Are research landscapes from submitted project proposals or from the S&T literature similar ? A comparison using text mining and clustering

Ivana Roche, Nathalie Vedovotto, Edgar L. Schiebel, Marianne Hörlesberger, Dominique Besagni, Claire François

International Association for Management of Technology
IAMOT 2013 Proceedings

ARE RESEARCH LANDSCAPES FROM SUBMITTED PROJECT PROPOSALS OR FROM THE S&T LITERATURE SIMILAR? A COMPARISON USING TEXT MINING AND CLUSTERING

Ivana ROCHE

CNRS, INIST, UPS76, 2 allée du Parc de Brabois
CS 10310, 54519 Vandœuvre-lès-Nancy Cedex, France
ivana.roche@inist.fr

Nathalie VEDOVOTTO

CNRS, INIST, UPS76, 2 allée du Parc de Brabois
CS 10310, 54519 Vandœuvre-lès-Nancy Cedex, France
nathalie.vedovotto@inist.fr

Edgar SCHIEBEL

AIT Austrian Institute of Technology GmbH
Tech Gate Vienna, Donau-City-Strasse 1, 1220 Vienna, Austria
edgar.schiebel@ait.ac.at

Mariane HÖRLESBERGER

AIT Austrian Institute of Technology GmbH
Tech Gate Vienna, Donau-City-Strasse 1, 1220 Vienna, Austria
marianne.hoerlesberger@ait.ac.at

Dominique BESAGNI

CNRS, INIST, UPS76, 2 allée du Parc de Brabois
CS 10310, 54519 Vandœuvre-lès-Nancy Cedex, France
dominique.besagni@inist.fr

Claire FRANÇOIS

CNRS, INIST, UPS76, 2 allée du Parc de Brabois
CS 10310, 54519 Vandœuvre-lès-Nancy Cedex, France
claire.francois@inist.fr

This work aims at studying and comparing, within the scientific field of Information and Communication Technologies (ICT), two types of scientific production, both asking for consequent research efforts. The first considered data set is a corpus of records extracted from a bibliographic database and representing the results of research works published in the scientific and technological literature. The second one is a corpus of records extracted from a database collecting the information related to the proposals answering the calls for projects launched under the aegis of the European Commission in relation to the Seventh Framework Programme (FP7). After the application of a text mining approach operated with tools coming from the NLP (natural language processing) domain, a clustering step supplies a representation of each corpus by producing a

thematic map of clusters. Then, with the help of an expert, a content analysis is produced allowing comparing the map and the content of the clusters obtained for each of the two corpora under two criteria: the distribution of the developed works by topic and their potential applicability. This work intends to answer the question: Are the works published by the community of ICT researchers in scientific and technical literature and those developed in projects submitted for funding equivalent in terms of their potential applicability?

Keywords: text mining, content analysis, natural language processing, clustering, applicability, ICT

Introduction

Publishing papers in the scientific and technological (ST) literature is, for the researchers, the usual and efficient way to spread the results of their work and to submit them to the appreciation of the scientific community. Hence, bibliographic databases collecting this scientific production and returning it accessible under a structured shape provide a consistent source of current trends in science. On another note, researchers answer calls for projects issued by different funding agencies to get the means to develop their work in a better context. The European Commission, especially, develops an asserted action for promoting scientific research by launching a great number of calls for projects in many disciplinary fields and by operating a selection process. The submitted proposals are registered in a repository that represents a very reliable source of a scientific production pragmatically oriented towards funding searching.

These two types of scientific production are equally submitted to a mandatory selection step, either from the editorial board or from an expert panel mandated by the funding agency. If the editorial board essentially takes into account the scientific excellence of the submitted paper, with regards to the aims and scopes of the journal, the funding agencies add other selection criteria, as the potential applicability of the expected results of submitted proposals. For instance, the European Research Council defines the applicability as follows: "… *may well be concerned with both new knowledge about the world and with generating potentially useful knowledge at the same time. Therefore, there is a much closer and more intimate connection between the resulting science and technology, with few of the barriers that arise when basic research and applied research are carried out separately."* [EC, 2005].

