

*An overview of the history of Science
of Science in China based on the use of
bibliographic and citation data: a new
method of analysis based on clustering with
feature maximization and contrast graphs*
**Jean-Charles Lamirel, Yue Chen, Pascal
Cuxac, Shadi Al Shehabi, Nicolas Dugué,
et al.**

Scientometrics

An International Journal for all
Quantitative Aspects of the Science of
Science, Communication in Science and
Science Policy

ISSN 0138-9130

Scientometrics

DOI 10.1007/s11192-020-03503-8



Your article is protected by copyright and all rights are held exclusively by Akadémiai Kiadó, Budapest, Hungary. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at link.springer.com".



An overview of the history of Science of Science in China based on the use of bibliographic and citation data: a new method of analysis based on clustering with feature maximization and contrast graphs

Jean-Charles Lamirel¹ · Yue Chen² · Pascal Cuxac³ · Shadi Al Shehabi⁴ · Nicolas Dugué⁵ · Zeyuan Liu²

Received: 8 November 2019
© Akadémiai Kiadó, Budapest, Hungary 2020

Abstract

In the first part of this paper, we shall discuss the historical context of Science of Science both in China and at world level. In the second part, we use the unsupervised combination of GNG clustering with feature maximization metrics and associated contrast graphs to present an analysis of the contents of selected academic journal papers in Science of Science in China and the construction of an overall map of the research topics' structure during the last 40 years. Furthermore, we highlight how the topics have evolved through analysis of publication dates and also use author information to clarify the topics' content. The results obtained have been reviewed and approved by 3 leading experts in this field and interestingly show that Chinese Science of Science has gradually become mature in the last 40 years, evolving from the general nature of the discipline itself to related disciplines and their potential interactions, from qualitative analysis to quantitative and visual analysis, and from general research on the social function of science to its more specific economic function and strategic function studies. Consequently, the proposed novel method can be used without supervision, parameters and help from any external knowledge to obtain very clear and precise insights about the development of a scientific domain. The output of the topic extraction part of the method (clustering + feature maximization) is finally compared with the output of the well-known LDA approach by experts in the domain which serves to highlight the very clear superiority of the proposed approach.

Keywords Science of Science · China · World · Topic tracking · Feature maximization · Unsupervised learning · Diachronic analysis

Professor Liu Zeyuan—author of the paper decessed on Februray 8th, 2020, during the review process.

Jean-Charles Lamirel
jean-charles.lamirel@loria.fr

Extended author information available on the last page of the article

Introduction

“Science of Science” refers to research into scientific and technological knowledge and explores the fundamental laws of the development of science and technology. It developed in Poland as early as the 1910s. The approach of Polish researchers who moved from metaphysics to empirical research and from the analysis of a single scientific discipline to the overall study of science laid the theoretical foundations for Science of Science as a specific field of study in Poland (Chen et al. 2017). However, the book of the communist-oriented English researcher (Bernal 1939) “The Social Function of Science” is generally recognized as the symbol of the real birth of the Science of Science. This book was directly influenced by the “Hessen episode” (Zhao and Jiang 1988) which had its deeper origins in Marxist ideology as Marx argued that “the essence of science is just its social function”. Similar ideas seem to be recurrent and more recently (Zhao and Jiang 1983) also argued that science and society are closely related and cannot be separated from each other.

As a global, interdisciplinary subject, the main objective of Science of Science is to consider all scientific and technological knowledge and activities as a research subject in order to explore the fundamental laws of the development of science and technology. Its field of research should therefore include historical, philosophical, sociological and economic research on science. However, the development of Science of Science on the international scene has not been so straightforward. Figure 1, taken from (He et al. 2017), illustrates the scientific development route of Bernal Prize winners. It is divided into three different research axes, “Scientometrics”, “Science, Technology and Society (STS)” and “Sociology of Scientific Knowledge (SSK)”. Derek J. de Solla Price inherited and developed Bernal’s scientific ideas and paradigms, deepening and broadening the theory and methods of Science of Science with an emphasis on data and quantitative analysis of science (Liu et al. 2013). The American scientific sociologist R. K. Merton examined the relationship between science, technology and society (STS) as an independent research subject while excluding the possibility of sociological research into the content of scientific knowledge (Pu and Di 1998). As a result, research on the sociology of science exploring the “social perspectives” and “cognitive perspectives” of Science of Science has been continuously differentiated during the development of the field (as illustrated, for example, by the creation of the SSK, which focuses on the fields of “anthropology” and “ethics”). In fact, Science of Science has gradually moved away from the original paradigm of Bernal’s scientific theory.¹

In the specific context of China, Bernal’s book (Bernal 1939) has been the subject of great interest since its publication. In particular, the part mentioning China and highlighting the limits of the development of modern science in that country quickly attracted the attention of leading Chinese scientists such as Zhu Kezhen (1890–1974, President of Zhejiang University), Wu Xuezhou (1902–1983, Director of the Chinese Chemistry Institution)

¹ It is worth noting that very recent scientific contributions on Science of Science seem to indicate that the Science of Science research domain is re-broadening its scope to come back to the original Bernal’s paradigm. Examples of this are the works by Zeng et al. (2017) published in Physics Reports and Science by the System Science Research Team of Beijing Normal University in China, Fortunato et al. (2018) which was a collaboration project between authors from the University of Indiana (USA) and Leiden University (The Netherlands) or even recent high quality papers published by the Complex Networks Research Team of the Northeastern University (USA) (Huang et al. 2012; Wang and Barabási 2013; Shen and Barabási 2014; Sinatra et al. 2017).

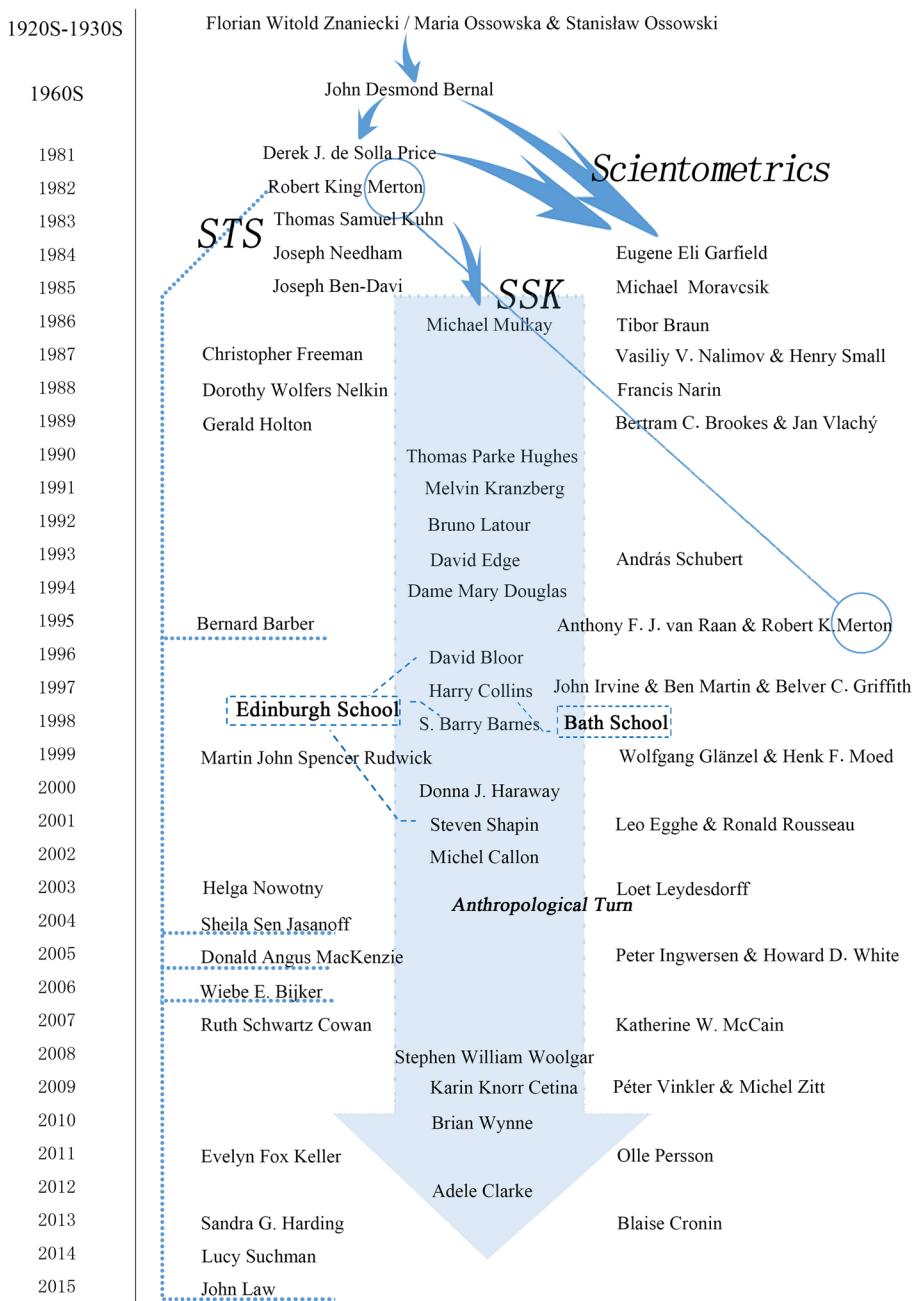


Fig. 1 The development route of Science of Science (He et al. 2017)

and Ren Hongjun (1886–1961, one of the founders of the China Science Society) (Qian and Li 2012) and the comments made there therefore quickly spread throughout China.

The formal birth of Science of Science in China, however, came from Tsien Hsueshen's initiative in the document entitled "Science and Technology" in 1977, which encouraged the creation of a new research space in China called "Science of Science" (Tsien 1979). Following on from Bernal's thinking, Tsien stressed that the field of Science of Science should be part of the social sciences (Tsien 1979, 1980). So far, three specific research institutes have focused their research on Science of Science namely those in Beijing, Tianjin and Shanghai. In addition, non-specialized institutes based in Beijing, such as CAST (China Association for Science and Technology), CASTED (The Chinese Academy of Science and Technology for Development of the Ministry of Science and Technology of the People's Republic of China), the Institute of Science and Development of the Chinese Academy of Sciences, the Chinese Institute of Engineering Development Strategies and many colleges and universities throughout China have also invested significant resources in basic and applied research into Science of Science.

