

Feature selection and graph representation for an analysis of science fields evolution: an application to the digital library ISTE^X

Jean-Charles LAMIREL¹ and Pascal CUXAC²

¹ SYNALP Team-LORIA, INRIA Nancy-Grand Est, Vandoeuvre-lès-Nancy, France
jean-charles.lamirel@loria.fr

² Inist-CNRS, 2, allée du parc de Brabois, 54519 Vandoeuvre-lès-Nancy, France
pascal.cuxac@inist.fr

Abstract. This paper presents an original approach based on a recent metric called feature maximization for developing accurate diachronic analysis tools. In such process, querying of bibliographic databases is firstly exploited to provide a thematic corpus of scientific publications covering a large time period. In a second step, two strategies based on contrast graphs generated by the use of feature maximization metric are proposed. The first one is based on the direct use contrast graphs who relates time periods and publication contents. The second strategy combines a preliminary step of clustering with the use of contrast graph generated by feature maximization applied on cluster contents to highlight the relation between topics represented in clusters as well as to embed them in a temporal path. Both techniques are parameter-free and knowledge agnostic. We illustrate the efficiency and the complementarity of the proposed technique by experimenting then on a dataset related to gerontology research extracted from the data collected by the ISTE^X project, a project whose aims is to construct a general purpose and open access database of scientific documents.

Keywords: Feature selection, graph-based approach, diachronic analysis, visualization, big data management.

1 Introduction

The ISTE^X¹ project's main objective is to provide the whole French higher education and research community with on-line access to retrospective collections of scientific literature in all disciplines.

On the basis of the initial platform services, we are currently working towards proposing new added-value services. One of our central concern is then to develop tools for highlighting the dynamics of the collection. Hence, the development of dynamic information analysis methods, like incremental clustering and novelty detection techniques, is becoming a central concern in a bunch of applications whose main goal is to deal with large volume of textual information whose content is varying over time, such as ISTE^X. The purpose of the analysis and diachronic mapping is to track, for a given domain, changes in contexts (sub-themes) and the evolution of vocabularies and actors

¹ <http://www.istex.fr/>

that materialize these changes in terms of appearances, disappearances, divergence or convergence. The applications relate to very various and highly strategic domains, including web mining, technological and scientific survey.

In order to identify and analyze the emergence, or to detect changes in the data, we have previously proposed two different and complementary approaches:

1. Performing static classifications at different periods of time and analyzing changes between these periods (time-step approach or diachronic analysis);
2. Developing methods of classification that can directly track the changes: incremental clustering methods (incremental clustering) and novelty detection methods (incremental supervised classification).

The development of direct methods being still an ongoing research, we present hereafter two original word-based methods relying on the first approach and using a metric called feature maximization we have recently developed (Lamirel and al. 2013). The goal of the two methods that are based on contrast graphs derived from this metric is to tackle with document belonging to the same scientific field in order to detect significant topic differences between documents related to different time periods:

1. Unlike common approaches based on graph analysis (Porter and Rafols 2009) (Sayama and Akaishi 2012), our first approach is a supervised approach that establishes a bipartite contrast graph between documents time stamps and documents salient keywords, those latter being extracted through a feature selection process based on feature maximization.
2. Our second approach is an unsupervised approach based on clustering. Thanks to this approach optimal number of clusters (i.e. topics) is extracted from the whole document dataset and relations between extracted topics and selected salient keywords are used to form the bipartite contrast graph. Documents timestamps are exploited in a second step to highlight diachronic changes and diachronic path between topics.

We first present our feature maximization metrics and related contrast graph exploited throughout our approach. In a next step, we describe our experimental data and associated preprocessing. Lastly, we highlight our results with the two proposed approaches and our conclusion.

2 Feature maximization, feature selection and contrast graph

2.1 Feature maximization

Feature maximization (F-max) is an unbiased cluster quality metrics that exploits the properties of the data associated to each cluster without prior consideration of clusters profiles. This metrics has been initially proposed in (Lamirel and al 2004). Its main advantage is to be independent altogether of the clustering methods and of their operating mode.

