Diachronic'Explorer : keep track of your clusters !

Nicolas Dugué, Jean-Charles Lamirel, Pascal Cuxac

Diachronic'Explorer : keep track of your clusters !

Nicolas Dugué, Jean-Charles Lamirel
LORIA-Synalp
Vandoeuvre lès Nancy, France.
e-mail : first.last@loria.fr

Pascal Cuxac
CNRS-INIST
Vandoeuvre lès Nancy, France.
e-mail: pascal.cuxac@inist.fr

***Abstract*—We introduce *Diachronic'Explorer*, a toolbox to produce and visualize diachronic results, which is based on a new complete theoretic framework for diachrony that we detail. This toolbox, which is dedicated to run diachronic algorithms from clustering results, allows also to explore complex diachronic results at all the granularity levels through a web application.**

I. Introduction

The ISTEX project compiles and provides access to scientific publications to French research. ISTEX-R project aims to highlight evolution of research topics using the ISTEX database. To that aim, we designed a new theoretical framework to track scientific fields changes and similarities across time : appearing and disappearing topics, splitting and merging topics. This framework is based on diachronic analysis, which aims to compare data or results from two distinct periods. Furthermore, this framework was implemented as a toolbox called *Diachronic'Explorer* that integrates all diachronic analysis steps and provides efficient visualizations.

In the context of our framework, we basically consider tp time periods, containing documents split into separate sets according to these time periods and described by a set of features F (words or expressions) : $D = \cup_{0 \leq i \leq tp} D_i$. For each time period, clustering algorithms detect cluster sets such as $CS = \cup_{0 \leq i \leq tp} C_i$. Each set of clusters $C_i = \{c_1, \cdots, c_{n_i}\}$ is constituted of $n_i = |C_i|$ clusters describing data D_i. Clusters gather similar documents and thus represent topics. Consequently, to monitor topic changes, our framework aims to track cluster changes and similarities across time periods.

The **first step** of our new framework consists in **labeling clusters** of every cluster sets. These labels will then be used to track cluster changes and similarities. Our labeling step consists in extracting prevalent features sets S_{c_j} for each cluster $c_j \in C_i$ of all the cluster sets $C_i \in CS$ using a feature selection method. We implement the *feature maximization* method proposed by Lamirel et al. [3] that has proven to be efficient for both supervised and unsupervised learning. The **second step** processes **diachronic analysis** between time periods using cluster labels. Basically, using S_{c_j}, the prevalent features sets of each cluster, we apply a diachronic algorithm based on MVDA, an unsupervised Bayesian process operating between views (here periods) for diachronic mining [4].

II. Feature selection for cluster labeling

The first step of our framework consists in **labeling clusters** using a feature selection method. The one we implement is non-parametric, only weakly influenced by feature scaling, and it allows to label clusters with ordered weighted features [3]. We apply it independently to each period. Thus, for each cluster set $C_i, 0 \leq i \leq tp$, two ratios are processed over each cluster $c_j \in C_i, 0 \leq j \leq n_i$ to give a weight to features of F for the cluster considered. On the first hand, *Feature Predominance* (Equation 1) aims to evaluate if a feature $f \in F$ allows to discriminate c_j from other clusters of C_i. On the other hand, *Feature Recall* (Equation 2) aims to evaluate whether f allows to faithfully describe data in c_j.

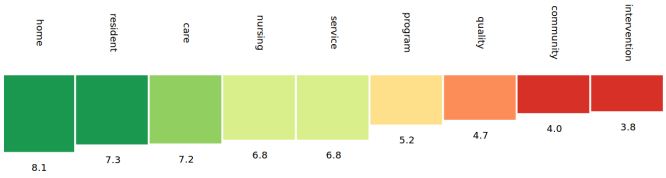

Fig. 1. Cluster labels extracted and visualized with *Diachronic'Explorer*.