In 2010, on the occasion of the 30th anniversary of the Chinese journal Science of Science and S&T Management, Liu Zeyuan highlighted the borders and main fields of Science of Science in China by mapping the literature in this field with the CiteSpace² tool (Liu 2017). Two major blocks of knowledge, corresponding to 2 complementary paths of development of Science of Science in China, were thus isolated—scientometrics, focusing on quantitative analysis, and scientific studies, focusing on philosophical analysis.

In this article, we exploit the research material of the latest 40 years in Science of Science in China and put in place a new method to understand and monitor both more clearly and more accurately the development in this field. Our objective is therefore to provide relevant indications on the origin of Chinese Science of Science, its structure and future directions through an original data analysis method, operating in a completely unsupervised manner, without any parameters and without external knowledge source.

The section "[Data collection and preprocessing](#)" of the paper presents the data collection and data preprocessing steps. The section "[Feature maximization as a global approach for data analysis](#)" focuses on the description of the feature maximization metric and of its associated feature selection process. This latter process is also illustrated with a simple example in that section. Section "[Data analysis process](#)" presents our experimental protocol and the principle of our original visualization method based on contrast graphs as well. Section "[Data analysis and visualization results](#)" is dedicated to the analysis of our experimental results by 3 experts³ of the domain. Section "[Comparison with LDA](#)" provides an additional comparison of the topic extraction capabilities of the method with the well-known LDA approach. Lastly, section "[Conclusions and discussion](#)" draws our conclusion and perspectives.

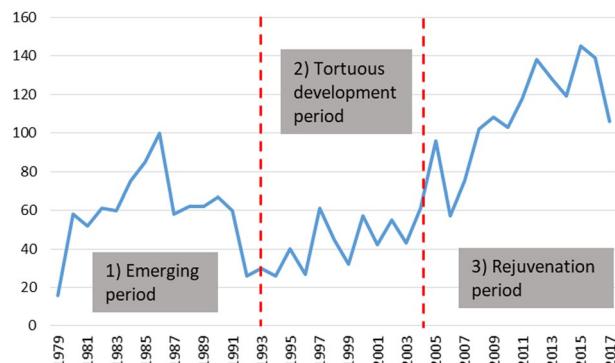
Data collection and preprocessing

Given the relatively blurred borders and vast scope of Science of Science (cf. section "[Introduction](#)"), it is not easy to make a complete and accurate extraction of literature relating to this domain. For this reason, in this paper we chose to focus on the evolution of the core content of the domain rather than trying to be exhaustive.

² <http://cluster.ischool.drexel.edu/~cchen/citespace/download/>.

³ See Acknowledgment section for more precise information on experts.

Fig. 2 Distribution of the number of papers per years in the experimental dataset and correspondence to historical periods for Science of Science in China



We queried the China National Knowledge Infrastructure (CNKI) database using “Science of Science” as the thematic term.⁴ We extracted 2401 articles belonging to journals referenced in the Chinese core journals list of Beijing University and in CSSCI⁵ (covering a research period until 2017-10-22). Data cleaning was carried out in a second phase to remove items that did not correspond to research documents (e.g. meeting notices, journal presentations or editorials). After this phase we retained 1334 journal articles. We used these to then retrieve 2677 cited articles (after removing duplicates), 1539 of which were published in specialized journals. We added these 1539 documents to our 1334 core documents to form our experimental data set of a total of 2873 research articles.

We then first looked for an indirect way to validate our data collection process. To do this, we tracked the distribution of articles according to their publication dates. The trend we observed (Fig. 2) was found to comply perfectly with Liu’s recent observations based on experience of the domain (Liu 2017) showing that Science of Science research in China has gone through three stages, namely a period of rapid growth (1977–1991) (1), a period of rather difficult development (1992–2003) (2) and a period of rejuvenation (2004–2017) (3).

In a second phase, the titles, abstracts and keywords of the 2790 articles were extracted.⁶

The indexation process was quite complex. It started with an initial dictionary of 9679 keywords gathered from the keyword field of the 2790 articles. We used NLPiR-ICTCLAS,⁷ a specific toolbox for Chinese language processing, for word segmentation and tagging of titles and article summaries. Due to the particularities of Science of Science, software cannot accurately segment certain domain terms expressed by complex multi-character words, such as “科学学”: “Science of Science”, “科学学研究”: “Science Research”, “科学逻辑学”: “Science Logistics”, “科学的社会功能”: “Science Social Function”. We therefore carried out a later reconstruction of these words.

Among the extracted words, we then used an ad-hoc Python programme to filter the elements labelled as nouns and deleted quantities (numbers, dates, etc.). Then, we carried out a second phase of term cleaning to remove empty and meaningless terms or those covering

⁴ Because the CNKI database was queried in Chinese, we use two different terms because “Science of Science” is described by two different terms in this language: “科学学” and “科学的科学”.

⁵ Chinese Social Sciences Citation Index.

⁶ For articles prior to 1997 that did not contain a summary or keywords we only used the information in the title.

⁷ <http://ictclas.nlpir.org/>.

Table 1 Summary of the lexicon processing steps

	1- Term merging	2- Void terms suppressing	3- Frequency thresholding (>5)
Initial vocabulary size	11931	11696	11571
Suppressed words	--	125	9995
Merged words	235	--	--
Resulting vocabulary size	11696	11571	1576

the whole context of the dataset (e.g. “search”, “analysis”, “year”) and also to merge terms with similar meanings (e.g. “著者分布” and “作者分布”: distribution of authors, “作者合作网” and “作者合作网络”: network of co-authors, “知识图谱” and “知识图谱分析”: knowledge mapping). Once these words were merged with the initial keyword dictionary, this resulted in a dictionary of 13,442 Chinese terms.

The dictionary of terms was then translated into English. Due to a poorer vocabulary in English than in Chinese, the translation was likely to generate new equivalent words (such as “知识地图”: “knowledge geography” and “可视化底图”: “basic visualization map”, “科技评估”: “S&T evaluation” and “科研评价”: “research evaluation”) that needed to be merged again. After this process, we obtained a dictionary of 11,931 English terms. Category labels (resp. “name”, “city”, “country”) were finally attached to the terms representing the corresponding entities (resp. person, place and country).

To eliminate the remaining noise we applied an additional cleaning pass detailed in Table 1. Firstly, we merged the remaining equivalent words into a single entry (for example, an author may appear with or without his or her first name such as “Merton” and “R.K. Merton”—an institution may appear with its acronym or in an expanded form such as “NSF” and “National Science Foundation”). Secondly, we deleted words or expressions whose meaning was unclear in English and corrected some translation errors. This last process led to us deleting 360 entries in the glossary (235 merged entries and 125 deleted entries). A frequency threshold of ⁸ was finally applied to remove low frequency words. The result was a final dictionary of 1576 terms which were used to re-index the articles.

Feature maximization as a global approach for data analysis

Most of our further data analysis of the Science of Science dataset is based on a feature selection approach relying itself on feature maximization metric (Lamirel et al. 2011). We shall first present this important metric before presenting the whole data analysis process. Feature maximization is an unbiased metric which can be used to estimate the quality of a classification whether it is supervised or not. In unsupervised classification (i.e. clustering), this measure exploits the properties (i.e. the features) of clusters’ associated data for different purposes (clustering labelling and cluster content highlighting, overall display

⁸ The frequency threshold of 6 was found empirically. It means the description space can be significantly reduced while allowing for accurate clustering (the quality of which was estimated both by the experts and by our quality measures presented in section “[Clustering and optimal model detection](#)”). No documents were deleted by this process.

of clustering results like on the contrast graph presented in this paper, optimal clustering model detection). Its main advantages are that it is free of parameters, totally independent of the clustering method and its operating mode, it works suitably in high dimensional spaces and represents a better compromise between discrimination and generalization than usual metrics (Euclidean, Cosine or Chi square, etc.).

Feature F-measure

Let us consider a partition C which results from a clustering method⁹ 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 the Feature Predominance $FP_c(f)$ which are themselves defined as follows:

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

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

avec

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

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 to the cluster c . Feature Predominance measures the ability of f to describe cluster c . In a complementary way, Feature Recall means f can be characterized 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 throw experiments (Lamirel et al. 2015) that the F-measure which is a combination of these two measures is only lightly influenced by feature scaling. Nevertheless, data must be standardized to guarantee full scale independent behaviour for this measure. Furthermore, the choice of the weighting scheme for data is not really limited by the approach but it is necessary to deal with positive values. The scheme therefore needs to be capable of figuring out the semantic significance and importance of the feature for the data.¹⁰

⁹ In this article, the features represent the words extracted from the title, abstract and keywords of the articles, the weights of the features are the adjusted frequency information associated with them and the unsupervised classification (clustering) is based on the GNG algorithm.

¹⁰ A feature with negative values can be separated into 2 different positive sub-features without loss of information. The first represents the positive part of the original feature and the second its negative part.

Feature maximization

In a supervised context, feature maximization measurement is capable of generating a powerful feature selection process. 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 parameter-free process that uses the capacity of Feature F-measure to discriminate between clusters ($FR_c(f)$ index) and also its ability to faithfully represent cluster data ($FP_c(f)$ index). The set S_c of features that are characteristic of a given cluster c belonging to a partition C is defined as:

$$S_c = \left\{ f \in F_c \mid FF_c(f) > \overline{FF}(f) \text{ and } FF_c(f) > \overline{FF}_D \right\} \quad (4)$$

with

$$\overline{FF}(f) = \sum_{c' \in C} \frac{FF_{c'}(f)}{|C_{/f}|} \text{ and } \overline{FF}_D = \sum_{f \in F} \frac{\overline{FF}(f)}{|F|} \quad (5)$$

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:

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

In other words, the features judged relevant for a given cluster are those whose representations are (1) better in this cluster than their average representation in all the clusters and (2) 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. This latter operation corresponds to a feature selection process.

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. The 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.

