Language-based Construction of Explorable News Graphs for Journalists

Rémi Bois and **Guillaume Gravier**
CNRS, IRISA & INRIA Rennes
263 Avenue Général Leclerc
35042 Rennes, France

Eric Jamet and **Maxime Robert**
CRPCC, Université de Rennes 2
Place du recteur Henri Le Moal
35043 Rennes, France

Emmanuel Morin
LS2N, Université de Nantes
2 Chemin de la Houssinière
44300 Nantes, France

Pascale Sébillot
INSA Rennes, IRISA & INRIA Rennes
263 Avenue Général Leclerc
35042 Rennes, France

Abstract

Faced with ever-growing news archives, media professionals are in need of advanced tools to explore the information surrounding specific events. This problem is most commonly answered by browsing news datasets, going from article to article and viewing unaltered original content. In this article, we introduce an efficient way to generate links between news items, allowing such browsing through an easily explorable graph, and enrich this graph by automatically typing links in order to inform the user on the nature of the relation between two news pieces. User evaluations are conducted on real world data with journalists in order to assess for the interest of both the graph representation and link typing in a press reviewing task, showing the system to be of significant help for their work.

1 Introduction

With content being massively made accessible grows the need for analytics and efficient organization of news collections so as to help users search and explore large amounts of content to gain knowledge and insight. Entity extraction and linking, along with topic and event detection, are now widely available to journalists in order to describe content and help search pieces of information. While these techniques are instrumental to content description and search, they are not sufficient to user-friendly exploration and navigation of a collection to gain insight, e.g., to summarize or to synthesize information. In the absence of a precise search intent, exploration is much more adapted than search.

News data have been extensively studied due to the relatively large accessibility and interest to both media professionals and general public, however mostly from the search angle. Typical search-based approaches consist in organizing datasets around clusters, in which similar or topically close news articles are grouped. The created clusters can be further processed to be displayed as threads (Ide et al., 2004), or according to temporal relations (Muller and Tannier, 2004). However, pitfalls appear when dealing with large timeframes, as the number of clusters to display becomes overwhelming. In this work, we rather focus on an exploration scenario without precise information need, where one has to get a comprehensive view on a topic in a limited amount of time, and for which the methods mentioned above are not suited. For this scenario, the usual approach consists in creating links between pairs of documents within the collection, allowing users to directly go from one news piece to another. By following links, the user is able to navigate the collection, choosing his next step among a limited set of links that are related to the news item he is currently viewing. Structures created by connecting pairs of news pieces can be seen as graphs, in which nodes correspond to documents, and edges are links between document pairs. Such collection structuring can lead to interesting applications, such as the ability to find a path connecting two arbitrary nodes, connecting the dots between two information pieces (Shahaf and Guestrin, 2010). In this context, we put forward the notion of *explorable* graphs linking news pieces in such a way that media professionals can easily find all relevant information on a topic by browsing the graph. Departing from standard approaches, e.g., $\mathcal{E}$-NN graphs, we propose a novel nearest neighbor graph construction algorithm based on lexical similarity that creates links in a reasonable number to avoid user overload and disorientation, yet ensur-

Proceedings of the 2017 EMNLP Workshop on Natural Language Processing meets Journalism, pages 31–36
Copenhagen, Denmark, September 7, 2017. ©2017 Association for Computational Linguistics

ing relevance and serendipitous drift. We further propose a typology of links between news pieces along with rules for automatic link categorization. These two elements, graph construction and link categorization, result in an explorable organization of large collections of news. We prove the interest of this organization to media professionals, and in particular that of link categorization, by means of user tests, where journalists were asked to write a synthesis on a particular topic in a limited amount of time.

2 Explorable news graph

Related studies on music recommendation have proven that explorability, or browsing capabilities, have a big impact on user experience (Seyerlehner et al., 2009) but, to the best of our knowledge, no attempts have been made at formalizing a list of necessary properties for explorable recommendations. We thus propose a set of intuitive properties that a graph should exhibit to be explorable:

Property 1: A link between two nodes indicates that those nodes are related in some way. The user should not be faced with senseless links that would lead to disorientation;

Property 2: There exists a path between any two given nodes. This ensures that the user can drift away from his original topic of interest and discover new information;

Property 3: The shortest path between any two given nodes should be reasonably small. The user can go from one topic to another in a relatively small number of steps;

Property 4: There is a reasonable amount of outgoing links for any given node. This ensures that the user is not overloaded by the number of proposed links;

Property 5: The amount of incoming links is proportional to the popularity of the node. The user should easily get access to the main topics of the collection.

The two main approaches to create graphs are the $\mathcal{E}$ nearest neighbors ($\mathcal{E}$-NN) and the K nearest neighbors (K-NN). They consist in linking each node to its closest neighbors–closeness being calculated by means of similarity measures–and rely on a fixed threshold that is either a number of neighbors K for K-NN or a similarity score $\mathcal{E}$ for $\mathcal{E}$-NN. In practice, finding their respective optimal thresholds, K or $\mathcal{E}$, is difficult and requires some annotation to estimate the ratio of irrelevant links, a process that is often complex and subjective (Ge et al., 2010). Moreover, graphs created with those methods exhibit some strong limitations in terms of explorability. K-NN graphs do not discriminate between news that are heavily discussed, and that could thus rightfully be linked to many other news pieces, and news that are reported by only a few medias, with few connections to other items. Using the same threshold K for the whole collection thus leads to links that are too few for some news items, and too numerous for others. The use of a distance threshold in $\mathcal{E}$-NN graphs skirts this issue by reducing the number of unrelevant links. However, $\mathcal{E}$-NN graphs tend to create very large hubs (Radovanović et al., 2010) , with a few nodes being connected to hundreds of others, causing navigation in such structures to be cumbersome.

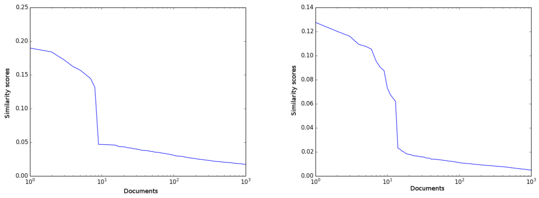
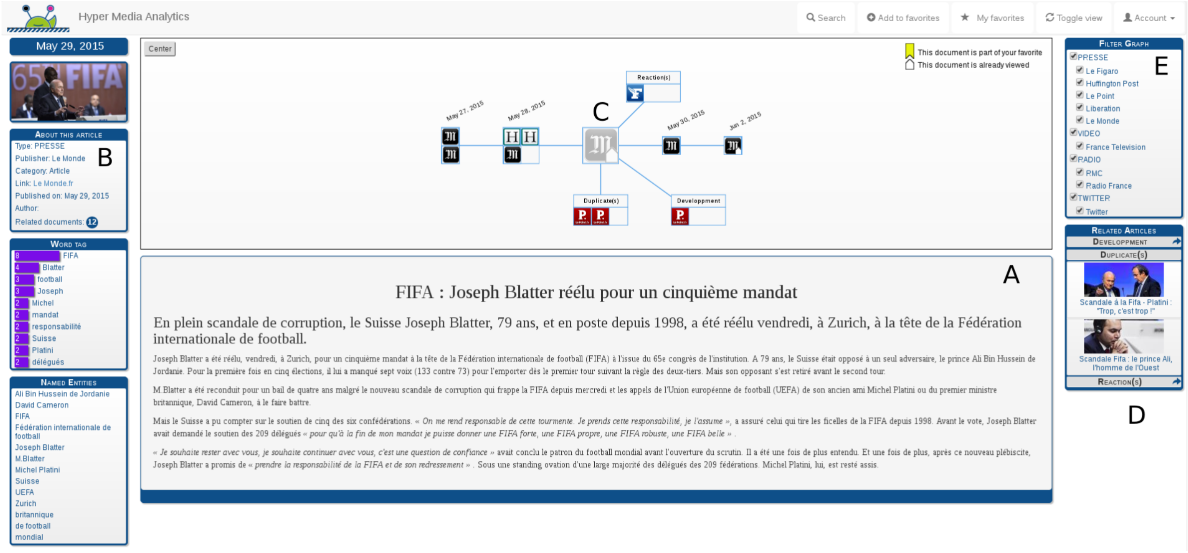


Figure 1: Illustration of similarity drops between close neighbors and far ones on two real-world examples.

Since the existence of a unique threshold for the entire collection leads to poorly crafted graphs, we propose a new method allowing to adapt the threshold on a per node basis, automatically deciding on the appropriate number of near neighbors by detecting a large gap in the representation space between close neighbors and far ones. Such gaps are known to happen naturally in large collections such as social graphs (Danisch et al., 2013) and are linked to the variations of the density of points in the representation space (Kriegel et al., 2011). For an item i corresponding to node v_i, the gap corresponds to a drop in the similarity between item i and other items sorted in descending order of similarity. Only items appearing before the gap are linked to item i. In our experiments, standard NLP approaches are used for lexical similarity scoring and drop detection. First, a tf-idf weighting and a cosine similarity measure allow us to obtain efficient similarity scores for document pairs. Then, after sorting in descending order all documents according to their similarity with a node/document of interest, we detect the largest drop in similar-



May 29, 2015

About this article

Type: PRESSE
Publisher: Le Monde
Category: Article
Link: Le Monde.fr
Published on: May 29, 2015
Author:
Related documents: 12

Word tag

8	FIFA
4	Blatter
3	football
3	Joseph
2	Michel
2	mandat
2	responsabilité
2	Suisse
2	Platini
2	délégués

Named Entities

Ali Bin Hussein de Jordanie
David Cameron
FIFA
Fédération internationale de football
Joseph Blatter
M.Blatter
Michel Platini
Suisse
UEFA
Zurich
britannique
de football
mondial

FIFA : Joseph Blatter réélu pour un cinquième mandat

En plein scandale de corruption, le Suisse Joseph Blatter, 79 ans, et en poste depuis 1998, a été réélu vendredi, à Zurich, à la tête de la Fédération internationale de football.

Joseph Blatter a été réélu, vendredi, à Zurich, pour un cinquième mandat à la tête de la Fédération internationale de football (FIFA) à l'issue du 65e congrès de l'institution. A 79 ans, le Suisse était opposé à un seul adversaire, le prince Ali Bin Hussein de Jordanie. Pour la première fois en cinq élections, il lui a manqué sept voix (133 contre 73) pour l'emporter dès le premier tour suivant la règle des deux-tiers. Mais son opposant s'est retiré avant le second tour.

M.Blatter a été reconduit pour un bail de quatre ans malgré le nouveau scandale de corruption qui frappe la FIFA depuis mercredi et les appels de l'Union européenne de football (UEFA) de son ancien ami Michel Platini ou du premier ministre britannique, David Cameron, à le faire battre.

Mais le Suisse a pu compter sur le soutien de cinq des six confédérations. « *On me rend responsable de cette tourmente. Je prends cette responsabilité, je l'assume* », a assuré celui qui tire les ficelles de la FIFA depuis 1998. Avant le vote, Joseph Blatter avait demandé le soutien des 209 délégués « *pour qu'à la fin de mon mandat je puisse donner une FIFA forte, une FIFA propre, une FIFA robuste, une FIFA belle* » .

« *Je souhaite rester avec vous, je souhaite continuer avec vous, c'est une question de confiance* » avait conclu le patron du football mondial avant l'ouverture du scrutin. Il a été une fois de plus entendu. Et une fois de plus, après ce nouveau plébiscite, Joseph Blatter a promis de « *prendre la responsabilité de la FIFA et de son redressement* » . Sous une standing ovation d'une large majorité des délégués des 209 fédérations. Michel Platini, lui, est resté assis.

Filter Graph

☑PRESSE
☑ Le Figaro
☑ Huffington Post
☑ Le Point
☑ Liberation
☑ Le Monde
☑VIDEO
☑ France Television
☑RADIO
☑ RMC
☑ Radio France
☑TWITTER
☑ Twitter

Related Articles

Developpment

Duplicate(s)

Scandale à la Fifa - Platini : "Trop, c'est trop !"

Scandale Fifa : le prince Ali, l'homme de l'Ouest

Reaction(s)

Figure 2: The LIMAH news exploration and analytics interface

ity among consecutive documents. The shallow lexical representation described above allows us to detect such drops, as illustrated in Figure 1, which do not appear when using semantic vectorial representations such as averaged word2vec or doc2vec (Mikolov et al., 2013).

Even with explorable graphs, the connection existing between two nodes can sometimes be puzzling to the user. We thus propose to characterize links between nodes according to a typology specifically crafted for news collections. News data depend a lot on chronology, which resulted in many approaches organizing collections as timelines so as to be able to follow the evolution of specific stories. The temporal relation is clearly the most important type of relations according to media professionals (Gravier et al., 2016). But it is insufficient alone, in particular when exploring large news datasets that include articles with very similar content from different newswires that tends to clutter timelines. Extending temporal relations, we used a typology consisting of 7 types of oriented links (Bois et al., 2015) defined as follows:

Near duplicate identifies a link between two nodes discussing the same event, where the target node provides little to no additional information compared to the source node;

Anterior/Posterior indicates a target node reporting on an event related to the source that occurred before (resp. after) the event of the source node;

Summary/Development corresponds to a link providing a subset (resp. superset) of information with respect to the source;

Reacts/In reaction to designates a reference (resp. followup) to another news piece in the collection.

In order to automatically categorize each link according to the above typology, we apply a set of handcrafted rules. Near duplicates are detected first based on a cosine similarity over tf-idf weighted terms. Summaries and developments are then detected by comparing documents' lengths. We then assign the reaction type by detecting cue phrases such as ”reacted to”, ”answered to”, or ”declared that”. Remaining links are considered as temporal relations and given the anterior/posterior type depending on publication dates.

3 Explorability evaluation and user validation

In order to assess for the explorability of graphs created with our novel method, we performed experiments on dataset (Gasparetti, 2016) composed of a five month monitoring of Google News over 4 categories (health, science, business, and entertainment), each of them containing around 15,000 articles. While this dataset provides a groundtruth based on clusters rather than pairing of documents, it can be used as a estimation of the correctness of our approach: elements that we link and belong to the same cluster can be considered as correct, and elements that we link but do not belong to the same cluster can be considered as incorrect. Since a perfect precision in these conditions would lead to a poorly explorable graph only composed of separate clusters, the goal here is rather to obtain explorable graphs while maintaining a high precision. Results revealed that not only our parameter-free method obtained good precision scores around the 70% mark, but also managed to regroup most nodes (over 93% of them) in a single component allowing users to drift away from topic to topic in a single walk. Results not reported within the scope of this paper show that our method builds graphs that offer much better trade-offs between precision and connectivity than K-NN and $\mathcal{E}$-NN graphs.

Interest to media professionals was evaluated by means of user testing involving journalism students. We ran experiments on a French news dataset gathered online. Documents were extracted over a 3 week period from a number of French newswires websites and include press articles, videos, and radio podcasts. Podcasts and videos underwent speech transcription so as to obtain textual representations. To deal with possibly long audio or video recordings, topic segmentation based on automatic transcripts (Guinaudeau et al., 2012) was used, each segment being treated as a document per se. In total, the resulting collection contains 4,966 news articles, 1,556 radio segments and 290 video segments. We ran our nearest neighbors algorithm on the collection as well as link categorization, creating 17,468 links in total: 10,980 temporal, 3,878 quasi-duplicates, 725 reactions, and 575 summaries/developments.

The starting point of the end-user interface[1] , called LIMAH for Linking Media in Acceptable Hypergraphs, is a full-fledged search bar using keywords. Search classically returns a list of documents ranked by relevance, from which the user can choose an entry point for navigation. Selecting an entry point brings the user to the content visualization and navigation part of the interface, composed of 5 parts, illustrated in Fig. 2. In this view, the user can initially see the entry point document itself (A) and the links that departs from it. In addition to the original content, metadata and keywords are displayed (B), as both were judged crucial in the preliminary usefulness studies (Gravier et al., 2016). Links appear in one of two ways. The graph view (C) quickly shows how related documents appear on a navigable section, facilitating the comprehension of the development of a story. Users can navigate the graph: a mouse-over on a node highlights the keywords in common with the entry point document; a click on a node enables viewing the content in zones A and C. To enable further exploration, a double click on a node defines the node as the new entry point and changes the graph and metadata displayed. For convenience, on the right side (D), links are also provided as a list of recommendations organized by link types, omitting chronological links that only appear on the graph section. At any time, filters listed in the top right section (E) allow selecting specific sources and a new entry point can be found from the search bar.

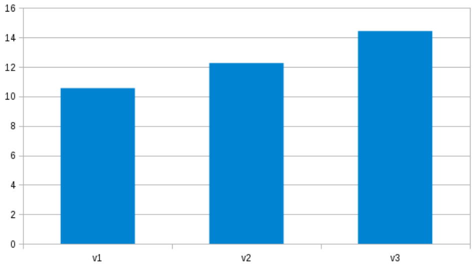


Figure 3: Knowledge extracted from the dataset depending on the version of the LIMAH interface.

In order to evaluate the interest of the graph structure and link typing to professionals, we compare three versions of the interface. Version 1 only provides the search engine, allowing for comparison with today's usage and with a technology that users are very familiar with. In this case, areas C, D, and E are hidden. Version 2 adds the recommendation and graph structure but converts all link types to temporal, organizing data in a linear fashion. Recommendations in zone D are thus uncategorized and every link in zone C is shown on a timeline. Version 3 corresponds to the whole interface, as presented above.

The study involved 25 journalism students in their last years of studies, split in three test pools of 8 to 9 people. The user test involved a pre questionnaire, an information gathering task, a post questionnaire, and a final open discussion in which users could provide feedback on their use of the tool. Users were shown a short video explaining how to use the interface, and received no additional support during tests. The information gathering task consisted in writing a synthesis about a particular subject in a limited amount of time, using the interface to find as much relevant informa-

[1] Demo available on http://limahweb.irisa.fr/texmix/

tion on the topic as possible. The chosen topic was Solar Impulse 2, a solar-powered aircraft that circumnavigated the globe from March 2015 to July 2016. Bad weather conditions necessitating the plane to land and consequences of this unexpected halt are reported in 17 articles in the dataset, representing a total of 68 distinct information pieces over a long timespan. As the dataset comes from a large set of newswires, some pieces of information are repeated, while others are mentioned by only one or two sources. Users had to complete this task in 20 minutes, a time long enough to fully read a few articles, but short enough to forbid reading totally most of them.

A preliminary manual annotation was performed on each document related to the Solar Impulse topic in order to list all individual facts and the documents in which they appeared. This annotation was used to assess for the exhaustiveness of the syntheses created by users. Exhaustiveness was measured by coding each synthesis according to the proportion of the 68 information pieces it contains. Figure 3 shows the average number of information pieces gathered by users for each version of the system under test. On average, versions 2 and 3 allowed to retrieve more information more efficiently. Results show that 10.57 (resp. 12.10 and 14.44) pieces of information were found for version 1 (resp. version 2 and version 3). Moreover, version 3 allowed to retrieve rarer pieces of information that appear in only a few documents in the collection. Surprisingly enough, the superiority of version 3 is not due to a higher amount of documents viewed by users. Rather, as shown in Figure 4, users of version 3 saw on average less documents than users of version 2, indicating that the better explorability lead to a better choice of which articles to read rather than an ability to read more of them.

During the open discussion following the tests, users from version 3 were mostly positive about their experience with the tool, calling it ”useful”, with a ”good accessibility”, and an ”interesting take on recommendation”. A few users mentioned a difficulty to handle the back and forth between the graph representation and the search interface.

4 Conclusion

Appropriate graph representations of news articles can help professionals gather information more efficiently, as evidenced by the study presented in this paper. In particular, we experimentally demonstrated that categorizing automatically hyperlinks established between articles further improves the amount and quality of the information retrieved while exploring to gain insight on a particular topic. We also proposed a parameter-free nearest neighbors algorithm that was shown to offer a better trade-off between relevance of the links and their number than standard nearest neighbors graph construction algorithms. Overall, organizing news collections in this way was proved to be helpful to journalists for their everyday work.

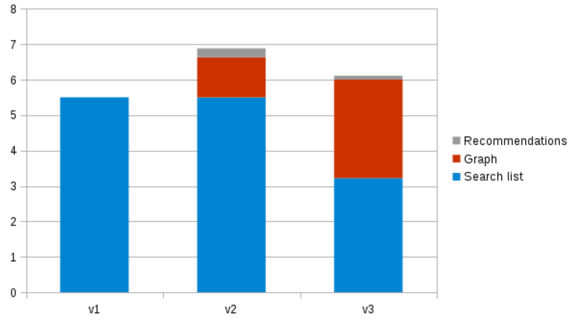


Figure 4: Number and origin of the articles viewed for the 3 versions of the LIMAH interface.

5 Acknowledgments

This work was funded by the CominLabs excellence laboratory, and financed by the National Research Agency under reference ANR-10-LABX-07-01. We are immensely grateful to Arnaud Touboulic for his key contribution in the design and implementation of the LIMAH interface.

References

Rémi Bois, Guillaume Gravier, Pascale Sébillot, and Emmanuel Morin. 2015. Vers une typologie de liens entre contenus journalistiques. In *22e Conférence Traitement Automatique des Langues Naturelles*. pages 515–521.

Maximilien Danisch, Jean-Loup Guillaume, and Bénédicte Le Grand. 2013. Towards multi-ego-centred communities: A node similarity approach. *International Journal of Web Based Communities* 9(3):299–322.

Fabio Gasparetti. 2016. Modeling user interests from web browsing activities. *Data Mining and Knowledge Discovery* pages 1–46.

Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *4th*

Conference on Recommender Systems. pages 257–260.

Guillaume Gravier, Martin Ragot, Laurent Amsaleg, Rémi Bois, Grégoire Jadi, Éric Jamet, Laura Monceaux, and Pascale Sébillot. 2016. Shaping-up multimedia analytics: Needs and expectations of media professionals. In *22nd MMM Conference, Perspectives on Multimedia Analytics*. pages 303–314.

Camille Guinaudeau, Guillaume Gravier, and Pascale Sébillot. 2012. Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation. *Computer Speech and Language* 26(2):90–104.

Ichiro Ide, Hiroshi Mo, Norio Katayama, and Shinichi Satoh. 2004. Topic threading for structuring a large-scale news video archive. In *International Conference on Image and Video Retrieval*. pages 123–131.

Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. 2011. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1(3):231–240.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. pages 3111–3119.

Philippe Muller and Xavier Tannier. 2004. Annotating and measuring temporal relations in texts. In *20th International Conference on Computational Linguistics*. pages 50–56.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research* 11(Sep):2487–2531.

Klaus Seyerlehner, Peter Knees, Dominik Schnitzer, and Gerhard Widmer. 2009. Browsing music recommendation networks. In *10th International Society for Music Information Retrieval Conference*. pages 129–134.

Dafna Shahaf and Carlos Guestrin. 2010. Connecting the dots between news articles. In *16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pages 623–632.