Appariement d'articles en ligne et de vidéos : stratégies de sélection et méthodes d'évaluation

Adèle Désoyer[1] Delphine Battistelli[1] Jean-Luc Minel[1]
(1) MoDyCo, CNRS, Université Paris Ouest - Nanterre La Défense
adele.desoyer@gmail.com, del.battistelli@gmail.com,
jean-luc.minel@u-paris10.fr

RÉSUMÉ

Dans cet article, nous proposons une méthode d'appariement de contenus d'actualité multimédias, considérant les exigences à la fois sémantiques et temporelles du besoin d'information. La pertinence d'une vidéo pour un article de presse est mesurée par deux indices, l'un saisissant la similarité de leurs contenus, l'autre la cohérence de leurs dates d'édition. Nous présentons également une méthodologie d'évaluation s'affranchissant des standards comparant les résultats du système à des résultats de référence, en soumettant les paires de documents proposées automatiquement à un panel d'utilisateurs chargé de juger de leur pertinence.

ABSTRACT

Pairing On-line News Articles to Videos : Selection Strategies and Evaluation Methods

Here, we propose a multimedia-content pairing method taking into account both semantic and temporal parameters : relevance is computed using on one hand content similarity, on the other hand publication time proximity. We present also an evaluation method in which documents pairs are presented to users for judgment.

MOTS-CLÉS : recherche d'information, articles en ligne, corpus vidéos, évaluation.

KEYWORDS: information retrieval, on-line news articles, videos collection, evaluation.

1 Introduction

La presse a connu ces dernières décennies de nombreuses transformations, notamment depuis la fin des années 90 lorsque les organismes de presse ont commencé à proposer leurs contenus au format numérique, accessibles directement sur le web (Zouari, 2007). Depuis, les modes de consommation de l'information continuent d'évoluer et, aujourd'hui, on observe dans quantité de sites de presse une offre multimédia de l'information, notamment des vidéos que les utilisateurs semblent apprécier puisque la tendance est à l'augmentation de ce mode d'accès à l'information (Newman *et al.*, 2015).

Face à la multitude d'articles publiés chaque jour, ainsi qu'une production quasi continue de vidéos d'information, comment faire le lien entre les données de ces différentes sources ? Les travaux que nous menons relèvent de cette problématique : avec pour partenaires d'une part des sites de presse publiant quotidiennement des articles qu'ils souhaitent assortir d'une vidéo ; d'autre part des producteurs de contenus vidéos alimentant également quotidiennement une collection, notre objectif

est d'apparier [1] l'une d'elles à chacun des articles traités. La tâche est doublement contrainte, à la fois par des critères de similarité de contenu et des critères temporels, afin de garantir à l'utilisateur final une cohérence documentaire dans son parcours de page web.

Dans cet article, nous présentons une méthode d'appariement article - vidéo basée sur des modèles de recherche d'information, en rappelant brièvement les travaux relevant de problématiques comparables. Nous discutons ensuite de l'évaluation des performances du système, mise en perspective avec les méthodes classiques d'évaluation en recherche d'information, en témoignant de la difficulté d'adapter ces protocoles aux données de notre étude, et en proposant un modèle adapté à notre problématique. Nous présentons finalement les conclusions et perspectives de ces travaux.

2 Association de contenus médias

Les données qui constituent notre corpus sont des documents vidéos décrits textuellement par un titre, un résumé et éventuellement une liste de mots-clés, provenant de différents producteurs de contenus, couvrant différentes thématiques. Au total, plus d'une cinquantaine de producteurs distincts nous font parvenir des vidéos. Parmi les plus actifs, on peut notamment citer Euronews, TF1, BFM, AFP, France 24 ou Europe 1 dont les contenus relèvent essentiellement d'actualités politiques, économiques et sociétales d'étendue nationale et internationale ; Fashion TV, Zoomin ou Public pour l'actualité *people* ; ainsi que d'autres plus marginaux tels que Doctissimo, RMC Sport, Journal du Geek, ... S'ajoute à cette hétérogénéité de contenus une diversité formelle puisque les tailles des résumés varient d'une phrase unique pour certains à une dizaine de paragraphes pour d'autres. Chacune des vidéos intégrées à la collection est également décrite par un ou plusieurs thèmes indexés dans une nomenclature thématique à 24 entrées [2], élaborée par les utilisateurs du système. En moyenne, et en fonction de l'actualité médiatique, ce sont 600 nouvelles vidéos qui sont indexées chaque jour au sein d'une collection en comptant plusieurs centaines de milliers.