As already mentioned, in clustering the active features in a cluster are selected features for which the contrast is greater than 1 in that cluster. Conversely, the passive features in a cluster are selected features present in the cluster's data for which the contrast is less than unity. As regards the principle of the method, this type of selected features inevitably has a contrast greater than 1 in one or several cluster(s) (see Eq. 7 for details). A simple way to exploit the features obtained is to use active selected features and their

Fig. 3 Principle of feature *F*-measure computation for sample data

Shoes Size	Hair Length	Nose Size	Class
9	5	5	M
9	10	5	M
9	20	6	M
5	15	5	W
6	25	6	W
5	25	5	W

$$FR(S,M) = \frac{27}{43} = 0.65$$

$$FP(S,M) = \frac{27}{78} = 0.35$$

$$FF(S,M) = \frac{2(FR(S,M) \times FP(S,M))}{FR(S,M) + FP(S,M)} = 0.45$$

Fig. 4 Principle of computation of the overall feature F-measure average and elimination of irrelevant features—here ‘Nose Size’

	$F(x,M)$	$F(x,F)$	$\overline{F(x,.)}$	$\overline{F(.,.)}$
Hair Length	0.39	0.66	0.53	
Shoes Size	0.45	0.22	0.34	
Nose Size	0,3	0,24	0,27	0.38

associated contrast for cluster labelling as we proposed in (Lamirel et al. 2015). We also used this idea further in the experimental context discussed in this paper.

Illustrative example

Below we give an example of the operating mode of the method on the basis of a toy-data-set encompassing two classes¹¹ (*Men (M)*, *Women (F)*) described with 3 features: *Nose_Size*, *Hair_Length*, *Shoe_Size*. Figure 3 shows the source data and how the F-measure calculation of the *Shoe_Size* feature operates in the *Men* class.

As shown in Fig. 4, the second step consists of calculating the average F-measure of each feature over the classes, and the overall average F-measure for the combination of all features and all classes. In this figure, notation $\overline{F(.,.)}$ stands for the overall average \overline{FF}_D presented in (Eq. 5) and notation $\overline{F(x,.)}$ stands for average of class x which is itself computed as:

$$\overline{F(x,.)} = \sum_{f \in S_x} \frac{FF_x(f)}{|S_x|}$$

Features with F-measures that are systematically lower than the overall average are eliminated and thus the *Nose_Size* feature is removed. The remaining features (i.e. selected features) are considered active in the classes in which their F-measure is above the marginal average:

¹¹ Behaviour of our measures is similar with classes or clusters.

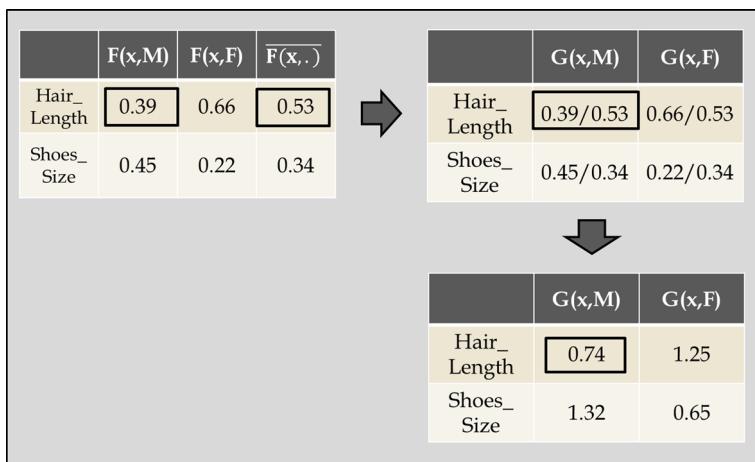


Fig. 5 The contrast computation principle for selected features

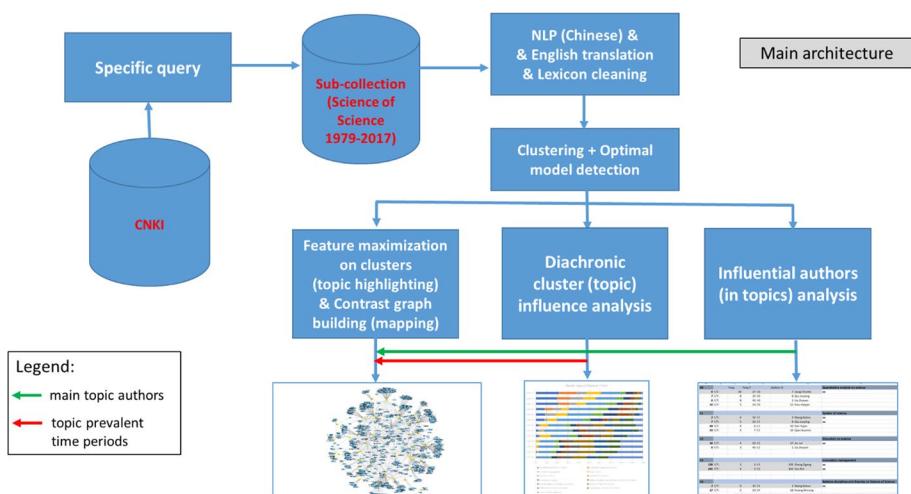


Fig. 6 Overall data analysis process

1. *Shoes_Size* is active in the *Men's* class,
2. *Hair_Length* is active in the *Women's* class.

The contrast ratio highlights the degree of activity and passivity of selected features as regards their F-measure marginal average in different classes. Figure 5 illustrates how the contrast is calculated for the example presented.

In the context of this example, the contrast may be considered to be a function that will have the following effects in virtual terms:

1. Increase the length of women's hair,

-
2. Increase the size of the men's shoes,
 3. Decrease the length of the men's hair,
 4. Reduce the size of women's shoes.

Data analysis process

The overall architecture of our experimental process is presented in Fig. 6. After preprocessing steps, the process uses clustering in combination with feature maximization to extract the main topics of research from the Science of Science dataset under study. We will show later on in this paper that the combination of a suitable clustering approach - like neural clustering based on growing gas (Fritzke 1995)—with feature maximization offers superior performances to alternative approaches for topic extraction like LDA (Blei et al. 2003). This is only the case as long as an optimal clustering model (i.e. a suitable number of clusters) can be properly identified from the analyzed data. We thus propose to exploit one of our recent and efficient approaches also based on feature maximization for the optimal model detection task (Lamirel et al. 2016). Processing the clustering results with a graph approach based on contrast is an original method presented in this paper. It enables the cognitive overload resulting from the representation of interactions in large datasets to be reduced and the dependencies between extracted topics through shared features with high contrast to be correctly calculated. The last part of our approach exploits external labels of data associated to clusters. Firstly, publication dates are used to perform a diachronic analysis of the activity of clusters (i.e. topics) and secondly, author information is used to highlight the most influential authors in the different topics. Dates and author information are also reported on the contrast graph. Details of the approach's different steps are given in the next sections.

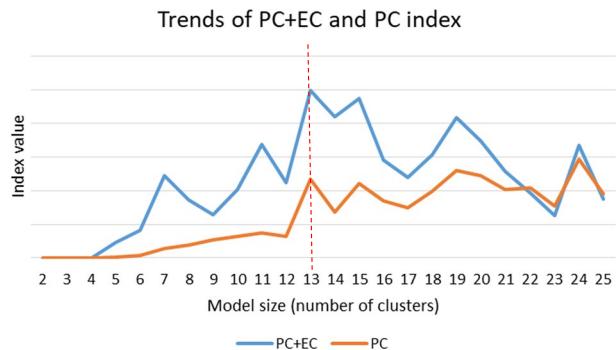
Clustering and optimal model detection

We use 2 different well-known clustering methods, namely k -means (MacQueen 1967), a winner-takes-all method, and GNG (Fritzke 1995), a winner-takes-most method with Hebbian learning. We have reported on the method which produced the best results in the following experiments. In all cases, the GNG method proved to be superior to the k -means method because of it is a Hebbian, incremental and winner take-most learning process which provides better independence from initial conditions and outliers and avoids producing degenerated clustering results. These kinds of results have also been observed in many of our former experiments (Lamirel et al. 2011).

The selection of the 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 estimators¹² do not produce satisfactory results in a realistic data context. They were also found to be sensitive to noise and to 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

¹² Like the Dunn index (Dunn 1974), the Davies-Bouldin index (Davies et Bouldin 1979), the Silhouette index (Rousseeuw 1987), the Caliński-Harabasz index (1974) or the Xie-Beni index (1991).

Fig. 7 Clustering quality evaluation (trends of PC and PC+EC indexes) and optimal model (13 clusters) highlighting. Index values have been rescaled for better visualization



optimal partition. This kind of partition is expected to maximize the contrast described by Eq. 7. Indeed, the more contrasted the clusters' features are, the more these clusters are found to be compact or separated. Hence, this approach leads to the definition of three different quality indexes: PC (Positive Contrast), EC (Extended Contrast), and CB (ComBined contrast).

Below there is an example the expression of the PC and EC index. The CB index represents a weighted combination of the other two. A more precise description of this approach as well as experiments comparing our indexes with other indexes on real-life data ranging from simple to complex data can be found in the reference (Lamirel et al. 2016).

The PC index's principle corresponds by analogy to that of intra-cluster inertia in the usual models. It is a macro-measure based on the maximization of the average weighted contrast of active features for optimal partition. For a partition comprising k clusters, it can be expressed as:

$$PC = \arg \max_k \left(\frac{1}{k} \sum_{i=1}^k \frac{1}{|s_i|} \sum_{f \in S_i} G_i(f) \right) \quad (8)$$

The EC index's principle corresponds by analogy to that of the combination between intra-cluster inertia and inter-cluster inertia in the usual models. This index is based on the maximization of the average weighted compromise between the contrast of active features and the inverted contrast of passive features for optimal partition:

$$EC = \arg \max_k \left[\frac{1}{k} \sum_{i=1}^k \left(\frac{|s_i| \sum_{f \in S_i} G_i(f) + |\bar{s}_i| \sum_{h \in \bar{s}_i} \frac{1}{G_i(h)}}{|s_i| + |\bar{s}_i|} \right) \right] \quad (9)$$

where n_i is the number of data associated with the cluster i , $|s_i|$ represents the number of active features in i , and $|\bar{s}_i|$, the number of passive features in the same cluster.

In our experiment, we vary the number of clusters in a range of up to 1/50 of the number of data. We reject size 1 models for two main reasons—firstly the quality indexes are not intended to produce results in this case and secondly these models correspond to a trivial clustering operation.