Consider a partition \mathcal{C} which results from a clustering method applied to a dataset D represented by a group of features F . The feature F-measure $FF_c(f)$ of a feature f associated with a cluster c is defined as the harmonic mean of the feature recall $FR_c(f)$ and of the feature predominance $FP_c(f)$, which are themselves defined as follows:

$$FR_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{c \in \mathcal{C}} \sum_{d \in c} W_d^f} \quad (1)$$

$$FP_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{f' \in F_c, d \in c} W_d^{f'}} \quad (2)$$

with

$$FF_c(f) = 2 \left(\frac{FR_c(f) \times FP_c(f)}{FR_c(f) + FP_c(f)} \right) \quad (3)$$

where W_d^f represents the weight of the feature f for the data d and F_c represents all the features present in the dataset associated with the cluster c . Feature Predominance measures the ability of f to *describe* cluster c . In a complementary way, Feature Recall allows to characterize f according to its ability to discriminate c from other clusters.

Feature recall is a scale independent measure but feature predominance is not. We have however shown experimentally in (Lamirel, Cuxac, et al. 2015) that the F-measure which is a combination of these two measures is only weakly influenced by feature scaling. Nevertheless, to guaranty full scale independent behavior for this measure, data must be standardized. Furthermore, the choice of the weighting scheme for data is not really constrained by the approach, but it is necessary to deal with positive values. Such scheme is supposed to figure out the significance (i.e. semantic and importance) of the feature for the data².

2.2 Feature selection

In supervised context, feature maximization measure can be exploited to generate a powerful feature selection process (Lamirel, Cuxac, et al. 2015). In our unsupervised (clustering) context, the selection process can be used to describe or label clusters according to the most typical and representative features. This process is a non-parametrized process that uses both the capacity of F-measure to discriminate between clusters ($FR_c(f)$ index) and its ability to faithfully represent the cluster data ($FP_c(f)$ index).

The set S_c of features that are characteristic of a given cluster c belonging to a partition \mathcal{C} is translated by:

$$S_c = \{f \in F_c \vee FF_c(f) > \overline{FF}(f) \wedge FF_c(f) > \overline{FF}_D\} \quad (5)$$

² A feature having some negative values can be separated in 2 different positive sub-features, the first one representing the positive part of original feature and the second one, its negative part.

$$\overline{FF}(f) = \sum_{c \in C} \frac{FF_c(f)}{|C_f|} \quad \text{and} \quad \overline{FF}_D = \sum_{f \in F} \frac{\overline{FF}(f)}{|F|} \quad (5)$$

where C_f represents the subset of C in which the feature f occurs.

Finally, the set of all selected features S_C is the subset of F defined by:

$$S_C = \bigcup_{c \in C} S_c \quad (6)$$

In other words, the features judged relevant for a given cluster are those whose representations are better than average in this cluster, and better than the average representation of all the features in the partition, in terms of Feature F-measure. Features which never respect the second condition in any cluster are discarded.

2.3 Contrast

A specific concept of contrast $G_c(f)$ can be defined to calculate the performance of a retained feature f for a given cluster c . It is an indicator value which is proportional to the ratio between the F-measure $FF_c(f)$ of a feature in the cluster c and the average F-measure \overline{FF} of this feature for the whole partition. Contrast of a feature f for a cluster c is expressed as:

$$G_c(f) = FF_c(f) / \overline{FF}(f) \quad (7)$$

The active features of a cluster are those for which the contrast is greater than 1. Moreover, the higher the contrast of a feature for one cluster, the better its performance in describing the cluster content.

2.4 Contrast graphs

In the mathematical field of graph theory, a bipartite graph (or bigraph) is a graph whose vertices can be divided into two disjoint and independent sets U and V such that every edge connects a vertex in U to one in V . Contrast graphs are bipartite graphs based on the relations between a set of features S and a set of labels L (Cuxac and Lamirel 2013). Theoretically, the set of labels L could represent any kind of information to which features can be related with and the set of features S is a subset of a global feature set F (i.e. the original feature space on which rely the data of a dataset) that has been obtained through a feature selection process, like feature maximization presented above. In the case of the use of feature maximization, the weight $c_{(u,v)}$ of an edge (u, v) , $u \in S$, $v \in L$ represents the contrast of feature u for a label v as, it is defined by equation 7.