Les articles, données d'entrée du système développé, nous arrivent sous forme d'URLs desquelles on extrait [3] les contenus utiles à sa représentation, données textuelles telles que le titre et le corps de l'article, et métadonnées telles que la date et le site diffuseur. Ces sites sont nombreux, mais comme pour les vidéos, seuls certains envoient régulièrement de nouveaux articles à traiter : parmi eux, des sites d'information généraliste tels que 20 minutes, Europe 1, Metro ou Atlantico ; des aggrégateurs de contenus tels que Portail Free ; des sites de presse quotidienne régionale tels que La Provence, La République du Centre, Le Progrès ou Midi Libre. Quelques milliers d'articles sont quotidiennement reçus de ces différentes sources pour traitement.

Nous construisons notre système d'appariement article - vidéo en divisant la tâche globale en deux modules distincts que sont : (1) un module de recherche d'information ; (2) un module de filtre et d'ordonnancement des résultats.

1. On parle ici d'appariement plutôt que d'association, car l'objectif est bien de proposer **une** vidéo pour chaque article, et non de retrouver toutes les vidéos de la collection qui y sont liées. En des termes de recherche d'information, les performances à atteindre relèvent de la précision à un document.
2. Il s'agit de thèmes génériques classiques de la presse écrite, tels que *Politique*, *International*, *Sports* ; *Santé-Beauté* ; ...
3. Via le programme de parsing html *Goose* (https ://github.com/jiminoc/goose/wiki), spécialisé dans l'extraction d'article

2.1 Recherche d'information

L'actualité en ligne et en direct étant devenue omniprésente, on dispose avec elle d'une ressource en données dont l'étendue pose de nouveaux défis. La fin des années 90 a vu émerger de nombreux travaux autour de la problématique de *Topic Detection and Tracking* (Allan *et al.*, 1998), particulièrement autour de corpus en anglais. Elle se donne pour objectif l'identification et le suivi d'actualités dans un flux d'informations. Pour satisfaire cet objectif, l'une des tâches fondamentales est la recherche en corpus de toutes les mentions relatives à un même fait. Différents enjeux guident les études sur corpus d'actualité, parmi lesquels on peut citer : la recherche en temps réel d'actualités similaires pour la génération de résumé multi-documents, décrit dans (Radev *et al.*, 2001) ; la classification de contenus d'actualité en fonction de l'autorité des sources y référant et de leurs thématiques, proposé dans (Del Corso *et al.*, 2005) ; l'association d'articles de presse à des contenus en ligne tels que des billets de blog, ou des contenus de réseaux sociaux. Ces derniers types de travaux présentent la tâche comme une problématique de recherche d'information, dont la requête est un article et le résultat un ensemble de contenus similaires. Tandis que (Ikeda *et al.*, 2006) s'attachent à la description de leur modèle de recherche d'information vectoriel, considérant l'hétérogénéité des structures des billets de blogs et des articles dans la représentation des contenus, (Tsagkias *et al.*, 2011) proposent une comparaison de différentes modélisations de l'article pour optimiser les performances du système construit.

En s'inspirant essentiellement de ces travaux, notre algorithme de recherche d'information est développé sur le classique modèle vectoriel de (Salton *et al.*, 1975), où les documents, articles et vidéos, sont représentés par le vecteur de leur termes [4], extraits de la chaîne de traitement suivante : (1) Tokenization ; (2) Lemmatisation ; (3) Filtre sur catégories syntaxiques (on ne conserve que les noms, verbes, adjectifs, numéraux et abréviations) ; (3) Repérage des entités nommées ; (4) Repérage des termes thématiquement marqués (par comparaison à une ressource terminologique construite manuellement et régulièrement mise à jour par de nouvelles entrées, intégrant des termes représentatifs d'une thématique particulière [5]) ; (5) Exclusion des *mots-vides*.

Les termes extraits de cette chaîne de traitement sont pondérés dans le vecteur via la méthode *TF*IDF*, et les entités nommées, ainsi que les termes extraits de la ressource thématique, sont privilégiés grâce à l'attribution d'un coefficient spécifique (*i.e.* on multiplie leur *TF* par 2 voire 3 selon que le terme apparaît dans le titre ou dans le corps de texte). C'est ensuite le calcul du cosinus de l'angle formé par le vecteur de l'article avec chacun des vecteurs vidéos de l'espace qui permet d'obtenir un score de similarité pour chacune des paires article-vidéo candidates.

2.2 Ordonnancement et seuil de scores

Considérant l'ancrage de nos données dans l'actualité, la pertinence d'une vidéo en réponse à un article dépend également de paramètres temporels. Prenons l'exemple ci-dessous présentant un résultat du système ; le score de similarité cosinus des deux documents est de 0.49, mais ils ne relatent manifestement pas le même événement au regard des 369 jours séparant leurs publications respectives.

