A new method for automatically analyzing the research dynamics: application on optoelectronics research

Jean-Charles Lamirel [1] and Raghvendra Mall[2]

[2] *lamirel@loria.fr*
LORIA, INRIA-TALARIS Project, 615 r. du Jardin Botanique, 54600 Villers-lès-Nancy (France)

[2] *raghvendra.mall@research.iiit.ac.in*
Center of Data Engineering, IIIT Hyderabad, 500032 Hyderabad, Andhra Pradesh (India)

Introduction

Diachronic analysis is based on the application of a clustering method on data associated with two, or more, successive periods of time, and on the study of the evolution of the clusters contents and of their mappings between the different periods. Analyzing the difference between time periods concerns different kinds of topics changes or similarities that could occur between the periods (appearing topics, disappearing topics, splitting topics, merging topics, stable topics). For that purpose, Shiebel and al. (2010) recently proposed to construct a matrix of keywords comparison which is based on the percentage of keywords of one period which pre-exist in the clusters of another period. Thanks to this matrix, it is then possible for an expert of the domain to highlight different cluster behaviors. An important limitation of this approach is that the process of comparison between clustering models must be achieved in a supervised way. Using the dynamic and unsupervised cooperation between clustering models, firstly introduced by Lamirel and al. (2004) in the *MultiView Data Analysis paradigm* (MVDA), we thus propose hereafter to fully automatize the process of diachronic analysis.

Method

Our approach is a *label-based approach*. In such an approach a second step of cluster labeling is thus achieved after the construction of the clustering model for each time period. The goal of the labeling step is to figure out which peculiar properties or labels can be associated to each cluster of a given time period. The identification of the topics relationships between two time periods is then achieved through the use of Bayesian reasoning relying on the extracted labels that are shared by the compared periods. To compute the probability of matching between clusters belonging to two time periods, we adapt the standard computation of Bayesian communication of the MVDA model to cluster label comparison, as it is described hereafter.

Let $P(t|s)$ be defined as the probability of activity of a target period cluster t knowing the activity of a source period cluster s. It can be expressed as:

$$P(t|s) = \frac{\sum_{l \in L_s \cap L_t} L_t F(l)}{\sum_{l \in L_t} L_t F(l)} \quad (1)$$

where L_x represent the set of labels associated to the cluster x, using a suitable cluster labeling approach, and $L_x \cap L_y$ represent the common labels, which can be called the label **matching kernel**, between the cluster x and the cluster y.

The average matching probability $P_A(S)$ of a source period cluster can be defined as the average probability of activity generated on all the clusters of the target period clusters by its associated labels:

$$P_A(s) = \frac{1}{|Env(s)|} \sum_{t \in Env(s)} P(t|s) \quad (2)$$

where $Env(s)$ represent the set of target period clusters activated by the labels of the source period cluster s.

The global average activity A_s generated by a source period model S on a target period model T can be defined as:

$$A_S = \frac{1}{|S|} \sum_{s \in S} P_A(s) \quad (3)$$

Its standard deviation can be defined as σ_s.

The **similarity** between a cluster s of the source period and a cluster t of the target period is established if the 2 following similarity rules are verified:

4) $P(t|s) > P_A(s)$ *et* $P(t|s) > A_s + \sigma_s$.

5) $P(s|t) > P_A(t)$ *et* $P(s|t) > A_t + \sigma_t$.

Cluster splitting is verified if there is more than one cluster of the target period which verifies the similarity rules with a cluster of the source period. Conversely, **cluster merging** is verified if there is more than one cluster of the source period which verifies the similarity rule with a cluster of the target period. Cluster of the source period that do not have similar cluster on the target period are considered as **vanishing clusters**. Conversely, clusters of the target period that do not have similar cluster on the source period are considered as **appearing clusters**.

Experiment and results

For our experiment, we reuse the original dataset of Schiebel et al. (2010) issued from the PROMTECH project. This dataset consisted of 3890 PASCAL records related to the topic of optoelectronics research over 6 years. To carry out the diachronic analysis, the dataset has been cut into two periods.

