Vers une typologie de liens entre contenus journalistiques

Rémi Bois, Guillaume Gravier, Pascale Sébillot, Emmanuel Morin

Vers une typologie de liens entre contenus journalistiques

Rémi BOIS[1] Guillaume GRAVIER[1] Pascale SÉBILLOT[2] Emmanuel MORIN[3]
(1) CNRS, IRISA & INRIA, Campus de Beaulieu, 35000 Rennes
(2) INSA, IRISA & INRIA, Campus de Beaulieu, 35000 Rennes
(3) Université de Nantes, LINA, 2 Rue de la Houssinière, 44300 Nantes
(1, 2) prenom.nom@irisa.fr (3) emmanuel.morin@univ-nantes.fr

Résumé. Nous présentons une typologie de liens pour un corpus multimédia ancré dans le domaine journalistique. Bien que plusieurs typologies aient été créées et utilisées par la communauté, aucune ne permet de répondre aux enjeux de taille et de variété soulevés par l'utilisation d'un corpus large comprenant des textes, des vidéos, ou des émissions radiophoniques. Nous proposons donc une nouvelle typologie, première étape visant à la création et la catégorisation automatique de liens entre des fragments de documents afin de proposer de nouveaux modes de navigation au sein d'un grand corpus. Plusieurs exemples d'instanciation de la typologie sont présentés afin d'illustrer son intérêt.

Abstract.
Towards a typology for linking newswire contents
In this paper, we introduce a typology of possible links between contents of a multimedia news corpus. While several typologies have been proposed and used by the community, we argue that they are not adapted to rich and large corpora which can contain texts, videos, or radio stations recordings. We propose a new typology, as a first step towards automatically creating and categorizing links between documents' fragments in order to create new ways to navigate, explore, and extract knowledge from large collections. Several examples of links in a large corpus are given.

Mots-clés : typologie, liens inter-documents, hypertexte, multimédia, presse.

Keywords: typology, linking documents, hypertext, multimedia, newswire.

1 Introduction

L'explosion de la quantité de sources d'informations disponible sur le web a rendu nécessaire l'utilisation d'outils permettant de guider la navigation des internautes, notamment dans le cas pratique de l'extraction d'information liée au domaine journalistique. Or, les outils existants aujourd'hui peinent à proposer à leurs utilisateurs une navigation qui soit à la fois intuitive et diversifiée. Une personne cherchant à étendre ses connaissances sur un évènement lié à l'actualité sera le plus souvent amenée à utiliser un moteur de recherche, outil certes performant mais laissant à son utilisateur la charge cognitive de faire le lien entre les différentes pages qu'il aura consultées. Certains professionnels, comme les attachés de presse, ont également besoin de compiler rapidement toute une série d'informations autour d'un même sujet. Une fois encore, les moteurs de recherche sont la seule solution à leur disposition, faute d'outils plus performants.

Le projet LIMAH [1], dans lequel cette étude se déroule, vise à répondre à cette problématique en créant de façon automatisée des liens explicites, fondés sur une similarité sémantique, entre des fragments de documents issus du domaine journalistique. Un grand corpus multimédia, contenant un mois de données journalistiques récupérées sur le web sous forme de vidéos (*e.g.* journal télévisé), de podcasts (*e.g.* chroniques radio), ou de pages HTML (*e.g.* lemonde.fr) a été constitué pour le projet. L'objectif est de permettre une navigation éclairée au sein de ce corpus, dans lequel une personne peut par exemple choisir de suivre l'évolution d'une actualité, ou au contraire d'organiser son parcours de l'information autour de résumés afin d'acquérir une connaissance rapide des principaux enjeux la concernant. Organisée sous forme d'hypergraphe, cette collection de fragments liés doit permettre de proposer une telle navigation, lors de laquelle l'utilisateur dispose de plusieurs choix, décrits de façon explicite, pour orienter son parcours.

1. Linking Media in Acceptable Hypergraphs `http://limah.irisa.fr/`

La première étape à considérer pour la création de tels hypergraphes est la caractérisation des liens à construire entre les fragments de documents. Si plusieurs typologies de liens ont déjà été proposées (*cf.* section 3.1), elles ne nous semblent pas suffisamment riches pour représenter la diversité des relations existantes. Dans cet article, nous proposons donc une typologie des liens dans un corpus multimédia ancré dans le domaine journalistique avec comme objectif la création et la catégorisation automatiques de ces liens.