We carry out crisp clustering. In this case each data is re-assigned to a single cluster after the clustering process and the usual form of the assignment function q_f which associates a data d to a given cluster is as follows:

C. 9# : Knowledge mapping on science	
5.376770 theme,	3.873852 international,
5.030978 research hot topics,	3.801943 expectation,
4.827424 literature,	3.778949 data,
4.734794 software,	3.721473 knowledge map,
4.697236 frontier,	3.648744 visualization analysis,
4.595268 development trend,	3.641972 tool,
4.401170 research topic,	3.557082 research situation,
4.342141 hotspot,	3.495185 trend,
4.159228 both at home and abroad,	3.411639 representative figure,
3.989917 science knowledge mapping,	3.327669 research direction,

Fig. 8 Example of a description of a cluster through the list of its most contrasted features (here terms). The cluster's related topic is knowledge mapping

$$af(d) = \arg \min_k \left(Dist\left(\vec{k}, \vec{d}\right) \right) \quad (10)$$

where $Dist$ represents a distance function (generally Euclidean distance is used), \vec{k} represents the description vector of cluster k and \vec{d} represents the description vector of document d .

By using the clusters' associated data, we select the model that optimizes the PC + EC combination of the above-mentioned indexes. This technique makes it possible to obtain the relevant number of clusters highlighting the main science research topics during the period under study. Expert analysis of the obtained results confirms that the clustering model chosen as optimal using our approach consistently accurately represents all the main research topics in the Science of Science field. Figure 7 presents trends in the evolution of the PC and EC indices and the optimal point (i.e. the optimal number of clusters or the optimal model) found for the model with 13 clusters. Figure 8 presents a description of a cluster based on its most contrasted characteristics and Table 2 presents the list of cluster titles that the expert characterized by exploiting the most contrasted elements.

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 and the set of features S is a subset of an overall feature set F (i.e. the original feature space on which a dataset's data relies) obtained through a feature selection process, like feature maximization as presented above. In the case of feature

Table 2 List and summary description of the clusters obtained

	Label (expert)	Content summary (most contrasted important terms)
Cluster 0#	Quantitative analysis on science	Bibliometrics, citation analysis, journal, indicator, quantity, impact factor, statistics analysis, data, SNA
Cluster 1#	Research evaluation	Efficiency, systems engineering, decision making, forecast, evaluation, administration, input and output, efficiency, sustainable development
Cluster 2#	Education on science and talent cultivation	Higher education, Ministry of Education, planning, talent cultivation, university
Cluster 3#	Innovation management	Enterprise, knowledge management, collaborative innovation, performance, competitive advantage, technological innovation, integration
Cluster 4#	Domain structure and peripheral disciplines on Science of Science	S&T studies, theory of science, technology theory, technology philosophy, dialectics of nature, library science, knowledge-based economy, history of science, discipline structure
Cluster 5#	Philosophical foundation on Science of Science	Philosophy, Marxist doctrine, reality, criticism, ontology, dialectics, human society, materialism, humanism
Cluster 6#	Discipline system	Definition, connotation, discipline system, research method, concept, principle, comparative research, system science, safety, safety principle, safety system
Cluster 7#	Research policy and impacts on society	Scientificification, S&T development, modern management, productivity, nation, world, emancipation of mind, socialism, social economic development
Cluster 8#	Subject attributes on Science of Science	Natural science, social science, modern science, regular pattern, development principle, edge, interdisciplinary research
Cluster 9#	Knowledge mapping on science	Research hot topics, software, hotspot, theme, frontier, development trend, knowledge map, data, visualization analysis
Cluster 10#	History on Science of Science	History of science, creator, JD.Bernal, Price, big science, Zhao Hongzhou, scientometrics, Soviet Union, world science, sociology of science
Cluster 11#	Publication on Science of Science	Journal, publication, S&T management, S&T system reform, S&T circle, editorial office, institute, S&T policy
Cluster 12#	Organization on Science of Science	Committee, leadership, Chinese Association for Science of Science, conference, symposium, academic exchange, Liu Zeyuan

For better clarity, cluster labels (i.e. key phrases) are added by the expert in the domain

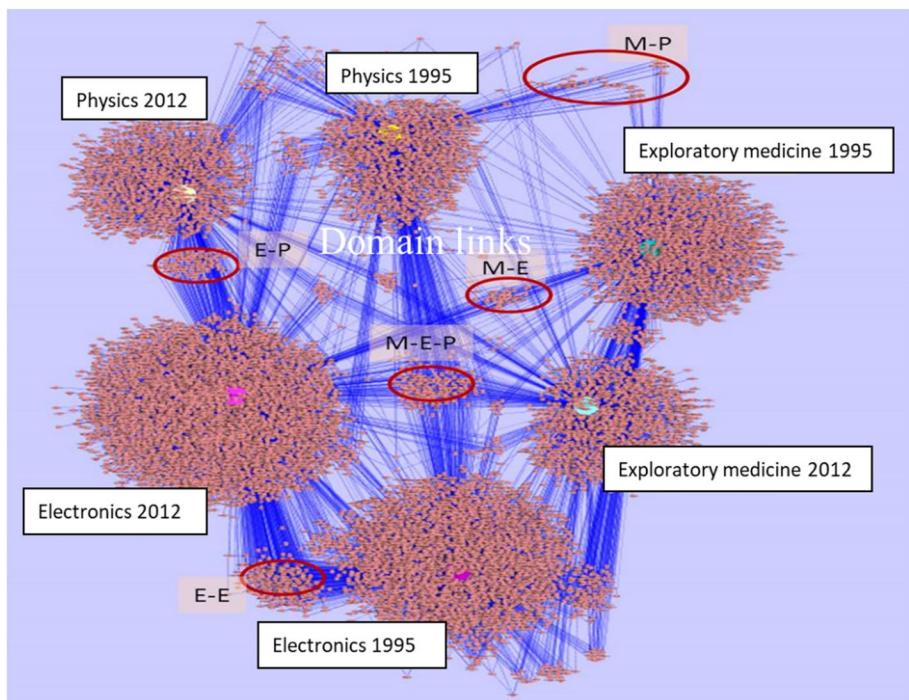


Fig. 9 An example of a contrast graph materializing the relationships between authors and scientific domains associated with time periods. The red circles highlight authors who are knowledge transmitters between domains and periods

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 and is defined by Eq. 7.¹³

These kinds of graphs have many interesting properties. Firstly, they reduce the cognitive overload produced with classical graphs' representation because of the associated feature selection process which reduces the number of potential connections. Secondly, they can be used to indirectly highlight relationships between labels whenever features have contrasted interaction with several labels. Thirdly, the combination of this approach with weighted force-directed model (Kobourov 2012) for graph representation highlights the central or most influential labels of the L set and makes it easy to identify the most densely connected labels through associated features with these latter appearing close together on the graph.

We 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 (2013). Figure 9 shows a resulting contrast graph where features represent the authors of research papers and labels represent a combination of time period and research domains. The authors who connect domains and time periods clearly appear on the

¹³ In Eq. 7, labels represent categories or clusters to which data are associated.

graph represented in the figure. They can be considered as knowledge transmitters which consequently clearly highlights their major social and scientific role.

Exploitation of complementary information through external labels

As (Attik et al. 2006) pointed out, external labels are information which is associated with data but does not play any role in the initial data analysis process. However, this information could include important clues to help enhance the precision of the analysis. In the case of the clustering process presented above, external labels can be exploited in a secondary step (i.e. after the clustering process) by evaluating their posterior distribution into clusters through clusters' associated data to provide complementary information about the latter data or related topics.

In the case of our Science of Science dataset we focus on two kinds of external labels namely papers' publication dates and papers' authors. Papers' publication dates are processed to carry out a diachronic analysis of topic activities highlighting the importance or activity of each topic in each time period with an activity considered either individually or relatively to the other topics. As is shown in the next section related to the analysis of the results, this approach helps to precisely understand the chronology of the research activity in an overall research field, like the Science of Science in our specific case. Information on papers' authors can be processed to highlight the most important contributors who drive or influence a main area of research (i.e. a topic). Such authors may even be considered to be central contributors if they led, influenced or coordinated several areas of research at the same time.¹⁴

In the context of our experiment, our external label analysis is based on two different measures—label frequency and label prevalence. Label frequency F_c^l of a label l of a type t in a cluster c can be defined as:

$$F_c^l = \text{Card}\{d \in D | af(d) = c \wedge l \in \text{Extlab}_t(d)\} \quad (11)$$

where Card is the set cardinal function, D is the whole set of exploited data, af the function defined at Eq. 10 (that provides the cluster associated to data d) and $\text{Extlab}_t(d)$ a function that provides the list of external labels of type t associated to data d .

Label prevalence is a cluster-based measure. A label l is prevalent in a cluster c if:

$$\nexists c' \in C, c' \neq c, F_{c'}^l > F_c^l \wedge \nexists l' \in L_c, l' \neq l, F_c^{l'} > F_c^l \quad (12)$$

where L_c is the set of labels occurring in clusters c through its associated data.

Prevalence is used to highlight the prior influence of a label. The consequences of this definition are that a label can be solely prevalent in one unique cluster and some clusters might not have any prevalent labels.

¹⁴ In many cases, each data can have several external labels of the same kind. For example, a research paper can have several different authors.