4. On considère comme requête l'intégralité du contenu des articles, (Tsagkias *et al.*, 2011) ayant conclu que cette représentation conduisait à de meilleures performances que celle considérant le titre seul.

5. Comme *ligue des champions* ou *open d'Australie* pour le sport ; *cours des comptes* ou *banque centrale européenne* pour l'économie, etc.)

Article publié le 05/02/2016 - *Athlétisme : Lavillenie pour reprendre son envol en salle à Karlsruhe*
Vidéo produite le 31/01/2015 - *Athlétisme Karlsruhe : Victoire pour Lavillenie - 31/01*

De récents travaux se sont intéressés à la prise en compte de la dimension temporelle dans l'ordonnancement de résultats de recherche d'information, en travaillant notamment sur la presse en ligne : ainsi, (Dong *et al.*, 2010) et (Dakka *et al.*, 2012) considèrent différents types de requêtes en fonction de leur sensibilité temporelle, et proposent différents scénarios de recherche d'information associés. (Dong *et al.*, 2010) décrivent un modèle de classification de l'ancrage temporel basés sur des caractéristiques d'horodatage et d'hypertexte spécifiques aux corpus web ; (Dakka *et al.*, 2012) observent la distribution dans le temps des résultats pertinents d'un article, pour en identifier les périodes importantes.

Dans notre étude, on dispose pour les articles comme pour les vidéos de la date de publication précise qui constitue une donnée temporelle fiable, et la très grande majorité d'entre eux décrivant des faits d'actualité, les articles sont tous considérés comme sensibles au critère temporel. Dans un premier temps, il s'agit donc de saisir le degré de proximité temporelle entre un article et une vidéo, puis de combiner ce score à celui de similarité thématique, pour ordonnancer les résultats de recherche d'information en fonction de ces deux critères.

S'agissant du score de proximité temporelle, noté *score_date*, nous proposons une fonction privilégiant en réponse à une requête la vidéo la plus récente lorsqu'elle existe, sans éliminer de façon drastique celles qui sont plus éloignées temporellement. L'unique variable de la fonction est la distance en nombre de jours séparant la date de publication de l'article et la date de production de la vidéo, et on observe sur sa courbe, présentée en figure 1, que les vidéos très récentes par rapport à un article (du jour-même ou de la veille) ont un *score_date* élevé (respectivement 2.2 et 1.7), et que les plus anciennes ensuite ont des scores variant de 1.7 à 0.95, valeur limite de la fonction.

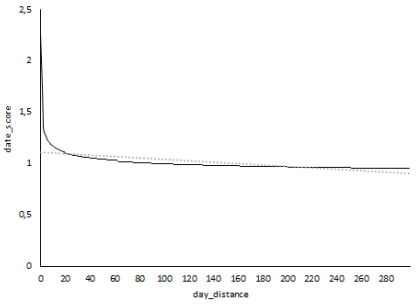

FIGURE 1 - Courbe de la fonction *score_date*

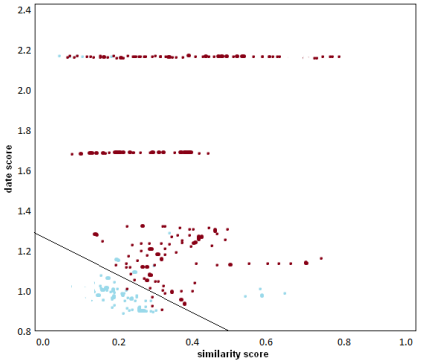

FIGURE 2 - Distribution des appariements en fonction des scores de similarité thématique et de proximité temporelle

S'agissant de la fonction d'ordonnancement des résultats, considérant le *score_date* et le *score_cosinus* d'une vidéo pour un article, nous avons observé les données d'une première série de tests du système, en distinguant deux ensembles que sont, d'une part, les appariements article/vidéo qu'un utilisateur a jugés pertinents, d'autre part, ceux qu'il a jugés non pertinents [6]. On considère alors ce problème d'ordonnancement comme relevant d'une classification binaire, dont les instances,

