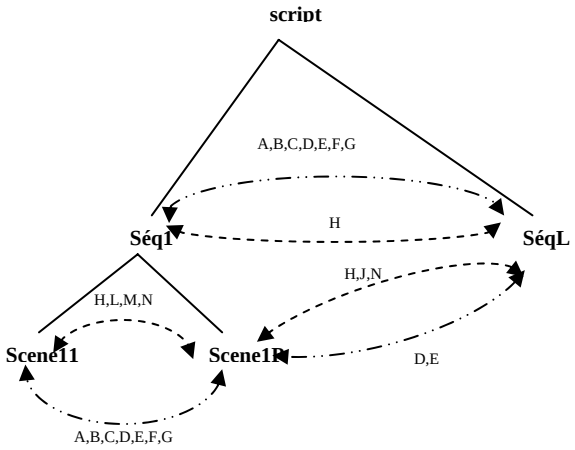

Figure 2: *Graphe représentatif des relations spatio-temporelles potentielles dans un script cinématographique*

```
<!ELEMENT Sequence (Scene *, lien_temporel_seq *, lien_temporel_scene *, lien_spatial_seq *, lien_spatial_scene *)>
<!ELEMENT Scene (Acetur ?, Intext ?, lieu ?, moment ?, lien_temporel_seq *, lien_temporel_scene *, lien_spatial_seq *, lien_spatial_scene *)>
<!ELEMENT lien_temporel_seq (relation_temporelle?, Sequence ?)>
<!ELEMENT lien_temporel_scene (relation_temporelle?, Scene ?)>
<!ELEMENT lien_spatial_seq (relation_spatiale?, Sequence ?)>
<!ELEMENT lien_spatial_scene (relation_spatiale?, Scene ?)>
<!ELEMENT relation_spatiale (#PCDATA)>
<!ELEMENT relation_temporelle (#PCDATA)>
```

Figure 3: *DTD intégrant les relations spatio-temporelles potentielles dans un script cinématographique*

5 Algorithme de recherche basé sur les relations spatio-temporelles

Dans notre approche, la recherche d'information dans un document cinématographique se base sur l'exploration de l'arbre d'un script enrichi selon la DTD de Figure 3. Le script structuré peut être représenté par un arbre de profondeur 6 contenant les données et méta données portant sur le document vidéo. Pour répondre à une requête utilisateur à travers le script, il s'agit de parcourir l'arbre à une profondeur égale à 6 pour pouvoir évaluer tous les « objets », « moments » et « lieux », susceptibles de contenir le résultat. De plus, ce parcours doit être répété en largeur pour couvrir toutes les séquences. Notons que la largeur de l'arbre dépend du nombre de séquences dans la vidéo.
Ainsi, grâce à l'enrichissement par les relations spatio-temporelles, la recherche d'information dans une vidéo cinématographique peut être optimisée en limitant le nombre de nœuds parcourus en profondeur et en largeur.

```
// Rechercher_scene est une fonction qui cherche l'information dans une liste de scenes.
// Recherche_relations_seq effectue une recherche à base de relation d'une information à partir d'une séquence
// Recherche_relations_scene effectue une recherche à base de relation d'une information à partir d'une scene

Algorithme Rechercher (Document_XML doc, string Information)

DEBUT
1. Trouvé ← faux
2. Type_rel ← Identification_relation(Information)

// recherche en profondeur Depth first search

3. TantQue Not EOF(doc) Faire
4.Seq ← doc.sequence // retourne le pointeur de la
                     //    première séquence
5. TantQue Seq <> NULL et trouve= Faux Faire
   6. Scene ← Seq.scene ;
   7. Trouvé = Rechercher_scene (scene,Information)
   8. Seq ← suivant(seq)
FinFaire

//Recherche en suivant les liens des relations

9.Si trouve Alors

        10. P_scene← tete(scene,type_rel)
        11. Tant que P_scene <> NULL Faire
                12. Recherche_relations_scene (P) ;
                13. P← P.suivant ;
        Fin faire
        14. P_seq← tete(seq,type_rel)
        15. Tant que P_seq<> NULL Faire
                16. Recherche_relations_seq(P)
                17. P← P.suivant ;
        Fin faire
FinSi

FIN
```

Figure 4: *Algorithme de recherche basé sur les relations spatio-temporelles*

Comme indiqué dans Figure 4, l'algorithme de recherche est composé de deux étapes :

1. une recherche en profondeur (depth first search) qui s'arrête au premier nœud ou feuille répondant à la requête utilisateur. (de la ligne 1 à la ligne 8)

2. Une recherche, à partir du nœud retrouvé à l'étape 1, à travers les liens spatio-temporels entre les scènes et les séquences du script. (de la ligne 9 à la ligne 17)

Par exemple, si l'utilisateur cherche « les moments filmés dans la maison ». Il suffit de trouver 1) la première scène ou séquence ayant comme lieu « dans la maison » ; puis 2) parcourir les listes chaînées traduisant les relations « recouvrement », « inclusion » ou « égalité » partant de ce nœud ou feuille initial.