Such kind of graphs have many interesting properties. First, they reduce the cognitive overload produced with classical graphs representation because of the associated feature selection process that reduces the number of potential connections. Second, they can be used to indirectly highlight relationships between labels, whenever features have contrasted interaction with several labels. Third, the combination of this approach with

weighted force-directed model (Kobourov 2012) for graph representation permits altogether to highlight central or most influent labels of the L set and to easily identify the labels that are the most densely connected through associated features, these latter appearing in close neighborhood position in the graph.

We have proposed a first original use of contrast graph in the case of the analysis of the transdisciplinarity between different research domains and time periods in (Cuxac and Lamirel 2013).

3 Data

Our experimental data is a collection of 9801 scientific papers in English language related to gerontology domain published between 1995 and 2010, extracted from ISTEK database by INIST documentary engineers specialized in the medical domain. After a tokenization step, the keywords are extracted from the abstracts by a part-of-speech method developed in Python. However, the NLP treatments are minimalist (just morpho-lexical and syntactic) and we thus don't use any vocabulary resource except a stop word dictionary.

We present on the following experimental section the two different approaches we have applied on the extracted metadata of our dataset, that are, the GRAFSEL approach which is a supervised approach based on a direct exploitation of the relations between document content extracted from keywords and document publication year to build a contrast graph and the CLUSTSEL approach that exploit a clustering process on the extracted document content and build up a contrast graph highlighting relation between cluster content with a further use of document publication years to highlight diachronic changes.

4 Experimental results

4.1 The GRAFSEL approach

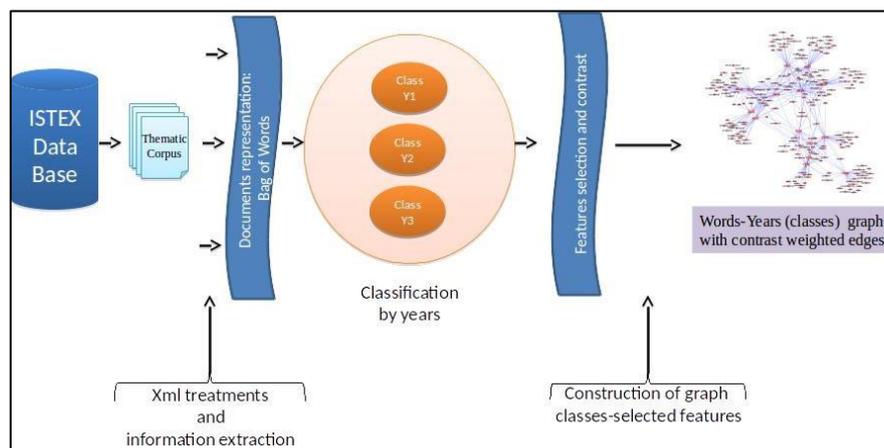


Fig. 1. Principle of the GRAFSEL approach.

Considering all the above-mentioned years, it is possible to detect invariant directions during each related period (figure 3). Furthermore, sudden changes of said directions suggest new scientific domains but also changes in professionals' practices. As we have formerly observed in figures 2, the direction corresponding to "nursing", "home", "care" appears in the year 2002-2003, and terms as "risk", "cancer", "mortality" and thereafter "exposure", "stress" emerge in the years 2006. Additionally, if the first years were solely marked by the term "women", the use of "people", "personality" in years 2003-2006 indicates that a humanization of care might appear. On its own side, the term "mice" is often used until 2001 and disappear after: it might figure out an indicator of changes in experimentation protocols.