$$FP_{c_j}(f) = \frac{W^f_{c_j}}{W_c} \quad (1) \qquad FR_{c_j}(f) = \frac{W^f_{c_j}}{W^f_{D_i}} \quad (2)$$

with $W^f_{c_j} = \sum_{d \in c_j} W^f_d$ the weighted sum of f across cluster c_j, $W^f_{D_i} = \sum_{d \in D_i} W^f_d$ the weighted sum of f across the data of the time period considered, $W_{c_j} = \sum_{f' \in F} W^f_{c_j}$ the weighted sum of all features across cluster c_j. The Feature *F-Measure* $FF_{c_j}(f)$ is the harmonic mean of these two values.

The Feature *F-Measure* is thus used to select features that are discriminant and representative for each cluster c_j using Equation 3.

$$S_{c_j} = \left\{ f \in F | FF_{c_j}(f) > \overline{FF_{C_i}(f)}, FF_{c_j}(f) > \overline{FF_{D_i}} \right\} \quad (3)$$

with $\overline{FF(f)}$ the mean F-Measure value of f across clusters of C_i where f is non-null, and $\overline{FF_{D_i}}$ the mean F-Measure across all features and all data of the considered period. Figure 1 shows how features selected for a cluster with *Diachronic'Explorer* are used as **labels** to describe the cluster in the visualization module.

III. Diachrony with MVDA

Once the representative labels of each cluster extracted, our second step consists in applying MVDA **diachronic analysis** between each time periods pairs. MVDA was firstly experimented as a fully unsupervised approach by Lamirel [4],

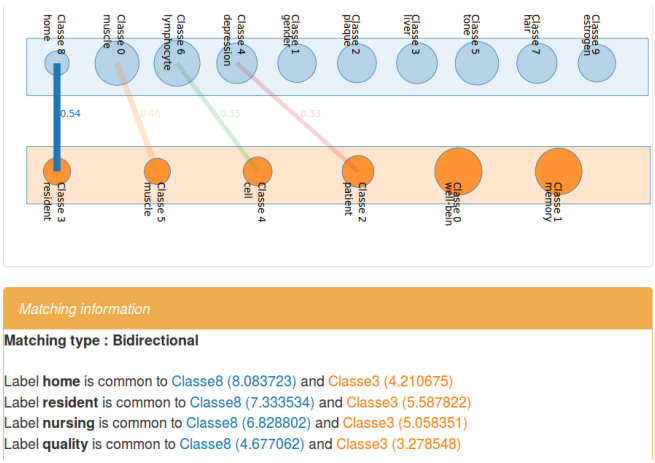

Fig. 2. Exploring matchs between two sets of clusters using *Diachronic'Explorer*. The *matching kernel* of *Classe8* and *Classe3* is displayed in the yellow box while the match strength between these clusters is represented with the thickness of the link between them.

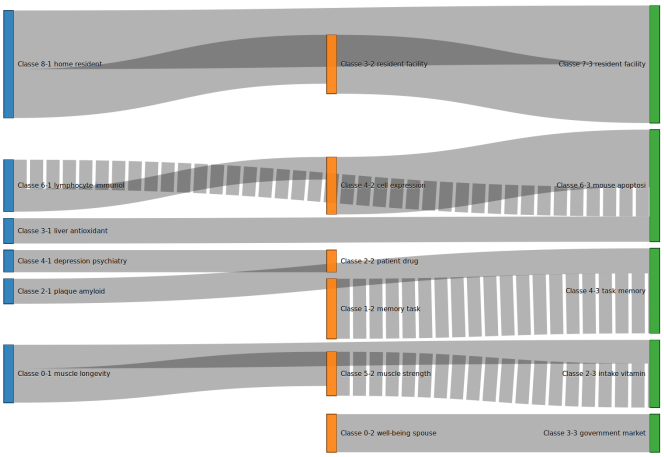

Fig. 3. Exploring cluster labels of a cluster set using *Diachronic'Explorer*. Each column/color describes a period where rectangles stands for labels clusters. Grey flows between the colored rectangles represent different kind of matchs (asymmetric/symmetric) and their strength.