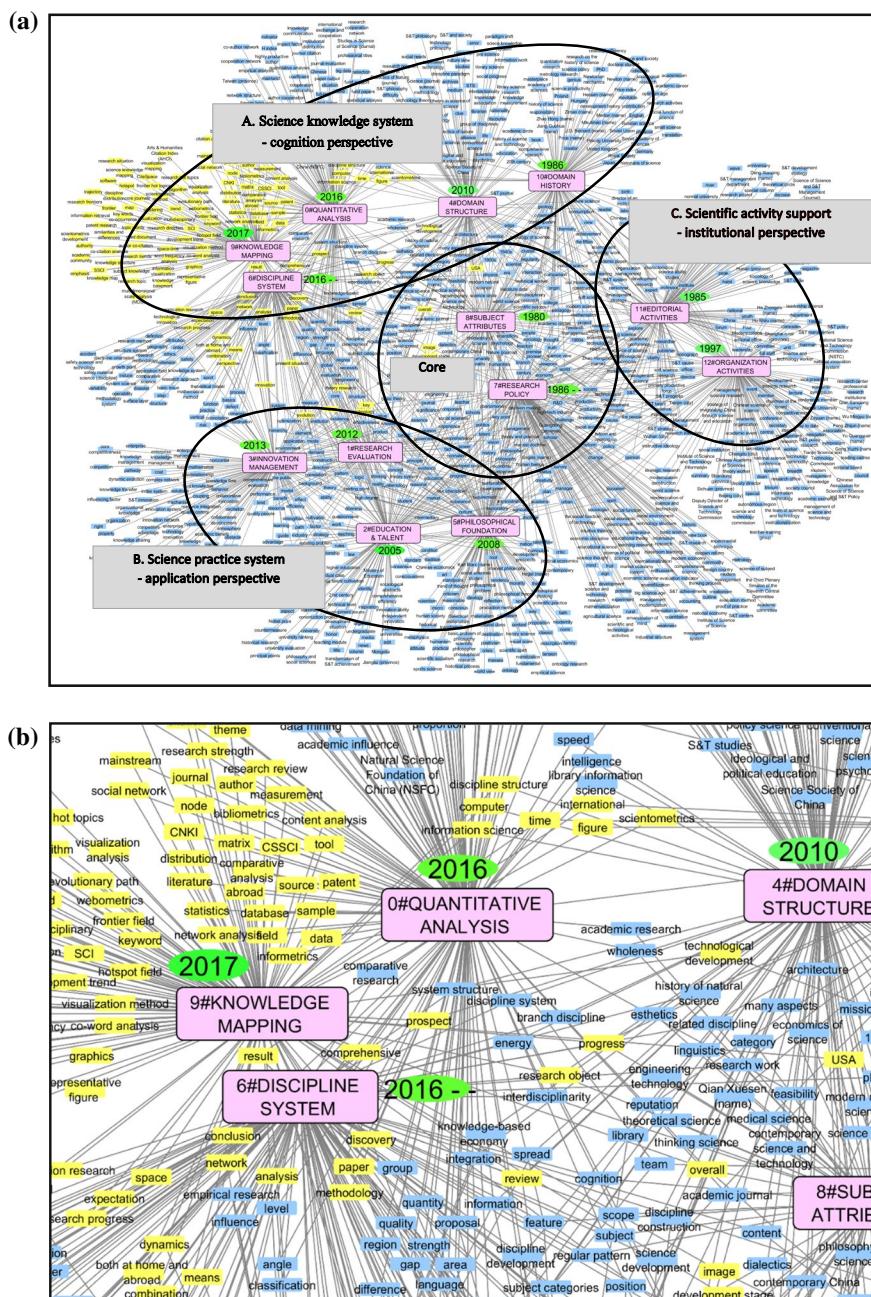


Fig. 10 **a** Global contrast graph representing main topics and domain structure in Science of Science in China. (Cluster 9# is highlighted, and the detailed information on clusters is shown in Table 2). **b:** Zoom presenting a specific area of the global graph of this figure

Data analysis and visualization results

General topic structure of the Science of Science domain

In the specific case of our experiment on Science of Science data, we propose to build a contrast graph between a set of clusters (set L) representing the main research topics in a given field extracted by the clustering process and the most contrasted features (set S) issued from the cluster descriptions (see Fig. 8). This approach combines clustering and contrast graphs in an original way and is particularly useful for Science experts¹⁵ to understand the construction of their domain, highlighting the most central topics in the domain (domain generators) as well as those which are the most connected. In the resulting graph, only the edges with a contrast greater than 1.4 are retained for the representation (1074 of the 1576 terms used for clustering are kept). Figure 7 represents the entire graph and Fig. 8 represents a sub-section highlighting closely related domains that share many characteristics.

The spatial distribution of the 13 topics is shown in Fig. 10 (a zoomed extract is also presented in Fig. 10b). According to all of our 3 experts, this graph highlights a very clearly interpretable structure of the Science of Science field in China. In such a model, highly interconnected topics will tend to appear at the centre of the representation (see section "Contrast graphs"). In our case, this information on the core domains is represented by two complementary topics—"8# Subject attributes of the Science of Science domain" and "7# Research policy and impacts on society".

The explanation given by the experts is as follows:

1. Science of Science is a reflective field and must therefore study its own development along with the models of science themselves (topic 8#).
2. Science of Science is an applied science that guides practice. Therefore, it must be practice-oriented, study the social function of science and serve research policy (topic 7#).

Three main fields of research appear around the basic information corresponding to usual scientific activities namely "A. Scientific knowledge system", "B. System of Practice of Science" and "C. Support of scientific activity". These form the complete logical structure of Science of Science from the points of view of cognition, application and institutional structures respectively.

The field "A. Scientific Knowledge System" is associated with the five related topics "10# History on Science of Science", "4# Domain structure and peripheral disciplines on Science of Science", "6# Discipline System", "0# Quantitative Analysis on science" and "9# Mapping of knowledge on science". The historical evidence (10#) served as a starting point for the first studies of the structure of Science of Science and peripheral disciplines and for the further construction of the field (4#). Finally, this evidence helped develop quantitative approaches to scientific research (0#). The recent emergence of new data management techniques and software technologies has led to the development of elaborate visualization approaches (9#). Although the discipline system field (6#) is obviously related to the structure of the field (4#), it is a more independent research field related to the study of research methodology, comparative approaches in research and security sciences.

¹⁵ In this case the opinion of our domain experts themselves (see "Acknowledgment" section for expert descriptions).

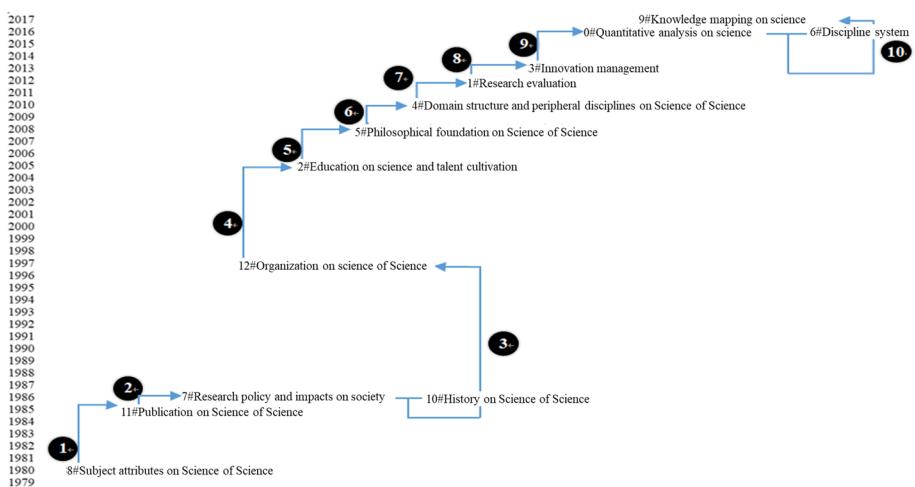


Fig. 11 The pathmap of topic change in Chinese Science of Science

The field “B. System of scientific practice” is composed of four main related topics: “1# Research evaluation”, “3# Innovation management”, “2# Educational sciences and talent culture” and “5# Philosophical foundations of the Science of Science domain”. Today, innovation strategy is one of the Chinese government’s major concerns because it aims to link science and technology to the economy and value-added applications for the development of Chinese society. The role of science and technology in social practice is therefore logically materialized by innovation management (3#) and this field has indeed developed considerably in recent years (see also Fig. 11). Research evaluation (1#) (analysis of scientific inputs and outputs, scientific decision-making) and encouraging education and talent (2#) play complementary roles in innovation strategy. As a result, these areas have also become essential issues in China’s scientific practice system. Finally, the entire system of scientific practice is clearly guided by the philosophical foundations (5#) inherited from Marx’s philosophy and Engels’ dialectic of nature.

The field “C. Support of the system of scientific activity” is composed of two main themes—“#11 Publications on Science of Science and “#12: Organization on Science of Science”. This field is clearly linked to the management of the production of scientific research (#11: publications, research periodicals), as well as to the organization of activities in the field (#12: learned societies, conferences and colloquia). These tasks support the successful development of the domain and ensure its sustainability.

The evolution of Science of Science

In the last 40 years, 13 research themes have been observed in Chinese Science of Science. The way they evolved, as materialized using the publication dates of the articles analyzed (see section "Exploitation of complementary information through external labels"), is also very clearly highlighted by our method and presented in Fig. 11. It is additionally justified by the analysis of the experts as described below.

In the 1980s, activity in Science of Science was only just beginning in China. The most widely discussed topic in the academic world was the issue of domain-specific attributes

(8#). At that time, researchers tried to identify the nature and general patterns of the field combining Bernal's thinking with the actual Chinese context (7#). When the National Science Conference was held in 1978, China's science and technology system began to enter a period of reform and three major journals in the field of Science of Science (11#) were successively created—Science Research Management (1978), Science of Science and S&T Management (1980) and Studies in Science of Science (1983). These journals were supported by the government in their early stages and quickly attracted a large number of articles which meant the publication and management of research results became an important topic.

At the same time, academics have studied the history of Science of Science (10#) to find evidence of the theoretical basis and construction of the discipline in China by using the work of the main foreign actors in the field. This is how they laid the theoretical foundations of research policy in China.

Then, Science of Science in China developed more slowly for a period until the third council of the Chinese Association for Science of Science and Science and Technology Policy was established in 1997 (12#). The doctoral programme launched in the same year at Dalian University of Technology helped improve the discipline's institutional system and therefore science teaching methods and promoting talent became topical themes in 2005 (2#).

Philosophical ideas play a key role in the educational system to guide practice and this point appeared more important in promoting and supporting professional talent in science (5#). Thus, in 2008, Chinese academic circles became more interested in the philosophical origins of Science of Science, and more particularly in the foundations of Marx's philosophy.

The early development of the method of scientific knowledge mapping in China (Chen and Liu 2005) paved the way for a new field of research aimed at obtaining information on the structure of the fields and peripheral disciplines of Science of Science (4#) such as scientific and technological studies, technological theory, technological philosophy, library science or knowledge economy.