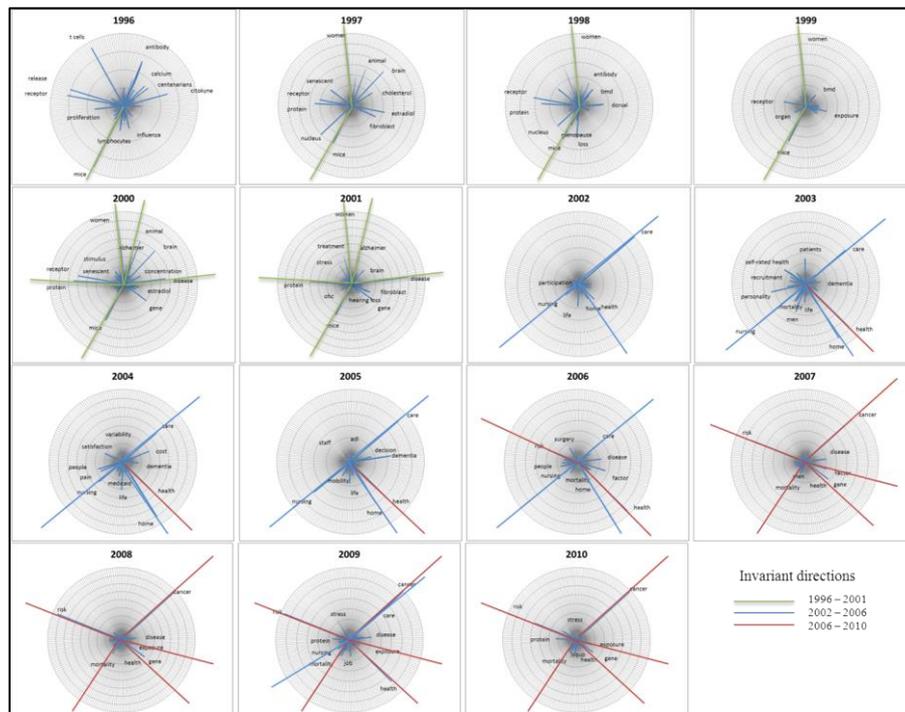


Fig. 3. Radar charts and invariant directions for all years

This short discussion shows that the use complementary modes of representation obviously enables a quick and simple view of the evolution of a thematic corpus through time.

4.2 The CLUSTSEL approach

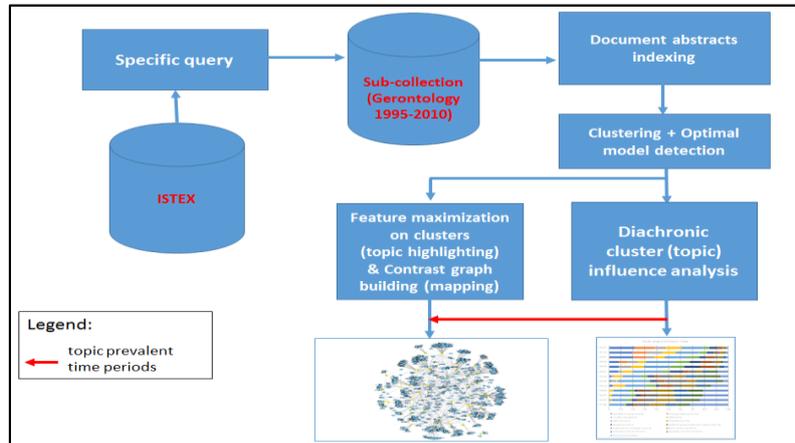


Fig. 4. Principle of the CLUSTSEL approach.

The overall principle of our unsupervised approach, we named CLUSTSEL is presented in figure 4:

1. The papers being represented their extracted keywords, we cluster their descriptions using a clustering algorithm. Several experiments are achieved by varying the number of expected clusters and consequently obtaining clustering models of various sizes.
2. A clustering quality measurement based on feature maximization is exploited to find out the optimal model among all the ones that have been generated.
3. The clusters of the optimal model being represented by the keywords extracted from their associated papers, we select keywords related to each cluster and compute the strength of the relations (i.e. the contrast) between selected keywords and clusters exploiting the feature maximization metric shortly described in section 3;
4. The graph highlighting the relationships between the clusters (i.e. the topics) and the selected keywords is built by weighting its links with the formerly obtained contrast values;
5. Papers publication years are used to find out dominant period of the topics as well as to build up a diachronic chart figuring out the comparative influence of topics during each year.