Dans un premier temps, nous décrivons le type de corpus sur lequel se fonde cette étude (*cf.* section 2). Puis, nous exposons la typologie proposée et la façon dont elle a été construite (*cf.* section 3). Nous montrons ensuite que cette typologie est adaptée au corpus étudié au travers d'exemples issus de ces données (*cf.* section 3.3). Nous concluons ce travail en exposant les possibilités offertes par cette typologie (*cf.* section 4).

2 Corpus

La recherche d'informations liées à l'actualité sur le web met en œuvre de très nombreux types de documents. La plupart des journaux papiers sont en effet disponibles dans une version électronique, le plus souvent accessible gratuitement, et parfois enrichie de contenus multimédias (*e.g.* vidéos explicatives, illustrations, graphiques). Les stations radio et chaînes de télévision mettent également à disposition des internautes leurs émissions, sous forme de podcasts pour les premiers et de vidéos à la demande pour les secondes. Les utilisateurs se tournent également de plus en plus vers une information communautaire, où les réactions à un évènement deviennent aussi importantes que la description de l'évènement lui-même. Les titres des articles de la presse écrite sont tweetés, leurs en-têtes publiés sur Facebook, et les internautes interagissent directement pour commenter non seulement l'information, mais également la façon dont celle-ci est transmise. Le réseau social Twitter est notamment largement utilisé pour s'informer, comme le montrent plusieurs études (Kwak *et al.*, 2010).

Pour appréhender ce domaine, disposer d'un corpus représentant cette masse et cette diversité semble nécessaire. Nous nous fondons donc sur un mois de données journalistiques, mélangeant des ressources issues des chaînes de télévision (France Télévision) ou des stations de radio (Radio France), des articles de blogs ou de presse écrite (Le Monde, Le Figaro), ainsi que les commentaires qui leur sont associés directement sur leur site ou via les réseaux sociaux (Twitter et Facebook). Parmi ces sources se trouvent des documents engagés (articles de blogs, chroniques radios) ainsi que des documents plus neutres (articles de presse). Sont également présentes des parodies diffusées dans des émissions (Le petit journal, Les guignols de l'info) ou transmises via des réseaux sociaux.

Dans le cadre du projet LIMAH, l'ensemble des données vidéo ou audio sont transcrites automatiquement et seules ces transcriptions sont utilisées pour la création de liens. Les documents utilisés ont donc une qualité variable selon leur modalité (*e.g.* article de presse écrite *vs* transcription d'une émission radio) ou leur provenance (*e.g.* abréviations dans des textes issus de Twitter). Ce corpus de travail totalisant plusieurs centaines de gigaoctets de données n'a pas vocation a être diffusé.

3 Construction d'une typologie de liens adaptée au domaine journalistique

Dans cette section, nous décrivons un ensemble de typologies et expliquons leurs faiblesses pour notre cas d'usage, avant de proposer une nouvelle typologie plus adaptée. Notre description des typologies existantes s'articule autour des différentes communautés qui les ont proposées. Ces communautés ont souvent des buts différents, et choisissent donc de lier différents éléments (*e.g.* évènements, thèmes, documents) avec différents types de liens.

3.1 Typologies existantes

La création de liens entre documents, ou fragments de document, a été explorée par différentes communautés. La communauté du traitement automatique des langues s'est principalement intéressée, au travers de corpus journalistiques, à lier des évènements entre eux, le plus souvent via deux types de liens : un lien de similarité (les deux documents présentent le même évènement) et un lien de causalité temporelle (un évènement en provoque un second) (Nallapati *et al.*, 2004; Rennison, 1994; Muller & Tannier, 2004). Ces relations temporelles facilitent le parcours d'une collection de documents en proposant une navigation chronologique permettant de recomposer l'évolution d'une série d'évènements. Il nous semble néanmoins que l'utilisation de ces deux seuls types de liens est peu adaptée à un corpus multimédia, dans lequel de nom-

breux évènements sont décrits simultanément par différentes sources, commentés sur les réseaux sociaux et repris par des blogueurs. Un typage plus fin nous paraît donc nécessaire. Une seconde approche explorée par cette communauté consiste à relier des thèmes (*topics*) émergeant des documents. Une fois encore, le but principal consiste à regrouper ces thèmes et à proposer un parcours chronologique de ceux-ci (Ide *et al.*, 2004), avec les mêmes limites que celles décrites précédemment.

