Pairing Online News Aticles to Videos using Graph Mining

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

Telecom-ParisTech school, Paris, France[1]

issaad.fouzia@gmail.com, albert.bifet@telecom-paristech.fr

MediaBong company, Paris, France[2]

km@mediabong.com

Paris-Sud university, Orsay, France[3]

Abstract

In this poster, we propose a new method of pairing videos to online articles based on graph database. To extract information from text data, we consider three aspects : Semantic similarity, temporal parameter and content's theme using graph mining. We present the method and its evaluation according to user judgement.

Introduction

Motivations

Online press diffuses information in multi-way such as text, pictures and videos where the text accompanied with multimedia videos are increasing. This is due to the market economical evolution and the competition to satisfy the customer in order to ensure as many viewers as possible. In order to get the best performance with a market witch is always in increasing, a challenge is placed:How to associate automatically the most significant video to an article in the shortest possible time?

Objective

Our objective is to develop a new method to improve the association of videos to articles according to their content and freshness to better reach the concerned public.

Data presentation

Dataset essentially contains videos and articles which have a title and description.

- Videos are given by producers; More than 500 videos are in French and 700 videos are in English. They are indexed every day.
- Articles are available online in editors websites in two languages French and English. A thousands of articles are processed every day.

Nowadays, the graph contains more than 468K videos in French, 246K videos in English and 100K articles.

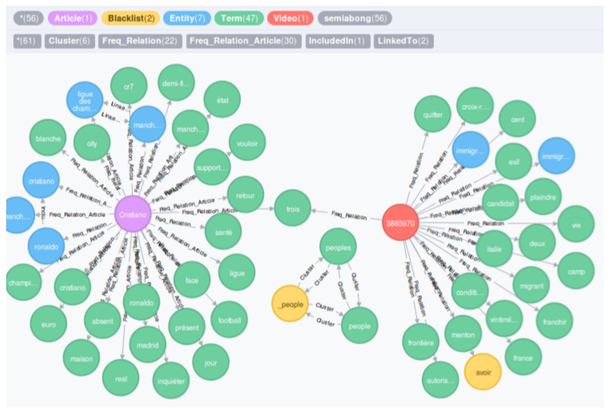

Figure 1: Illustration of graph structure.

Data are linked with edges to be processed after. Edges have TF[1] property , whose DF[1] appears in a term/entity property. Edges between terms or entities give words co-occurrences. It is constructed according to Levenshtein distance [2], Jaccard metric[4] and the shared videos quantity.

Data processing

The construction of pairs <article, video> is equivalent to calculate the similarity between the input article and videos which is called semantic score *sc*. It uses cosine similarity method [3]. 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:

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

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

The following figure gives the range of values output of the previous function judged to be acceptable.

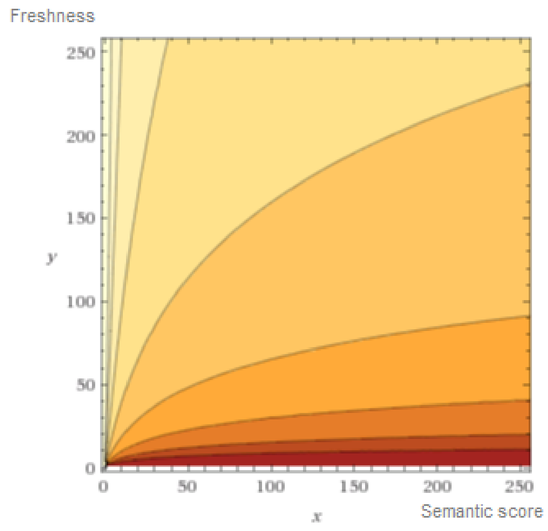

Figure 2: Response density according to the freshness and score.

Process

Once videos and the on-line article are parsed and saved on graph database, this latter is used as input to the automatic classifier and it follows this process below.