For clustering, we exploit 2 different usual clustering methods, namely k-means (MacQueen 1967), a winner-take-all method, and GNG (Fritske 1995), a winner-take-most method with Hebbian learning. The GNG method proved to be superior to k-means method because of (altogether) Hebbian, incremental and winner take-most learning process providing better independence to initial conditions and avoiding producing degenerated clustering results. Similar results have been already reported in

(Lamirel, Mall, et al. 2011). The selection of optimal model relies on feature maximization metrics presented in the former section. Our former experiments on reference datasets show that most of the usual quality estimators do not produce satisfactory results in a realistic data context, are sensitive to noise and perform poorly with high dimensional data (Kassab and Lamirel 2008). A more accurate method is thus to exploit feature maximization and more especially information related to the activity and passivity of selected features in clusters to define clustering quality indexes identifying an optimal partition. This kind of partition is expected to maximize the contrast described by eq. 7. The method is more precisely detailed in (Lamirel, Dugué, et al. 2016).

In the specific case of your experiment we propose to build up a contrast graph between a set of clusters representing the main research topics of the domain that have been extracted by the clustering process and the most contrasted features issued from the clusters' descriptions. This approach that combines clustering and contrast graph in an original way highlighting the most connected topics.

In the case of our experiment we focus on one type of external labels that are papers' publication years. Papers' publication years are exploited to perform a diachronic analysis of the topics' activity, highlighting the importance of each topic in each time period, either this activity is considered individually (see figure 7) or relatively to the other topics (see figure 6). As it is shown in the next part related to the analysis of the results, this approach helps to precisely understand the chronology of the research activity of a global research domain, like in our specific case.

In the context of our dataset we obtained an optimal model comprising 12 clusters (i.e. topics). The spatial distribution of 12 topics presented on the graph of figure 6 highlights clearly interpretable structure of the domain. Such graph provides generic although detailed representation of the domain-related research topics whilst highlighting the main relationships between the said topics whenever those topics appears as close neighbors on the graph. As an example, the 'Homecare' topic is directly related to logically connected topics like "Physical performance" and "Health condition". Similarly, "Menopause related problems", an early topic, appears to be accurately related to "Cancer studies" and "Gene senescence" that figure out more recent and more general research topics.

On its own side, diachronic representations that are presented in figure 6-7 can then be used to get a better understanding of the gerontology development from early research ("Menopause related problems", Hearing loss, "Age change" general studies) to more up to date research ("Home care", "Risks factors", "Physical performance", "Health condition", "Sociology of health") that fits well with the global changes regarding health politics. In that context research on "Physical performance" becomes the most prominent in the recent years and seems thus clearly represent a central focus because of its obvious influence on the other recent research areas.

Last but not least, research on "Neurodegenerative diseases" (Alzheimer, Parkinson, ...) seems to have split into two parts by generating a new specialized area related to "Memory performance".

In such a way, results provide by CLUSTSEL approach appears clearly complementary to the ones obtained by the GRAFSEL approach. Hence, the two methods provide similar results although they highlight those ones with different levels of generality.

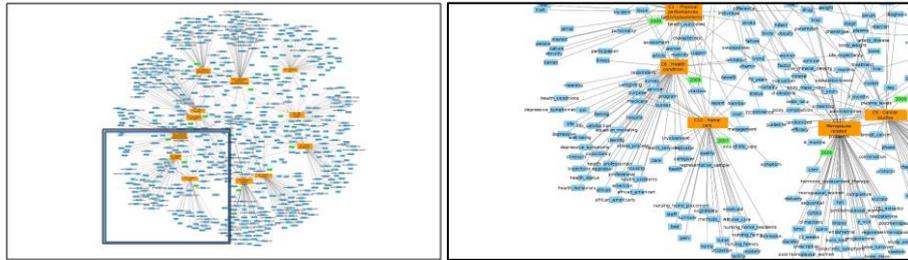


Fig.5 Global contrast graph based on optimal clustering (left) and focus (right) on the left-down part (blue square).