In 2012, Chinese academics stated that scientific activities are a system in themselves. Therefore, to increase effectiveness, these activities must be evaluated and planned using approaches involving scientific and technical systems (1#). There are two main reasons for this change in context. The acceleration of decision-making in science and technology required the rapid and objective assessment of research input and outputs while many new approaches were developed in scientometrics which were to successfully complement traditional assessment methods.

In Yang Xiaolin's book, he relates the words of Wu Mingyu, one of the pioneers in the field of Science of Science in China: "People who engage in science of science should first emphasize the concept of innovation".¹⁶ In the same vein, shortly after its creation in 1992, the NSFC (National Natural Science Foundation of China) began to support research on innovation. As awareness of the central role of innovation in China increases, research themes related to innovation are rapidly expanding ("technological innovation", "global

¹⁶ Yang Xiaolin. Thirty years of science and Technology Policy Research—Wu Mingyu's oral autobiography [M]. Hunan Education Press 2015.

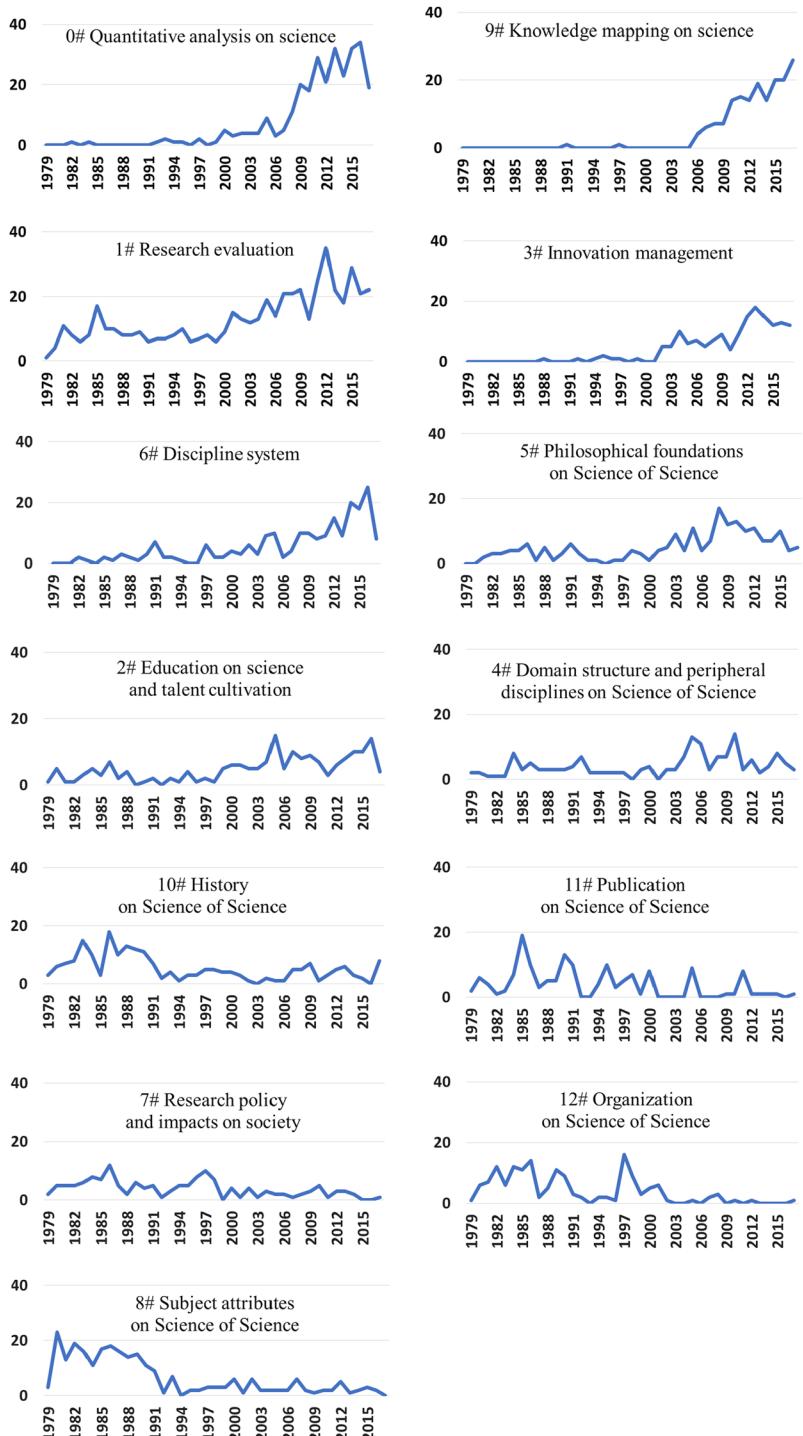


Fig. 12 Activity trends of the 13 research topics in Science of Science domain in China. Ascending trends are presented first

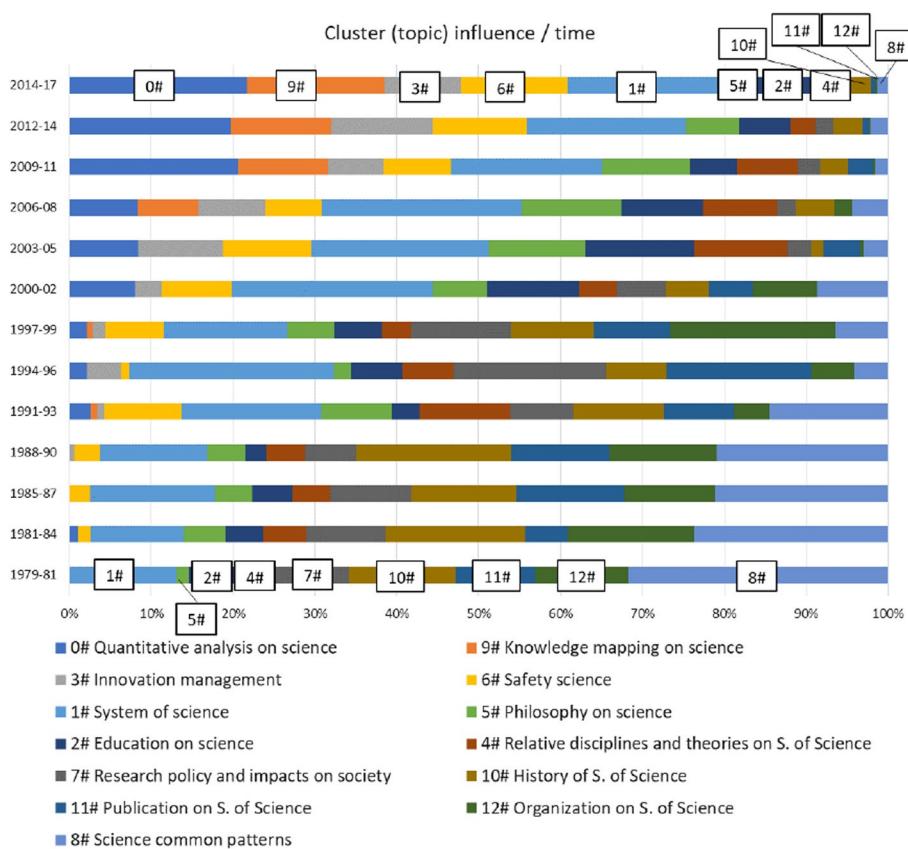


Fig. 13 Coordinated influence of research topics in Science of Science in China (by 3-years periods)

innovation”, “independent innovation”, “collaborative innovation”, “disruptive innovation”, etc.), making innovation management (3#) a very popular theme in 2013.

In 2016, quantitative scientific analysis (0#) and the disciplinary system became more popular, and in 2017, relevant research on science knowledge mapping (9#) became a highlight of scientific research in Science of Science in China.¹⁷

Figure 12 shows the distribution of the number of papers per year in all thematic groups. This kind of approach can be used to highlight specific periods of thematic activity. These include a growth in activity indicating emerging topics (0#, 9#, 1#, 3#, 6#), seed topics initiating the field and characterized by strong initial activity growth followed by a long time activity decrease (#8), topics that have had a significant and localized period of maturation/activity in the analyzed period (#10), or even topics presenting local peaks of activity that can be explained by particular events in the historical development of the field. The takeover of Science of Science initiated by the creation of the third council of the China

¹⁷ From 2009, 5 training seminars on knowledge mapping were held in WISELAB (Dalian University of Technology) broadcasting the methods and thinking widely in China. This approach also lead to the present paper using specific mapping tool to highlight the structure and the evolution of Science of Science domain in China.

Association for Science and Science & Technology in 1997 corresponds to such a local peak in topics #6 and 12# for example.

A more overall representation of the influence of each cluster (i.e. topic) in different periods (using 3-year blocks) can be derived from previous distributions. This representation presented in Fig. 13 can then be used to better understand the laws of Science of Science development in China. This point of view can especially help distinguish between important but accidental topics that have a chance of developing in the short term and rational important topics that play a major role in the construction of the domain in the long term.

The topics “0# Quantitative analysis of science”, “9# Science knowledge mapping” and “3# Innovation management” did not appear at the beginning of scientific research in the field of Science of Science in China and it was only in recent years that the status of these topics became increasingly important. The establishment of the dominant position of the topic “0# Quantitative analysis of science” shows that Science of Science has reached maturity as a subject. The importance of the topic “9# Science knowledge mapping” indicates that Science of Science has become an open subject, integrating computational approaches and information visualization technologies. The growing prosperity of the topic “3# Innovation management” shows that Science of Science is an increasingly practice-oriented domain that emphasizes the economic value of science and technology, and shows its strategic position in China today. In comparison, research topics involving the attributes of the field (8#), the construction of scientific organization and publication processes (#11) and the management of scientific research results (#12) have gradually weakened which also indicates that scientific research in Science of Science is gradually becoming mature and standardized in China.

Comparison of Chinese and word contexts in Science of Science

We briefly summarize hereafter the main differences between the results of our Science of Science study performed in China and those of former studies that have been performed in the international community, like by Fortunato et al. (2018) and Zeng et al. (2017).