but here, we use cluster labels instead of indexer keywords. This process is theoretically based on Bayesian reasoning [4] whose operating mode can be summarized as this in our context : the diachronic analysis is processed by computing matching probabilities between clusters of distinct sets (thus time periods) using the selected labels they share or not. The matching probability of a cluster s from a set C_{src} and a cluster t from a set C_{tgt} is computed as shown in Equation 4.

$$P(t|s) = \frac{\sum_{f \in S_s \cap S_t} FF_t(f)}{\sum_{f \in S_t} FF_t(f)} \quad (4)$$

The matching probability between t and s is processed symmetrically. Finally, an asymmetric match from s to t is detected if the matching probability $P(t|s)$ is greater than the average matching probability of s with the clusters of C_{tgt}. A symmetric match exists if both asymmetric match from s to t, and from t to s are detected. In such a case, the set of features describing this match is called the *matching kernel* (Figure 2). These different kind of match detected by *Diachronic'Explorer* allow us to track and describe changes and similarities between cluster sets of distinct periods.

IV. Experimental results and visualizations

Experiments and visualizations shown in this papers were conducted on a corpus constituted of bibliographic data from the ISTEX project [6]. The ISTEX database is queried to extract papers related to research in medical care between years 1996 and 2010. This results in a dataset of 9779 papers. Because the goal is to observe evolution of this field across time, a community detection algorithm [5] exploiting the relationship between documents indexes and the publication years is used to extract meta-periods. Obtained meta-periods are 1996-2000, 2000-2005, 2006-2010 [1]. GNG clustering [2] is launched several times on the data of each meta-period to extract an optimal number of clusters with the help of ad-hoc quality indexes [4]. Then, using *Diachronic'Explorer*[1], we produce and then visually explore the diachronic results with powerful dynamic visualizations using advanced Javascript[2]. Visualization tool of *Diachronic'Explorer* makes possible to navigate easily from one granularity level to another : cluster and their labels (Figure 1), match between two periods (Figure 2), overall dataset evolution (Figure 3).

V. Conclusion

Integrating a new framework using powerful labeling method and unsupervised Bayesian reasoning among clusters sets, *Diachronic'Explorer* allows to produce diachronic results and provides a web application to visualize these results, navigating easily from one granularity level to another.

Acknowledgments. ISTEX receives assistance from the French state managed by the National Research Agency under the program *Future Investments* bearing the reference ANR-10-IDEX-0004-12.

References

[1] Cuxac, P. and Lamirel, J. C. Analysis of evolutions and interactions between science fields: the cooperation between feature selection and graph representation. 14th COLLNET Meeting (2013)
[2] Fritzke, B. A growing neural gas network learns topologies. Advances in Neural Information Processing Systems 7, 625–632 (1995)
[3] Lamirel, J.-C., Cuxac P., Chivukula A.S., Hajlaoui K.: Optimizing text classification through efficient feature selection based on quality metric. Journal of Intelligent Information Systems, 2013, 1–18 (2014)
[4] Lamirel, J.-C. : A new approach for automatizing the analysis of research topics dynamics: application to optoelectronics research. Scientometrics 93(1), 151–166 (2012)
[5] Pons, P. and Latapy, M.: Computing Communities in Large Networks Using Random Walks. Journal of Graph Algorithms and Applications, 10(2): 191–218 (2006)
[6] Dugué, N. and Tebbakh, A. and Cuxac, P. and Lamirel, J.-C.: Feature selection and complex networks methods for an analysis of collaboration evolution in science: an application to the ISTEX digital library. 5th International Symposium ISKO-MAGHREB (2015)

[1]Demo - Choose "2tiers1tiers" experiment - http://107.170.67.36:3000/
Git - https://github.com/nicolasdugue/istex-demonstrateur
[2]https://d3js.org/