La communauté du multimédia s'est également intéressée à la création automatique de liens entre documents. Le plus souvent, ces liens ne sont pas typés et ne servent qu'à mettre en lumière une relation non explicite entre deux documents (Eskevich *et al.*, 2012). Ces liens non explicites se révèlent particulièrement utiles pour de petites collections, ou pour le développement de moteurs de recommandation. Cette absence de typage limite néanmoins les usages possibles et rend difficile une navigation éclairée. D'autres travaux dans ce même domaine mettent d'ailleurs en avant le besoin d'un typage pour faciliter le parcours de grandes collections (Cleary & Bareiss, 1996). Cette dernière étude expose une partie de la typologie qu'elle utilise où les 8 types les plus fréquents sont décrits sous forme de questions. Nous y trouvons des types classiques du domaine journalistique tels que les relations de causalité, mais aussi des liens de type « conseils : comment puis-je capitaliser sur cette situation ? » qui sont difficilement instanciables sur un corpus d'actualités.

La communauté des sciences de l'information et de la communication s'est aussi penchée sur le rôle des liens hypertextes. L'un de ces travaux a notamment influencé la typologie que nous proposons (Ertzscheid, 2002). Cette étude, bien que très générique et se plaçant dans un contexte éloigné du domaine journalistique, met en avant plusieurs grandes catégories de liens dont nous nous inspirons (*e.g.* une relation de récurrence, peu abordée dans les autres travaux, mais qui nous paraît pertinente dès lors que le corpus considéré est volumineux).

3.2 Typologie proposée

Dans le cadre du projet LIMAH, il nous paraît intéressant de lier des informations entre elles, plutôt que des événements trop fermés, ou des thèmes trop larges. Une information est définie par le Larousse comme « tout événement, tout fait, tout jugement porté à la connaissance d'un public plus ou moins large, sous forme d'images, de textes, de discours, de sons ». Nous choisissons d'étendre cette définition en constatant qu'une information peut également correspondre à une série d'évènements ou de faits. Un exemple concret est un sujet d'un journal télévisé. Ce sujet peut durer plusieurs minutes et réunir différents évènements présentés en un tout cohérent, qu'on appelle l'information. Cette notion de liens entre informations a déjà été exploitée dans d'autres travaux avec succès (Shahaf & Guestrin, 2010). Il s'agit donc de lier entre eux des fragments de documents, chacun de ces fragments représentant une information.

Nous proposons trois grandes catégories de liens, dont deux sont divisées en sous-catégories. Pour chacun des liens présentés, un lien inverse existe de telle sorte que tout fragment de document lié à un autre est à la fois source et cible d'un lien. Cette relation double peut se caractériser par des liens non orientés (*e.g.* un lien de type quasi duplicat est non orienté) ou par des liens duaux (*e.g.* le lien dual du développement est le résumé). Les types proposés ne sont pas exclusifs, un lien entre deux documents pouvant disposer de plusieurs types (*cf.* section 3.3). Les trois catégories retenues sont :

la récurrence : répétition d'une information. Le contenu est similaire mais peut être présenté de diverses manières, indépendamment de la modalité utilisée ;
l'extension : enrichissement d'une information. L'extension peut correspondre à un enrichissement en volume, avec un contenu plus large, ou bien à une extension temporelle correspondant à un suivi d'information ;
la réaction : l'information est commentée par un nouvel intervenant.

La récurrence est la relation la plus fréquemment rencontrée. Elle peut être envisagée sous trois formes :

le quasi duplicat : l'information est répétée, de façon similaire, sans ajout ou suppression notable ;
la citation : une référence à une information délivrée précédemment est incluse ;
la parodie : une information est reprise et détournée.

Nous considérons la parodie comme une forme de récurrence car elle reprend une information identique et change son traitement à des fins de divertissement. L'information traitée reste néanmoins la même.

L'extension enrichit une information en la développant ou en exhibant un lien temporel avec une autre information. Elle peut donc se préciser selon les deux sous-catégories suivantes :

le développement : l'information est développée, son contenu est plus important ;
la postériorité : une relation de suivi temporel est exhibée entre les deux informations.

La réaction concerne l'ensemble des commentaires qui peuvent être apportés sur une information, que ceux-ci aient lieu dans un milieu contrôlé (*e.g.* diffusion de la réponse d'un homme politique à une critique adverse) ou libre (*e.g.* réaction d'un internaute sur Twitter). Nous choisissons de ne pas offrir de sous-catégories à la réaction, bien qu'il soit possible d'utiliser les typologies existantes en analyse d'opinion pour affiner ce type (Ekman, 1992).