1. *The states of social background and S&T development are different.* The foundation work of Science of Science by Bernal, The Social Function of Science (Bernal 1939) was directly derived from the Soviet scholar Boris Hessen's report on the social and economic roots of Newton's mechanics at the 2nd International Congress of History of Science (IUCHS) in 1931 (Hessen 1931) and it should be noted that Bernal admitted he was influenced by the “impact of Marxism”. One of the important viewpoints in his book is that “science can be planned” and he focused more on how to rebuild after the war. One landmark of the establishment of Chinese Science of Science came in 1977 when Tsien Hsueshen was the first to propose the creation of a new discipline of “科学的科学” (Science of Science) in his long essay “Modern science and technology”. It occurred simultaneously with China's Economic Reform and opening of society and another was the advent of the “spring of science” in China (The National Science Conference was held in the early spring of 1978). Han Bingcheng, who worked in the Embassy in the

- Soviet Union, simplified “科学的科学” into “科学学” according to Chinese language habits for disciplines (1978) and he introduced the development of science of the Soviet Union for the first time and was the first to use the concept of “Science of Science”.
2. *The main development path of the core research is different.* The development path of Science of Science worldwide mainly involves the “sociology of science” and “scientometrics” which is clearly shown in Fig. 1. In China, the development path mainly involved the “philosophy of science and technology” and “scientometrics” with the latter prospering more in recent years. Worldwide research has concentrated more on science and technology’s role and its influence on social and economic development and summarized the law of science and technology development through quantitative analysis. However, in China, research focuses more on how to carry out scientific work and understand the social function of science according to Marxism and Engels’ natural dialectics. The construction of China’s scientific output database was comparatively late (e.g. CNKI was founded in 1999) means that it is somewhat difficult to access to international scientific output data in China as scientometrics did not begin flourish until nearly 20 years later.
 3. *The technology innovation system is an important hot topic in the Science of Science but the research perspective is slightly different.* Chinese scholars of Science of Science entered the field of technological innovation in the 1980s and emphasized the source role of science and technology in the innovation system from the perspective of science & technology management, technological economics, philosophy of science and technology and system science. The content was broadened from focusing on “science and technology activity itself” to the interaction between science-technology-economy and research policy. Indeed a large amount of research results with Chinese characteristics are adopted by government departments and then become national strategy. In that context, some international scholars, especially in the field of scientometrics, like Loyd Leydesdorff (Etzkowitz and Leydesdorff 1997), have carried out empirical research on the triple helix innovation theory using the data of papers, patents and economy.
 4. *The development paths in China and abroad tend to recently converge.* The development of information science and data analysis technology combined with a deeper understanding of the complexity of science and technology activities, scientific big data analysis and complexity scientific analysis have become hot topics in the field of Science of Science at home and abroad. Science of Science has returned to its own method research which has encouraged advances in Scientometrics 2.0 which has become the consensus view of development both in China and worldwide.

Comparison with LDA

As our method’s topic extraction capabilities represent a central point for the overall analysis process to be correctly achieved, we went on to compare it with the LDA approach (Blei et al. 2003), a state-of-the-art method for topic extraction. We use similar data as that which was used for the clustering process as input for LDA, namely 2790 articles represented by a bag of words model of 1576 terms selected with a frequency threshold of 6.¹⁸ Gensim¹⁹ implementation of LDA is chosen for our experimentation.

¹⁸ See section “Data collection and preprocessing” for more details.

¹⁹ <https://radimrehurek.com/gensim/>.

OCF: Quantitative analysis on science	ICF: Research evaluation	2CF: Education on science and talent cultivation	3CF: Innovation management	4CF: Domain structure and peripheral discipline on SoS
0.022 journal 0.022 analysis 0.020 paper 0.019 citation 0.019 result 0.019 data 0.018 SNA 0.016 citation analysis 0.016 bibliometrics 0.015 author	0.015 system 0.012 mechanism 0.011 methodology 0.011 effect 0.010 basics 0.010 science 0.010 construction 0.009 theory 0.009 development 0.009 application	0.021 countermeasure 0.019 project 0.018 education 0.017 university 0.015 higher education 0.014 planning 0.014 aspect 0.013 China 0.013 development 0.012856 problem	0.029 mechanism 0.030 enterprise 0.027 knowledge management 0.021 innovation 0.018 collaborative innovation 0.017 analysis 0.016 process 0.016 knowledge 0.016 model	0.023 SAT studies 0.019 dialectics of nature 0.018 book 0.012 science of science 0.012 era 0.012 library 0.012 technology theory 0.011 subject 0.011 discipline construction 0.011 theory in science of science
5CF: Philosophical foundations on SoS	6CF: Discipline system	7CF: Research policy and impact on society	8CF: Subject attributes on SoS	9CF: Knowledge mapping on SoS
0.035 philosophy 0.025 history 0.020 reflection 0.019 Karl Marx 0.017 thought 0.016 context 0.016 theory 0.015 problem 0.015 marxist doctrine 0.015 comprehension	0.033 connotation 0.023 concept 0.022 definition 0.022 safety science 0.021 safety 0.019 system 0.019 discipline system 0.018 basics 0.018 principle 0.017 research method	0.018 modern science 0.017 management 0.017 S&T 0.016 science of science 0.015 S&T development 0.015 Institute of Science of Science 0.014 development 0.014 science 0.013 content 0.012 world	0.027 natural science 0.027 social sciences 0.022 science 0.020 regular pattern 0.020 scientific 0.019 science of science 0.017 modern science 0.015 development 0.015 subject 0.014 S&T	0.037493 knowledge map 0.032971 research hot topics 0.030203 literature 0.026027 visualization analysis 0.023870 field 0.021070 visualization 0.020661 analysis 0.020526 data 0.019496 CiteSpace 0.019492 information visualization
10CF: History on SoS	11CF: Publication on SoS	12CF: Organization on SoS		
0.032 J.D. Bernal 0.026 Price 0.022 science of science 0.019 social function of science 0.018 professor 0.017 science 0.017 history of science 0.016 scientometrics 0.015 modern science 0.014 United Kingdom	0.057 S&T management 0.045 magazine 0.034 Science of Science and S&T Management (journal) 0.031 S&T system reform 0.028 comrades 0.021 science of science 0.017 reform 0.016 National Science and Technology Commission (NSTC) 0.016 reader 0.015 S&T policy 0.015 problem	0.053 Chinese Association for Science of Science and S&T Policy 0.037 Chinese science of science conference 0.034 conference 0.033 academic symposium 0.032 national 0.029 S&T 0.023 comrades 0.021 Beijing (city) 0.018 science of science 0.017 National Science and Technology Commission (NSTC)		

Fig. 14 Results of topic extraction obtained by the combination of clustering and feature maximization (top 10 ranked terms and topic titles provided by experts). The blue colour is used for titles which were difficult to establish by experts with the single use of top ten topic words

For the sake of comparison with our method, the number of topics to be extracted by LDA is fixed at the optimal number of clusters identified by our clustering quality criteria (i.e. 13). The training chunk size and the number of passes of training of LDA were made to vary. The other LDA parameters are set to their default or automatic values (according to (Hoffmann et al. 2010) for grid search of optimal model). The best result as regards varying parameters is kept. For both the LDA and clustering with feature maximization methods, the topics are presented to the experts in the form of their 10 highest ranked terms. The ranking criteria used for the clustering results is the Feature F-measure value (eq. 3).

The further role of the experts was to check and approve topics through suitable identification of their content using the 10 highest ranked terms for both LDA and clustering with feature maximization topic lists. Suitable identification is assumed to derive from a topic title. The results of the process are presented in Fig. 14 (clustering with feature maximization) and Fig. 15 (LDA).

As a result, all our 3 experts agreed on the fact that they can easily label all the topics generated by clustering with feature maximization with the help of the 10 top terms

0LD: Sociology of science	1LD: Philosophical foundations on SoS ?	2LD: S&T policy ?	3LD: Research evaluation ?	4LD: Subject attributes on SoS ?
Equiv (~):SCF ----- 0.068 sociology of science 0.066 planning 0.056 opportunity 0.050 Merton 0.041 explanatory 0.039 science view 0.028 mainstream 0.036 root 0.035 big science 0.033 social economy	Equiv (~):SCF ----- 0.126 philosophy 0.054 reason 0.045 sociology 0.036 Karl Marx 0.032 a book 0.033 creativity 0.031 academic circle 0.031 mathematics 0.030 figure 0.027 national economy	Equiv (~):ICF ----- 0.060 SAT management 0.056 SAT policy 0.034 Soviet Union 0.032 National Science and Technology Commission (NSTC) 0.031 expert 0.027 for 0.026 Tianjin 0.025 SAT system 0.024 Chinese Academy of Sciences 0.023 SAT system reform	Equiv (~):ICF ----- 0.070 research 0.058 science 0.048 development 0.042 subject 0.038 problem 0.039 for 0.032 methodology 0.032 system 0.027 aspect 0.026 basics	Equiv (~):SCF ----- 0.104 natural science 0.095 social sciences 0.086 people 0.063 cognition 0.055 marxist doctrine 0.044 comparison 0.042 addition 0.038 form 0.024 economic development 0.022 spread
5LD: History on SoS ?	6LD: ??	7LD: ??	8LD: ??	9LD: ??
Equiv (~):ICF ----- 0.077 scientometrics 0.067 Price 0.067 professor 0.061 interest 0.057 actual 0.052 decision making 0.035 behavior 0.033 scientifilization 0.026 act 0.025 Zhao Hongzhou	0.084 Journal 0.073 engineering 0.058 Qian Xuesen 0.051 survey 0.046 the people 0.043 urban 0.042 technological science 0.042 event 0.035 Japan 0.030 management system	0.077 technology 0.051 effect 0.047 knowledge 0.044 policy 0.030 mechanism 0.029 evolution 0.029 industry 0.027 mode 0.023 strategy 0.021 pathway	0.091 science of science 0.052 China 0.048 society 0.048 SAT 0.034 economy 0.033 development 0.026 history 0.023 comrades 0.019 achievement 0.019 national	0.056 institute 0.056 personnel 0.043 logic 0.042 key 0.035 theory system 0.031 literature 0.029 data 0.028 publication 0.022 dilemma 0.021 technological development
10LD: ??	11LD: ??	12LD: ??		
0.049 paper 0.048 field 0.041 China 0.040 situation 0.037 research 0.031 influence 0.030 regular pattern 0.026 author 0.023 research achievement 0.022 international	0.122 magazine 0.117 human beings 0.074 psychology 0.058 reader 0.051 frame 0.049 definition 0.037 discipline 0.032 safety system 0.027 limit 0.020 interaction	0.046 significance 0.037 knowledge 0.026 pattern 0.024 tradition 0.024 organization 0.024 innovation 0.023 task 0.023 hierarchy 0.023 contribution		

Fig. 15 Results of topic extraction obtained by LDA (top 10 ranked terms and topic titles provided by experts). Blue is used for titles which were difficult or impossible (??) to establish by experts with the single use of top 10 topic words. A single question mark also means unclear content and a double question mark means no possible title (incoherent content). Green is used to represent the potential corresponding topics on the clustering with feature maximization (CF) side

(see Fig. 14). Conversely, they found that generating explanations on most of the topics produced by LDA is pretty much an impossible task as only 6 topics made some sense including 5 with very unclear meaning (see Fig. 15). Additionally, for these 5 “unclear topics” they also point out they can find clear descriptions with the clustering with feature maximization method.