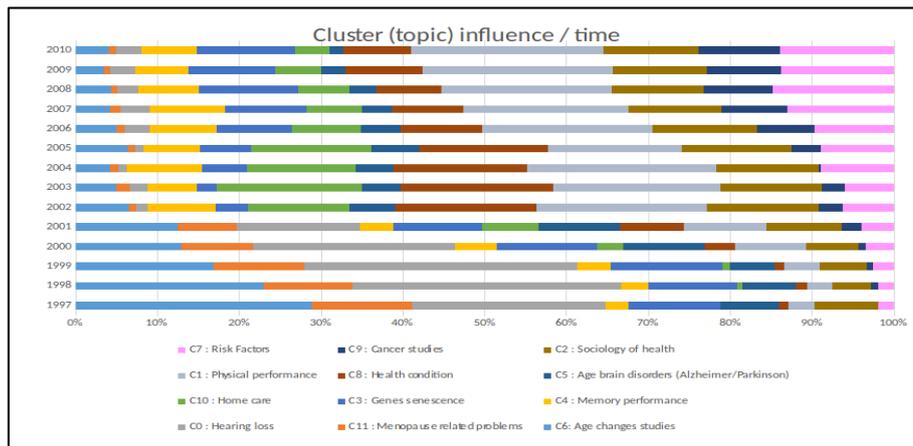


Fig. 6 Clusters influence by years.

5 Conclusion

We have presented an original overall methodology for the diachronic analysis of large and heterogeneous text collections based on feature maximization and associated contrast graphs. The originality of that approach comes from the fact that the nodes of the obtained graphs result from the combination of a feature selection processes and a classification or a clustering process, depending on the chosen option. Thus, one main advantage of the approach is to avoid cognitive overload in the current case of management of high dimensional data. Another of its main advantage is to be altogether parameter-free and knowledge/language-agnostic. Our first experimental results

obtained from the analysis of a realistic dataset extracted from the ISTEEX bibliographic database are promising. Hence, they prove to be easily interpretable by an expert of the analyzed domain. Moreover, the supervised and unsupervised options of our approach provide similar results that can be considered of different levels of generality.

One further and encouraging domain of investigation would concern to check the scalability of our approach to the context of massive data analysis.

6 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. (2013). Analysis of evolutions and interactions between science fields: the cooperation between feature selection and graph representation. 14th COLLNET Meeting, August 15-17, 2013 Tartu, Estonia
2. Dubey A., Ho Q., Williamson S. and Xing E. P. (2014), Dependent nonparametric trees for dynamic hierarchical clustering, NIPS 2014: 1152-1160.
3. Fritzke, B. (1995). A growing neural gas network learns topologies. In *Advances in neural information processing systems* (pp. 625–632).
4. Kassab, R., & Lamirel, J.-C. (2008). Feature-based cluster validation for high-dimensional data. In *Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications* (pp. 232–239). ACTA Press.
5. Lamirel, J.-C., Al Shehabi, S., Francois, C. & Hoffmann, M. (2004). New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping, *Scientometrics*, vol. 60, n° 3, 2004.
6. Lamirel, J.-C., Mall, R., Cuxac, P., & Safi, G. (2011). Variations to incremental growing neural gas algorithm based on label maximization. In *Neural Networks (IJCNN), The 2011 International Joint Conference on* (pp. 956–965). IEEE.
7. Lamirel, J.-C. (2012). A new approach for automatizing the analysis of research topics dynamics: application to optoelectronics research *Scientometrics* (2012) 93: 151-166 , October 01, 2012
8. Lamirel, J.-C., Cuxac, P., Chivukula, A. S., & Hajlaoui, K. (2015). Optimizing text classification through efficient feature selection based on quality metric. *Journal of Intelligent Information Systems*, 45(3), 379–396. doi:10.1007/s10844-014-0317-4
9. Lamirel, J.-C., Dugué, N., & Cuxac, P. (2016). New efficient clustering quality indexes. In *Neural Networks (IJCNN), 2016 International Joint Conference on* (pp. 3649–3657). IEEE.
10. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). Oakland, CA, USA.
11. Porter, A. L. and Rafols, I. (2009) Is science becoming more interdisciplinary? Measuring and mapping six research fields over time, *Scientometrics*, vol. 81, no 3, p. 719-745, 2009
12. Sayama, H. and Akaishi, J. (2012) Characterizing Interdisciplinarity of Researchers and Research Topics Using Web Search Engines, *Plos One*, vol. 7, no 6, p. e38747, 2012