6. *cf* section 3 pour préciser ce qu'on l'on considère comme appariements pertinent et non pertinent

représentant des paires article/vidéo, seraient décrites par un vecteur de caractéristiques à deux dimensions (*score_date* et *score_cosinus*) et seraient à catégoriser soit dans la classe des appariements pertinents, soit dans celles des appariements non pertinents. C'est sur un ensemble restreint de données, d'à peine un millier d'appariements récupérés sur une semaine d'évaluation, et échantillonnés de manière à considérer la diversité des thématiques et des producteurs de contenus caractérisant nos données, que nous apprenons ce modèle. Étant donné le peu de variables considérées dans la description des instances d'apprentissage, on oriente notre choix de modèle vers un SVM linéaire, qui semble adapté à la séparation de deux classes de données dans un espace bidimensionnel tel que le nôtre. La figure 2 en propose une représentation, dans laquelle on distingue deux ensembles de points : Les bons appariements (correspondant aux points rouges) se concentrent majoritairement aux valeurs de *score_date* correspondant à une production vidéo datant du jour de la publication de l'article (*score_date = 2.2*) ou de la veille (*score_date = 1.7*). Parallèlement, la majorité des mauvais appariements (correspondant aux points bleus) sont concentrés dans une portion inférieure de l'espace, que l'hyperplan tracé - calculé par le SVM - permet d'isoler. Ainsi, on considère désormais qu'une vidéo est pertinente pour un article si la situation de la paire qu'ils forment dans cet espace est supérieure au séparateur, et le degré de pertinence d'un appariement est proportionnel à la distance séparant sa représentation ponctuelle de cette droite : plus il en est loin, meilleur il est.

3 Évaluation des performances du système

Le paradigme de Cranfield, décrit dans (Cleverdon, 1967), reproduit en conditions expérimentales une situation de recherche d'information, et la majorité des travaux de ce domaine évaluent les performances de leurs systèmes par cette méthode : elle nécessite de sélectionner *a priori* un ensemble fini de requêtes, auxquelles est respectivement associé un ensemble fini de résultats jugés pertinents par un évaluateur, extraits d'une collection figée de documents. L'ensemble de test ainsi constitué sert de base de comparaison aux résultats automatiques, qui sont évalués en termes de précision (taux de résultats automatiques correspondant à des résultats de référence) et rappel (taux de résultats de référence retrouvés par le système). Cette rapide description permet de constater que ce protocole d'évaluation n'est adapté ni à notre problématique (qui est de ne retourner qu'une vidéo pertinente pour un article), ni à nos données dont le dynamisme est une des caractéristiques principales : Comment sélectionner pour un article un ensemble de résultats sachant que la collection de vidéos évolue sans cesse ? De précédents travaux - (Buckley & Voorhees, 2004; Sakai & Kando, 2008) - discutaient déjà des biais de cette méthodologie, concernant en particulier la subjectivité et la non-exhaustivité des résultats de référence pour une requête. Par ailleurs, contrairement aux systèmes de recherche d'information classiques, la collection de documents n'est pas exhaustive relativement à toutes les requêtes traitées : autrement dit, tous les articles reçus n'ont pas nécessairement de vidéo associée en base. Le système développé doit considérer cette spécificité, et préférer ne proposer aucun résultat pour de tels articles, plutôt qu'un ou plusieurs résultats erronés.

Pour évaluer les performances de notre système d'appariement article/vidéo, on présente ses résultats à un ensemble de 5 utilisateurs spécialistes des contenus médiatiques. Une interface dédiée à cette tâche propose, pour chaque article, la liste des 5 meilleures vidéos retournées par le système (si toutefois elles existent), et l'utilisateur a alors plusieurs choix devant lui : (1) Sélectionner une des vidéos proposées automatiquement ; (2) Sélectionner une vidéo hors des propositions du système, s'il considère qu'aucune n'est pertinente ; (3) Ne rien sélectionner s'il estime qu'aucune vidéo de la collection n'illustre correctement l'article. Ces choix manuels forment un ensemble de référence,

auquel sont comparés les résultats automatiques, que l'on évalue comme des résultats de classification binaire, dont les deux classes sont celle des articles ayant des résultats en base (notée VIDEO), et celle des articles n'ayant aucun résultats en base (notée NO VIDEO). Les éléments de la classe VIDEO sont ensuite subdivisés en différents cas, selon le degré de corrélation entre les résultats automatiques et la sélection manuelle : on considère comme très bons les cas où l'utilisateur sélectionne la première vidéo proposée par le système (noté *First*) ; comme moins bons les cas où il sélectionne l'une des autres du top 5 (noté *Top 5*) ; comme mauvais ceux où il sélectionne une vidéo qui n'a pas été proposée automatiquement (noté *Outside*). La table 1 présente les résultats de cette évaluation sur 31966 articles traités entre février et avril 2016 : on peut alors mesurer le rappel et la précision de chacune des deux classes : la classe VIDEO obtient une précision de 0.75, et un rappel de 0.65 ; la classe NO VIDEO obtient une précision de 0.68, et un rappel de 0.78. On peut également mesurer un score s'apparentant à une mesure d'exactitude, en comparant le taux de résultats automatiques correspondant exactement à la sélection manuelle, relativement à l'ensemble des cas : on obtient ainsi un score de (6501+12610)/31966 = 0.6 pour cette métrique.