La différence entre l'algorithme proposé et l'algorithme de recherche classique (en profondeur exhaustif) se résume dans deux points. Premièrement, le parcours de l'arbre représentatif du script s'arrête à la profondeur deux du nœud scène puisque toutes les informations spatio-temporelles ont été déplacées à ce niveau et au niveau séquence.

Deuxièmement, une fois le résultat est obtenu pour une première séquence ou scène, le parcours en profondeur de l'arbre est interrompu et sera suivi d'un parcours en largeur des liens spatiaux et /ou temporels dans le script. Ainsi, cet algorithme de recherche assure une optimisation au niveau temps de réponse. Les liens spatio-temporels étant des liens bidirectionnels, le résultat obtenu devrait être le même que par un parcours en profondeur exhaustif de l'arbre enrichi.

6 Conclusion

Dans ce papier, nous avons présenté une nouvelle approche d'enrichissement automatique des scripts cinématographiques par des méta données décrivant les relations spatiales et temporelles existant entre les différents composants du document cinématographique. Nous avons aussi proposé un algorithme de recherche multi modal, dans ce document, basé sur les relations spatio-temporelles intra document.

Les résultats d'une requête utilisateur étant dépendants du niveau d'alignement du script à sa vidéo (dicté par les liens d'alignement avec le flux vidéo) et étant donné qu'il est fixé dans notre cas au niveau scène, nous avons opté à ramener les relations entre les différents constituants de la vidéo à des relations entre scènes, entre séquences ou scènes/séquences.

Ainsi, l'algorithme de recherche que nous avons proposé exploite des relations à un niveau plus haut que les feuilles de l'arbre enrichi et permet d'accélérer la retrouvaille de l'information. Notre approche est en train d'être complétée pour tenir compte des relations autres que celles spatio-temporelles pour répondre à d'autres types de requêtes utilisateurs. En outre, nous sommes en train d'explorer une méthode de détection ascendante en partant d'une analyse du flux vidéo. Cette détection nous permettrait aussi de créer des points d'ancrage dans le script vers des segments vidéo plus fins que la scène.

Par ailleurs, une évaluation expérimentale est en cours pour pouvoir quantifier la pertinence et la qualité de la recherche selon l'algorithme proposé.

Références

[1] A.G. HAUPTMAN, M. G. CHRISTEL, *« Successful Approaches in the TREC video retrieval evaluation », proceedings of ACM multimedia 2004, October 10-16 2004, New York*

[2] BOUASSIDA N., BEN-ABDALLAH H., MAHDI W. *« VRS : Video Retrieval System based on script Structure Recognition », Actes des journées MediaNet 2002, Accès intelligent aux documents multimédias sur Internet, 17-21 juin 2002 , Tunisie.*

[3] FENDRI E., HAJJEM N., BEN-ABDALLAH H., *"Indexation et recherche de vidéo à travers leur script", Actes des journées CORESA'06 : Compression et Représentation des Signaux Audiovisuels, Caen 9 et 10 Novembre 2006, p. 202-207*

[4] BOUASSIDA N., BEN-ABDALLAH H., MAHDI W., Liming Chen: *Script alignment based video retrieval. EGC 2001: 149-154*

[5] FENDRI E., BEN-ABDALLAH H., BEN HAMADOU A. *« FOCSE : A framework for Soccer Video Edition »* MediaNet 2004, Tozeur, Tunisie Décembre 2004.

[6] LARHER T., MILBEAU K., *« Les méta données », Internet au quotidien : Rechercher, CNDP Octobre 1999, p37-39*

[7] MAGHREBI W., « *système de recherche vidéo »*. Mémoire de DEA SINT FSEG de Sfax Tunisie. Mars, 2003.

[8] E.CLEMENTINI, P. FELICE, D. OOSTEROM ET P.VAN., *"A Small Set of Formal Topological Relationships Suitable for End-User Interaction", dans le Proceeding de la " 3rd International Symposium on Advances in Spatial Databases ",Berlin Heidelberg New York, juin 1993, Springer Verlag, LNCS n 692, p. 277-295.*

[9] J.F. ALLEN, *"Maintaining knowledge about Temporal Intervals", Communication ACM, Vol. 26, no 11, p. 837-843, novembre 1983.*

[10] AMOUS I., JEDIDI A., SEDES F. "A Contribution to Multimedia Document Modeling and Querying". Multimedia Tools Appl. 25(3): 391-404 (2005)