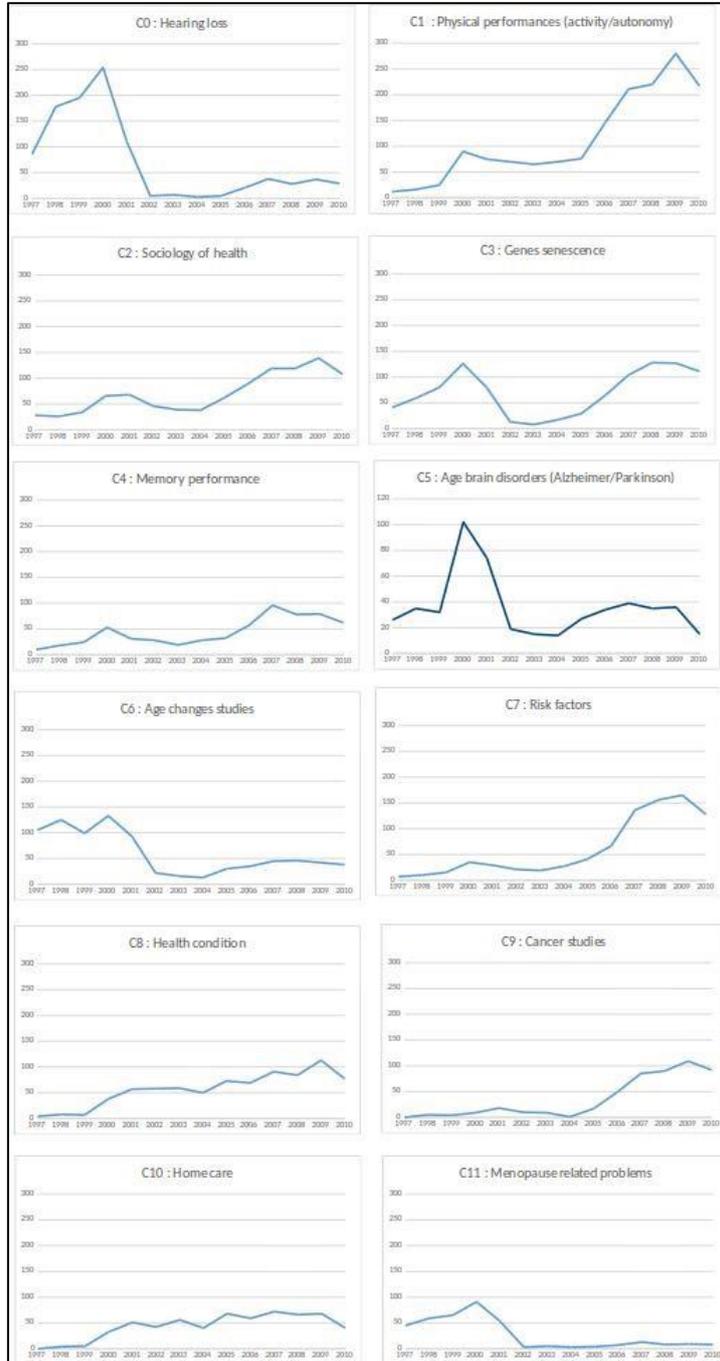


Fig. 7 Activity trends of the 12 topics.