Système \ Référence		VIDEO	NO VIDEO
VIDEO	*First*	6501	3604
	Top 5	398	
	Outside	3026	
NO VIDEO		5897	12610

TABLE 1 – Matrice de confusion présentant la répartition des articles traités et évalués entre février et avril 2016

4 Conclusions et perspectives

Nous avons proposé un modèle d'appariement d'articles en ligne et de vidéos d'information basé à la fois sur des méthodes de recherche d'information considérant la similarité des contenus, et des méthodes de *ranking* considérant la proximité temporelle, cherchant ainsi à répondre à la double contrainte posée par la requête. Les spécificités de notre corpus, notamment le fait qu'il s'agisse de Français, ainsi que celles de notre problématique dont l'objectif est la précision du premier document, a rendu nos résultats difficilement comparables à ceux de l'état de l'art disposant majoritairement de corpus en anglais et d'ensembles de test attestés. Nous envisageons donc dans la poursuite de ces travaux la mise en place d'un ensemble de test représentatif du corpus, en considérant notamment dans la sélection des requêtes la variation des thématiques et des sites diffuseurs. Si cet échantillonnage est important, c'est que les performances du système dépendent fortement des articles dont il doit satisfaire le besoin d'information : tous les sites diffuseurs n'ont pas le même niveau d'attente (certains demandent un résultat pour chaque article, quitte à ce qu'il ne soit pas son exacte illustration, tandis que d'autres, insistant sur la pertinence du lien, s'y refusent absolument) ; ni la même nécessité de fraîcheur du résultat ; ni la même prédominance d'entités nommées (par exemple, la majorité des articles liés à la thématique *International* semblent contenir des entités nommées de type *lieu* qu'une sur-pondération dans le vecteur des termes permettraient de représenter plus pertinemment ; de même pour les entités *personne* des articles *People*). Les perspectives actuelles de cette étude sont donc à la mise en place de parcours de recherche personnalisés, en fonction de l'article passé en requête.

Références

ALLAN J., CARBONELL J. G., DODDINGTON G., YAMRON J. & YANG Y. (1998). Topic detection and tracking pilot study final report.

BUCKLEY C. & VOORHEES E. M. (2004). Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 25–32 : ACM.

CLEVERDON C. (1967). The cranfield tests on index language devices. In *Aslib proceedings*, volume 19, p. 173–194 : MCB UP Ltd.

DAKKA W., GRAVANO L. & IPEIROTIS P. G. (2012). Answering general time-sensitive queries. *Knowledge and Data Engineering, IEEE Transactions on*, **24**(2), 220–235.

DEL CORSO G. M., GULLÍ A. & ROMANI F. (2005). Ranking a stream of news. In *Proceedings of the 14th international conference on World Wide Web*, p. 97–106 : ACM.

DONG A., CHANG Y., ZHENG Z., MISHNE G., BAI J., ZHANG R., BUCHNER K., LIAO C. & DIAZ F. (2010). Towards recency ranking in web search. In *Proceedings of the third ACM international conference on Web search and data mining*, p. 11–20 : ACM.

IKEDA D., FUJIKI T. & OKUMURA M. (2006). Automatically linking news articles to blog entries. In *AAAI Spring Symposium : Computational Approaches to Analyzing Weblogs*, p. 78–82.

NEWMAN N., LEVY D. A. & NIELSEN R. K. (2015). Reuters institute digital news report 2015. *Available at SSRN 2619576*.

RADEV D. R., BLAIR-GOLDENSOHN S., ZHANG Z. & RAGHAVAN R. S. (2001). Interactive, domain-independent identification and summarization of topically related news articles. In *Research and Advanced Technology for Digital Libraries*, p. 225–238. Springer.

SAKAI T. & KANDO N. (2008). On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, **11**(5), 447–470.

SALTON G., WONG A. & YANG C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, **18**(11), 613–620.

TSAGKIAS M., DE RIJKE M. & WEERKAMP W. (2011). Linking online news and social media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, p. 565–574 : ACM.

ZOUARI K. (2007). La presse en ligne : vers un nouveau média ? *Les Enjeux de l'information et de la communication*, **2007**(1), 81–92.