La typologie proposée est donc issue à la fois d'un réagencement de types couramment utilisés par la communauté, ainsi que de types rarement utilisés, mais pertinents dans le cadre d'un corpus multimodal diversifié. Elle reprend donc les relations classiques d'antériorité/postériorité, qui permettent de suivre une information d'un point de vue temporel, ou bien de source/citation, largement étudiées dans le cadre de corpus scientifiques (Nanba *et al.*, 2011; Thelwall, 2003), mais aussi des relations moins souvent exploitées telles que la parodie ou le quasi duplicat.

La figure 1 présente la typologie proposée par cette étude. Elle indique la nature des relations duales lorsqu'elles existent. Lorsqu'il n'y a pas de dualité, le lien est considéré comme non orienté. Ainsi, un lien d'antériorité entraîne nécessairement un lien inverse de postériorité, ou un lien de développement correspond toujours à un lien de résumé. Cette typologie nous paraît couvrir une très large majorité des liens possibles et permet d'envisager de nouveau moyens de parcourir une grande collection de documents.

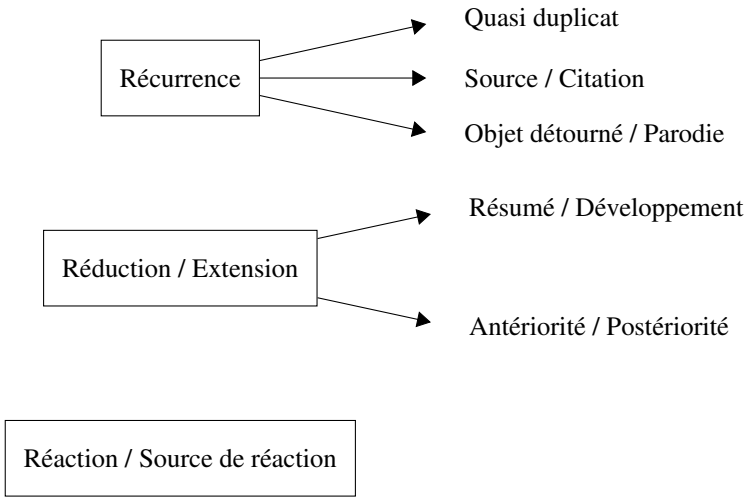

FIGURE 1: Typologie des liens entre informations

Cette typologie paraît adaptée au corpus décrit précédemment, comme le montrent les exemples développés dans la section suivante.

3.3 Exemples extraits du corpus

Nous exposons ici deux exemples extraits du corpus. Trois fragments de documents sont présentés. Le premier est un article du Figaro daté du 27 février 2015 et rapportant une allocution de Monsieur Manuel Valls lors d'un meeting électoral. Lors de cette allocution, M. Manuel Valls désigne l'extrême droite comme « l'adversaire principal ». Le deuxième est une partie d'une interview de Monsieur Florian Philippot. Durant cet entretien qui se déroule le 28 février 2015, M. Florian Philippot critique l'allocution de M. Manuel Valls et ses propos à l'encontre de son parti. Le troisième est un article du Point qui reprend par écrit l'interview de M. Florian Philippot. Également daté du 28 février 2015, l'article cite sa source et rapporte les paroles de M. Florian Philippot. La figure 2 montre les liens existant entre ces fragments de documents, en accord avec la typologie décrite précédemment.

La figure 3 reprend trois documents illustrant le dépôt de plainte de la ville de Paris après que la chaîne américaine Fox News ait qualifié certains quartiers parisiens de « no-go zones ». L'article du Point présente l'affaire tandis qu'un tweet résume l'article en reprenant l'en-tête tout en citant sa source. L'émission Le petit journal parodie l'information en décrivant le dépôt de plainte comme « une bataille entre Madame Anne Hidalgo, maire de Paris, et le premier amendement de la constitution américaine ».

Les liens duaux ne sont pas représentés sur les figures 2 et 3 pour des soucis de lisibilité.

