Diachronic’Explorer: Keep track of your clusters!
Nicolas Dugué, Jean-Charles Lamirel, Pascal Cuxac

To cite this version:
Nicolas Dugué, Jean-Charles Lamirel, Pascal Cuxac. Diachronic’Explorer: Keep track of your clusters!. RCIS, Jun 2017, Grenoble, France. pp.1 - 2, 10.1109/RCIS.2016.7549367. hal-01444752

HAL Id: hal-01444752
https://hal.archives-ouvertes.fr/hal-01444752
Submitted on 24 Jan 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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. Consequently, to monitor topic changes, our framework aims to compare data or results from two distinct periods, containing documents split into separate sets of all the cluster sets 

II. FEATURE SELECTION FOR CLUSTER LABELING

The first step of our 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 two 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].

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],
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:

\[
P(t|s) = \frac{\sum_{f \in S \cap T} FF_t(f)}{\sum_{f \in S} FF_t(f)}
\]

(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 paper 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 time, is launched several times on the data of each meta-period. 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'Explorer1, we produce and then visually explore the diachronic results with powerful dynamic visualizations using advanced Javascript2. 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


1Demo - Choose “2tiers1tiers” experiment - http://107.170.67.36:3000/
2Git - https://github.com/nicolasdugue/istex-demonstrateur

https://d3js.org/