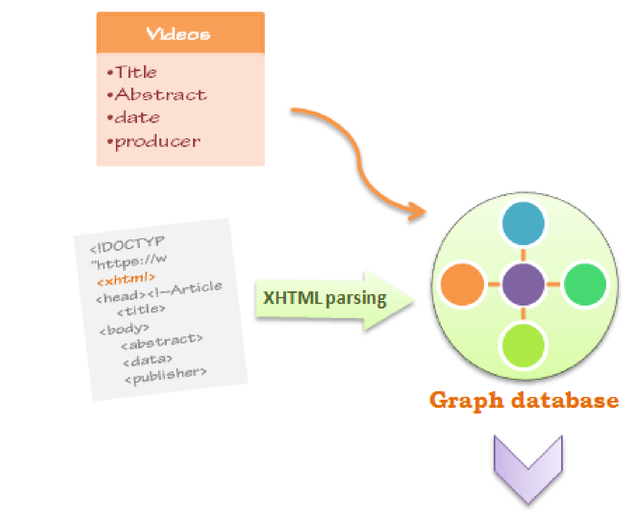

Figure 3: Data transformation

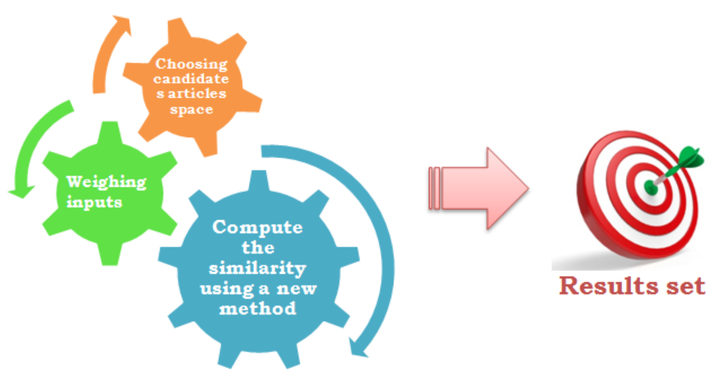

Figure 4: Data processing

Thematization

Results are filtered according the time score aspect and semantic score aspect. These dimensions are not sufficient. To improve results, the proposed videos need to have the same theme as the input article. To do this, themes videos are given in the MediaBong database but online articles themes are not known. In order to know online articles themes, Support Vector Machine learning algorithm is applied and article themes are predicted.

Results and evaluation

User evaluation results gives three data sets:

- Pertinent pairing are given with green points
- Not pertinent pairing are given with red points
- Accepted pairing are given with cyan points

The following evaluation is a Media Actuality subset. It gives the best separator according to time score and semantic score.

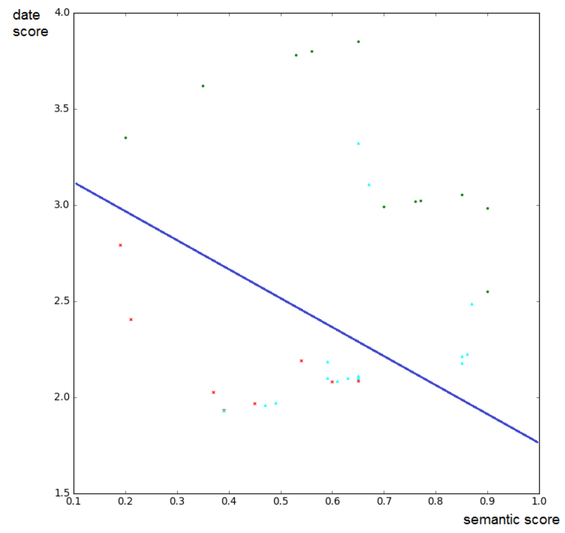

Figure 5: Scores separation

Thematization of articles is done with SVM algorithm using different thresholds separation.

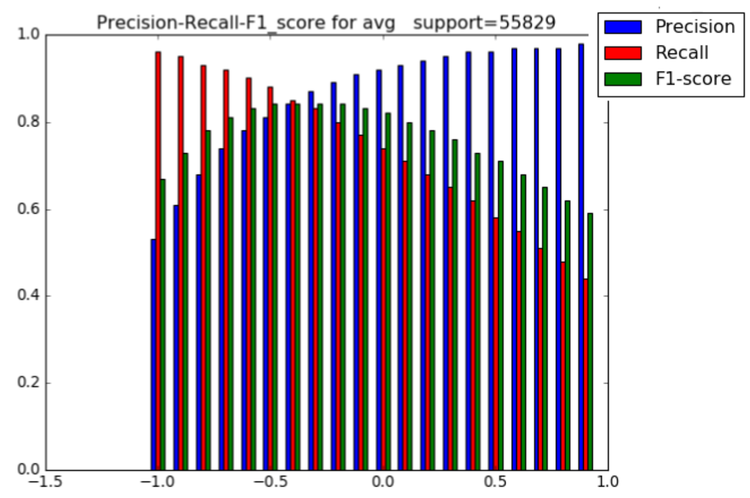

Figure 6: Themes prediction results

The whole users evaluation of the system output on the summer 2016 are given in the following table.

	Selection		No selection
Proposition	First	2937	1359
	Top5	262	
	Outside	2217	
No-proposition		2724	3975
Auto-validate		26756	

Figure 7: Confusion matrix

Conclusions and perspectives

We constructed a system of online pairing articles to videos, based on graph mining with French and English datasets. Then, we applied a learning method: SVM linear on articles to predict themes. As perspective, we think about separating parameters of each theme, for example Cooking theme do not have the same freshness sensibility degree as Politics once. To improve the classifier, the parameters classification separation according to themes is necessary.

References

[1] 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

[2] V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, 1966. Original in Russian in Doklady Akademii Nauk SSSR, 1965.

[3] Amit Singhal. Modern Information Retrieval: A Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. 2001

[4] 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.