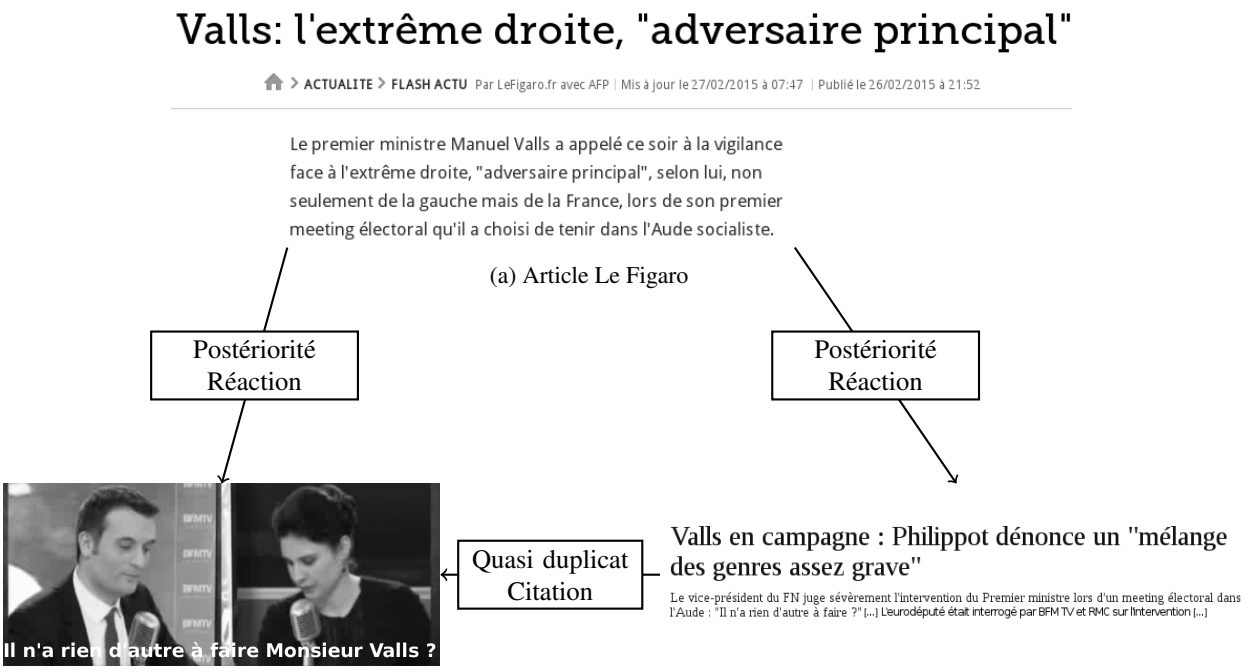

FIGURE 2: Divers liens entre trois informations

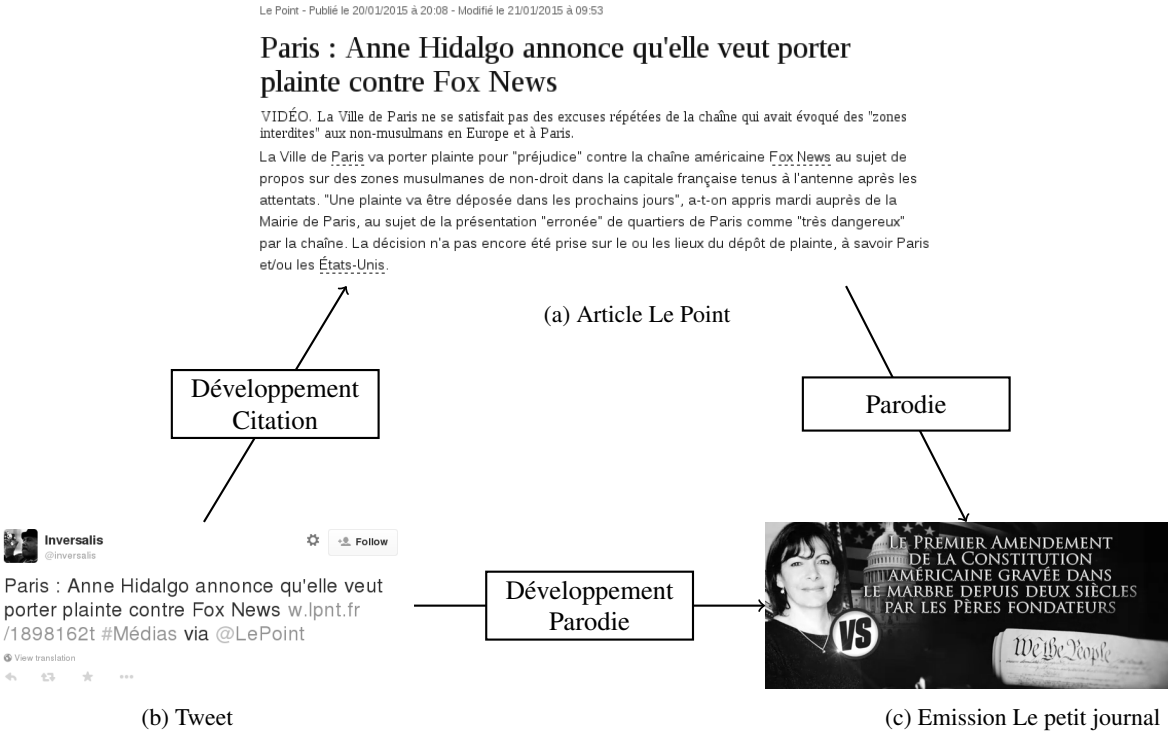

FIGURE 3: Liens de parodie et de développement

Les liens créés entre ces documents ou fragments de documents (*e.g.* dans la figure 3 on trouve des documents entiers tels le tweet ou l'article du Point ainsi qu'un fragment de l'émission télévisuelle « Le petit journal ») sont indépendants de la modalité de ces derniers. La figure 2 montre ainsi un article reprenant les propos tenus par M. Florian Philippot lors d'une interview. L'article n'apporte pas davantage d'informations que sa source, et un lien de quasi duplicat est donc créé entre les deux fragments. On peut néanmoins envisager que certains liens soient plus fréquents pour certaines modalités. Ainsi, les tweets, de par leur limite de 140 caractères, ont peu de chances de développer une information présente dans un autre média.

4 Conclusion et travaux futurs

Dans cet article, nous avons proposé une typologie pour catégoriser les liens entre des informations de type journalistique. Cette typologie offre la possibilité d'enrichir le parcours des utilisateurs en s'écartant de la logique des moteurs de recherche pour aller vers une navigation éclairée dans un ensemble de documents dont les différentes informations sont reliées entre elles.

Notre prochain objectif consiste à développer les algorithmes qui permettront de créer et de catégoriser ces liens de façon automatique. L'un des enjeux consiste à rendre ces algorithmes efficaces sur des gros volumes de données, mais aussi, à terme, d'envisager les moyens de mettre en place un système dynamique, permettant de créer de nouveaux liens pertinents au fur et à mesure que des documents lui sont présentés. Nous souhaitons également confronter notre typologie et les systèmes qui la mettront en œuvre à une réalité terrain, et préparons donc une campagne de retours d'utilisateurs afin de mieux cerner les avantages et les limites de tels liens.

La création automatique de liens typés pour enrichir un corpus multimodal est une tâche complexe et la typologie proposée est un premier pas pour encourager le développement de tels algorithmes.

Remerciements

Ce travail a bénéficié d'une aide de l'État attribuée au labex COMIN LABS et gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-07-01.

Références

CLEARY C. & BAREISS R. (1996). Practical methods for automatically generating typed links. In *Proceedings of the the seventh ACM conference on Hypertext*, p. 31–41 : ACM.

EKMAN P. (1992). An argument for basic emotions. *Cognition & emotion*, **6**(3-4), 169–200.

ERTZSCHEID O. (2002). *Le lieu, le lien, le livre : les enjeux cognitifs et stylistiques de l'organisation hypertextuelle*. PhD thesis, Université de Toulouse 2.

ESKEVICH M., JONES G. J., CHEN S., ALY R., ORDELMAN R. & LARSON M. (2012). Search and hyperlinking task at MediaEval 2012. *CEUR Workshop Proceedings*, **927**.

IDE I., MO H., KATAYAMA N. & SATOH S. (2004). Topic threading for structuring a large-scale news video archive. In *Image and Video Retrieval*, p. 123–131. Springer.

KWAK H., LEE C., PARK H. & MOON S. (2010). What is twitter, a social network or a news media ? In *Proceedings of the 19th international conference on World wide web*, p. 591–600 : ACM.

MULLER P. & TANNIER X. (2004). Annotating and measuring temporal relations in texts. In *Proceedings of the 20th international conference on Computational Linguistics*, p. 50–56 : ACL.

NALLAPATI R., FENG A., PENG F. & ALLAN J. (2004). Event threading within news topics. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, p. 446–453 : ACM.

NANBA H., KANDO N. & OKUMURA M. (2011). Classification of research papers using citation links and citation types : Towards automatic review article generation. *Advances in Classification Research Online*, **11**(1), 117–134.

RENNISON E. (1994). Galaxy of news : An approach to visualizing and understanding expansive news landscapes. In *Proceedings of the 7th annual ACM symposium on User interface software and technology*, p. 3–12 : ACM.

SHAHAF D. & GUESTRIN C. (2010). Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 623–632 : ACM.

THELWALL M. (2003). What is this link doing here ? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information research*, **8**(3).