Our selected clustering technique is the "Growing Neural Gas" (GNG) (Fritske, 1995) neural method. For each period, many different clustering models are generated, letting varying the expected number the number of clusters. The best clustering model (i.e. the optimal number of clusters) for each period, as regards to the values of the unbiased *Recall-Precision* quality indexes defined by Lamirel et al. (2004), are finally kept for the further processes of labeling and time comparison.

The exploited cluster labeling technique is a high performance labeling method based on cluster data properties information maximization, recently proposed by Lamirel et al. (2010).

Table 1. Global time comparison results.

TIME PERIOD	NBR GROUPS	NBR MATCH	NBR DISAPEAR	NBR APPEAR	NBR SPLIT	NBR MERGE
1996-1999	43	33	10	-	7	-
2000-2003	50	38	-	12	-	3

Table 1 summarizes the results of the final step of time periods comparison, in terms of identification of correspondences and differences. It should be noted that the number of splitting of clusters of the first period into the second period is more important than its converse number of merging, which indicates a diversification of the research in the field of optoelectronics during the second period.

The similarities between the clusters of the various periods are identified by shared groups of labels (i.e. **matching kernels**), which we have also named **core-labels**. These **core-labels** illustrate in a specific way the nature of the temporal correspondences. On the one hand, small temporal changes can be identified in the surrounding context of these labels, and on another hand, the more important temporal changes can be materialized by the isolated clusters whose labels do not take part in any **core**.

As the examples of Figure 1 illustrate it, the method makes it possible to very clearly reveal the developments of the research topics in context. The first report highlights the development of research works on **polymer films** and the passage of the **theoretical studies** to the **practical applications** (from **optical polymers** to **polymer films**). The second report highlights the disappearance of research on **optical fibers** during the second period.

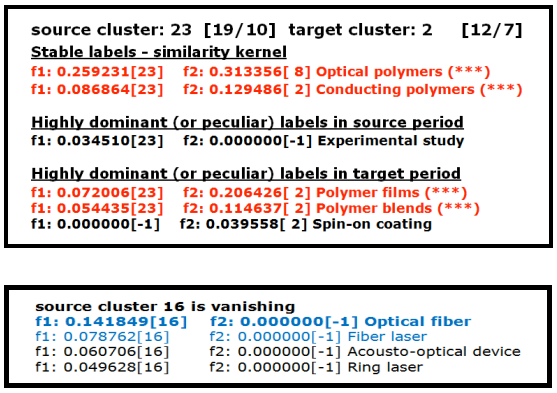

Figure 6. Two examples of automatic reports of correspondence and differences between periods.

Conclusion

We show in this paper the feasibility of an unsupervised incremental approach based on a time-step analysis of bibliographical data. This analysis has been carried out thanks to the exploitation of a specific model of data analysis managing multiple views on the data, namely the MVDA model. It was also based on the exploitation of original and stable measures for evaluating the quality and the coherence of the clustering results, and even for precisely synthesizing (i.e. labeling) the content of the clusters.

References

Al Shehabi S., Lamirel J.-C. (2004). Inference Bayesian Network for Multi-topographic neural network communication: a case study in documentary data. Proceedings of ICTTA, Damas, Syria, April 2004.

Frizke B. (1995). A growing neural gas network lerans topologies. Tesauro G., Touretzky D. S., leen T. K., Eds., Advances in neural Information processing Systems 7, pp 625-632, MIT Press, Cambridge MA.

Lamirel J.-C., Al-Shehabi S., François C., Hoffmann M. (2004). New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping. Scientometrics, 60(3).

Lamirel J.-C., Ta A.P. and Attik M. (2008). Novel Labeling Strategies for Hierarchical Representation of Multidimensional Data Analysis Results. IASTED (AIA), Innsbruck, Austria, February 2008.

Schiebel E., Hörlesberger., Roche I., François C., Besagni D. (2010). An advanced diffusion model to identify emergent research issues: the case of optoelectronic devices. Scientometrics. Vol. 83, N° 3, pp 765-781, 2010.