Funding agencies are often faced with a great number of applications, which have to be evaluated within limited laps of time. In this context, informetric methods can offer a "helping hand" to either support the decision-making process or to evaluate its outcome. In fact, the informetric evaluation could be witnessing a significant attention in the rising need to get a grip on science output and efficiency.

Informetric methods allow us to produce a content analysis-based approach, in order to evaluate the applied orientation of a researcher's production. This work finds its origin in a previous study developed within the framework of a European project [Holste *et al.*, 2012] which goal was to support the selection process of research projects submitted for financing to the ERC (European Research Council) and in which we developed an analytical methodology based on the informetric modeling of the criteria used by their scientific experts. The potential applicability was one of these criteria, which we studied in a previous work at the single proposal's level [Roche *et al.*, 2012]. The obtained results could be used as an ex-ante assessment in a selection process, providing a decision-aid tool.

One way of making the distinction between fundamental and applied research was introduced by Donald Stokes [Stokes, 1997], who defined a two dimensions chart, "the Pasteur's Quadrant". This

label is given to a class of scientific research developments that both seek fundamental understanding of scientific problems together with seeking to be beneficial to society. The works of Louis Pasteur, a French chemist and physicist, pioneer in microbiology, are thought to exemplify this type of method, which bridges the gap between "basic" and "applied" research. The Pasteur's Quadrant (figure 1) characterizes four distinct classes of research works:

- pure fundamental research, illustrated by the work of Niels Bohr, early 20th century atomic Danish physicist;
- careful observation, with great curiosity about particular phenomena, exemplified the work of the astronomer Tycho Brahe, who collected the data used by Johannes Kepler to establish that the orbits of the planets were elliptical;
- pure applied research, exemplified by the work of Thomas Edison, North-American inventor and businessman;
- application-inspired fundamental research, described as "Pasteur's Quadrant".

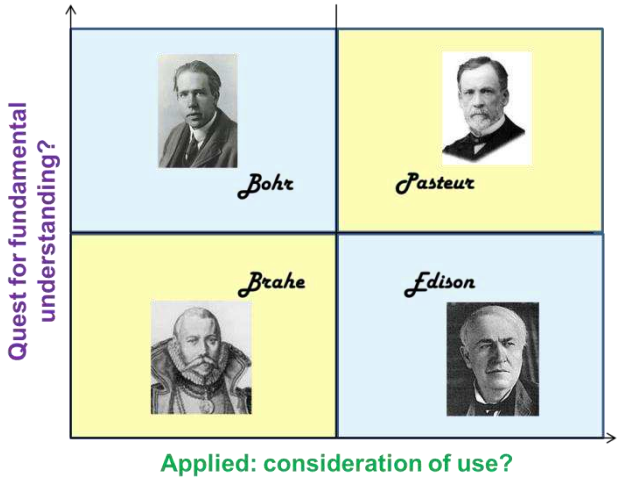

Figure 1: Pasteur's Quadrant

Our main purpose is to point out ex-post discrepancies, convergences, antagonisms or complementarities between the ICT scientific production published in ST literature and the ICT scientific production submitted in response to a call for projects by a funding agency, in terms of their potential applicability.

Data extraction

A first corpus (PC) has been extracted from PASCAL, a multidisciplinary bibliographic database produced by INIST-CNRS, providing broad multidisciplinary coverage and containing nowadays about 20 million bibliographic records resulting from the analysis of the scientific and technical international literature published predominantly in journals and conference proceedings. Moreover each PASCAL record is indexed, either manually by scientific experts or automatically based on a content analysis, by both keywords and thematic categories from a classification scheme. Our study is based on these indexing keywords.
The query operated in this work is done by a scientific expert and focuses on the ICT field, which gathers topics such as Computer science, Automation, Electronics, Telecommunications, