Pairing Online News Articles to Videos using Graph Mining:Experiments on French and English Data

Faouzia YESSAD[1,2,3], Khalid MEHL[2], and Albert BIFET[1]

[1] Telecom-ParisTech school, Paris, France
issaad.fouzia@gmail.com, albert.bifet@telecom-paristech.fr
[2] MediaBong company, Paris, France
km@mediabong.com
[3] Paris-Sud university, Orsay, France

Abstract. Here we propose a method of pairing videos to online articles. Data extracted and used for our study is the text content in order to analyse and take into consideration three aspects : Semantic similarity, temporal parameter and theme of the content using graph mining. We present the method and its evaluation according to user judgement.

Keywords: Information retrieval, graph mining, machine learning, large scale data

1 Introduction

In recent years, the online press sees an evolution, especially in 2014 when the numeric press encountered an explosion [6]. Given the evolution of multimedia and its market and the overall economy, the need for an advertising is increasing more and more. The goal of any company in creating advertises is to influence customers and push them to pass to purchase. For this, the customer needs to click on the video and/or watch it until the end. To enable this, we developed a new method to improve the video chosen according to the content of the article-videos and its date and thus better target the concerned public.

The existing system in MediaBong designed by Adele Desoyer [1] based on the vectorial model . This model proves its efficiency with a French corpus, but had some difficulties to get good results with English corpus [1]. In order to improve this point and go further in the perspective of the previous work, we propose a new method based on Graph Mining. We will present results, then discuss about them and finally a conclusion and perspectives will be given.

2 METHOD

Given dataset essentially contains videos and articles which have a title and description. Videos are given by producers such as Euronews, TF1, BFM or FashionTV and articles are available online websites of editors such as 20minutes, Europe1, Metro or Midi Libre.

Data forming the corpus are inserted in the graph. Relationships between data are explained in the following section. Nowadays, the graph contains more than 468K videos in French, 246K videos in English and 100K articles in French and more than 600 videos in each language are indexed every day.

Data are linked with edges to be processed after. Multiple types of edges : each case of edge has a specific property. Edges between video and its terms (resp. Entities[4]) or between Article and its terms (resp. Entities) have TF[2] property , whose DF[2] appears in a property of terms and entities nodes. Edges between terms or entities are made according to the position of words in the context; either constructed according to Levenshtein distance [3] or have property computed using Jaccard metric[5], or cosine similarity[4].

The construction of the pairing <article, video> is equivalent to calculate the similarity between the input article and videos witch is called semantic score *sc*. The selected videos need to get a hight *sc* and to be fresh according to the temporal aspect. Computing the final similarity score follows the formula:

$$\sqrt{sc^2 + \sqrt{\frac{1}{log_{10}(\sqrt{I+2})}}} \tag{1}$$

[4] Entity is a term appearing in the title

Where I is defined by: interval_day(I) = max((article_date - video_date).days,0)

The temporal aspect is initiated by Paul A. M. [5]. The theme of the article is predicted using linear SVM to filter only videos with the same theme as the article

3 Evaluation

In order to evaluate the system, a request is sent with an article as the input and it gives the five best videos displaying in users interface. The average run time of a request takes 10 seconds. The evaluation is given by human selection; they can select one of the proposed videos, or select a video apart from those proposed by the system (outside) if they think that videos proposed are not relevant, or in a third case they didn't select anything because there was no relevant video. Results were given when the system evaluation was done from 1st to the end of July 2016 whose theme prediction was not yet applied. System proposition and human selection are illustrated in the following table. The label 'Proposition' means the system proposes videos to users, if 'No proposition' then the system didn't

	Selection		No selection
Proposition	First	1085	497
	Top 5	84	
	Outside	600	
No proposition		712	1493
Auto-validate		11002	

Fig. 1. Confusion matrix of videos selection on July 2016

find any video in connection with article. Selection (resp. No selection) presents the number of selected videos by users (resp. didn't selected). Auto-validate is the number of videos automatically validated without proceeding by user, it concerns videos which have a semantic score greater than threshold previously defined.

4 Conclusion and perspectives

We constructed a system of online pairing articles to videos, based on graph mining with French and English datasets. We then applied a learning method: SVM linear on articles to predict themes. The learning process is launched every day on the whole data, then learned data may be processed again every day. As perspective, we think about applying a real time learning on the graph whose can proceed learning on articles only one time. Another perspective is to separate parameters of each theme, for example the theme Cooking has not a strong relationship with time whereas political news are very fresh, and named entities are frequent in sport theme. To improve the classifier, the separation of classification parameters according to themes is necessary.

References

1. A. Desoyer, D. Battistelli, J. Minel. Appariement d'articles en ligne et de videos : stratégies de sélection et méthodes d'évaluation. Actes de la conférence conjointe JEP-TALN-RECITAL 2016
2. Wu, H. C.; Luk, R. W. P.; Wong, K. F.; Kwok, K. L. (2008). Interpreting TF-IDF term weights as making relevance decisions. ACM Transactions on Information Systems
3. V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, 1966. Original in Russian in Doklady Akademii Nauk SSSR, 1965.
4. Amit Singhal. Modern Information Retrieval: A Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. 2001
5. Jaccard, Paul (1901), Comparative study of the floral distribution in a portion of the Alps and Jura, Bulletin of the Vaudoise Society of Natural Sciences, 37: 547-579.
6. La presse a moins souffert en 2014 que l'année précédente. LesEchos journal. JULIEN DUPONT-CALBO — in 16/02/2015 LesEchos.fr journal

[5] Paul Antoine Malezieux, Former employee of MediaBong company