An additional remark can be made about the very unstable level of generality of the results produced by LDA. Some topics look appear to refer to very specific and casual context (like topic 6LD, partly describing the main area of work of a specific scientist) while others refer to unclear areas with very large scopes (like topics 7LD to 12LD). Conversely, the results of the clustering with feature maximization method (i.e. all topics obtained) seem to reach a very balanced level of generality.

Finally, regarding the comparison of methods, clustering with feature maximization can highlight most of the important topics of the Science of Science domain as stated in section "Clustering and optimal model detection" whereas LDA failed to highlight any of the main topics like quantitative analysis, knowledge mapping, innovation

management, publication and organization in Science of Science, etc. Therefore LDA seems to be an unusable method in the context of our study especially for further fine-grained steps of detection of topic interaction and time-based topic analysis described in section "[Data analysis and visualization results](#)".

Conclusions and discussion

Science of Science in China as a practice-oriented fundamental theoretical research field was born with the reform and greater openness of the nation. In this work, we use elaborate and original methods of data analysis and knowledge mapping to objectively reveal the historical changes in Science of Science research topics in China and to reflect the central role of this field in the national development process. Our approach has also shown that the rapid development of the Chinese economy and its increasingly active practice of innovation have led to new research topics in the field of Science and Science.

The domain experts considered our approach to be a useful tool. The most specific discoveries they made with it are that research has moved from a period of pre-maturation of the subject to related disciplines and analysis of the structure of knowledge, from qualitative analysis to quantitative and visual analysis, from general research on the social function to more specific research on the economic and strategic function.

The combination of feature maximization and unsupervised learning and the joint use of contrast graphs for visualization is an original approach which we have proposed in this work. Our full-scale experiments were approved by experts in the field and showed that, without supervision, parameters or the support of any external source of knowledge, this method could very effectively reveal the research themes, their interactions and changes in a very complex research field such as Science of Science in China. In this article, we propose in particular a method for visualizing the analysis results using line maximization. This method is very suitable for large-scale data analysis in large dimensions. It also tolerates the integration of a wide range of additional information that can enrich analytical results and provide clarity and precision of results that current competing methods cannot provide. For example, as we have shown, methods such as LDA, which could potentially replace the proposed approach for the part concerning topic extraction, severely suffer from the dependence on parameters that are very difficult to control and working hypotheses that are difficult to verify on the distribution of words especially in the case of analysis of complex data. Consequently, these problems severely limit the quality of their results (level of generality, accuracy and precision) in that context.

Finally, as we have shown, the results we obtained have already proved sufficiently meaningful for all the 3 experts we asked to review our work but conducting a more detailed qualitative analysis based on semi-directive questionnaires given to several expert reviewers would certainly allow a more thorough validation of these results. We intend this step to be the subject of further work.

Acknowledgements Two of the authors of this paper and an additional famous researcher play the role of the domain experts that have been mobilized for this work. Their help proved to be invaluable in carrying out our analysis of the results the proposed methodology: Professor Liu Zeyuan is one of the pioneers of Science and Science in China and one of the most important contributors to the field. He is one of the founders of the scientific societies in Science of Science in China and also the founder of WISELAB at Dalian University, a laboratory itself a pioneer in the quantitative study of Science of Science in China.

Professor Liming Liang is one of the pioneer researchers in Scientometrics in China. She represents the Chinese researcher with the largest international academic reputation in that domain. Professor Chen Yue is the current director of WISELAB. Under his influence, this laboratory developed the use of modern methods of science analysis, especially cartographic methods. Today, this laboratory is considered one of the three main Chinese laboratories in the field of Science and Science. We also wish to warmly thank Richard Dickinson for the careful proof-reading of our paper.

References

- Attik, M., Lamirel, J.-C., Al Shehabi, S. (2006). Clustering analysis for data with multiple labels. In *Proceedings of IASTED international conference on databases and applications (DBA)*, Innsbruck, Austria.
- Bernal, J. (1939). *The social function of science*. London: George Routledge & Sons Ltd.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.
- Calafiroski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics- Theory and Methods*, 3(1), 1–27.
- Chen, Y., & Liu, Z. (2005). The rise of mapping knowledge domain. *Studies in Science of Science*, 23(2), 149–154. (in Chinese).
- Chen, Y., Zhang, L., & Liu, Z. (2017). The prelude of the science of science in the world: The third Copernican revolution initiated in Poland. *Studies in Science of Science*, 35(1), 4–10. (in Chinese).
- Cuxac, P., & Lamirel, J.-C. (2013). Analysis of evolutions and interactions between science fields: the cooperation between feature selection and graph representation. In *14th COLLNET meeting*.
- Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2, 224–227.
- Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1), 95–104.
- Etzkowitz, H., Leydesdorff, L. (eds) (1997). Universities and the global knowledge economy. In *A triple Helix of university-industry-government relations [M]*. New York: Pinter.
- Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., et al. (2018). Science of science. <https://doi.org/10.1126/science.aao0185>.
- Fritzke, B. (1995). A growing neural gas network learns topologies. In *Advances in neural information processing systems* (pp. 625–632).
- He, Y., Chen, Y., Cui, Y., & Liu, Z. (2017). Research approaches and prospects in the subject science of science based on the analysis of Bernal Price. *Studies in Science of Science*, 35(8), 1121–1129. (In Chinese).
- Hessen, B. M. (1931). *The social and economic roots of Newton's principia*. London: Science at the Cross Roads.
- Hoffmann, M. D., Blei, D. M., & Bach, F. (2010). Online learning for Latent Dirichlet allocation. In *24th conference on neural information processing systems (NIPS)*, pp. 856–864.
- Huang, J., Cheng, X. Q., Shen, H. W., et al. (2012). Exploring social influence via posterior effect of word-of-mouth recommendations. In *ACM International CONFERENCE ON WEB SEARCH AND DATA Mining*. ACM, 2012:573–582.
- 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.
- Kobourov, S. G. (2012). *Spring embedders and force directed graph drawing algorithms*. arXiv preprint [arXiv:1201.3011](https://arxiv.org/abs/1201.3011).
- 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. <https://doi.org/10.1007/s10844-014-0317-4>.
- Lamirel, J.-C., Dugué, N., & Cuxac, P. (2016). New efficient clustering quality indexes. In *2016 International joint conference on neural networks (IJCNN)*, pp. 3649–3657. IEEE.
- Lamirel, J.-C., Mall, R., Cuxac, P., & Safi, G. (2011). Variations to incremental growing neural gas algorithm based on label maximization. In *The 2011 international joint conference on neural networks (IJCNN)*, pp. 956–965. IEEE.
- Liu, Z. (2017). Feng Zhijun's puzzle: What is the core theory of the science of science? *Studies in Science of Science*, 35(5), 655–660. (in Chinese).

- Liu, Z., Chen, Y., & Zhu, X. (2013). D.J. Price's contribution to theory of the science of science. *Studies in Science of Science*, 31(12), 1762–1772. (in Chinese).
- 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.
- Pu, G., & Di, R. (1998). The cognitive turn of sociology of science. *Journal of Dialetics of Nature*, 5, 29–34. (in Chinese).
- Qian, W., & Li, X. (2012). J.D. Bernal and China. *Science & Culture Review*, 16–32. (in Chinese)
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
- Shen, H. W., & Barabási, A. L. (2014). Collective credit allocation in science. *Proceedings of the National Academy of Sciences of the United States of America*, 111(34), 12325.
- Sinatra, R., Wang, D., Deville, P., et al. (2017). Quantifying the evolution of individual scientific impact. *Science*, 354(6312), aaf5239.
- Tsien, H. (1979). Science of science, studies in science and technology system. *Marx's Philosophy Philosophical Researches*, 1, 20–27. (in Chinese).
- Tsien, H. (1980). On the establishment an development of Marxist science of science for the foundation of research management. *Research management*, (1), 3–8.
- Wang, D., & Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science*, 342(6154), 127–132.
- Xie, X. L., & Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8), 841–847.
- Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., et al. (2017). The science of science: From the perspective of complex systems. *Physics Reports*, 714–715, 1–73. <https://doi.org/10.1016/j.physrep.2017.10.001>.
- Zhao, H., & Jiang, G. (1983). Great facts, great subjects. *Science of Science and S&T Management*, 3(3). (in Chinese)
- Zhao, H., & Jiang, G. (1988). Hessen episode and the origin of science of science. *Studies in Science of Science*, 6(1), 14–23. (in Chinese).

Affiliations

Jean-Charles Lamirel¹  · Yue Chen² · Pascal Cuxac³ · Shadi Al Shehabi⁴ ·
Nicolas Dugué⁵ · Zeyuan Liu²

Yue Chen
chenyue@dlut.edu.cn

Pascal Cuxac
pascal.cuxac@instfr

Shadi Al Shehabi
shadialshehabi@gmail.com

Nicolas Dugué
nicolas.dugue@univ-lemans.fr

Zeyuan Liu
liuzy@dlut.edu.cn

¹ SYNALP Team-LORIA, INRIA Nancy-Grand Est, Vandoeuvre-lès-Nancy, France

² WISELAB, Dalian University of Technology, Dalian, China

³ INIST-CNRS, Vandoeuvre-lès-Nancy, France

⁴ University of Turkish Aeronautical Association, Ankara, Turkey

⁵ LIUM, Université du Mans, Le Mans, France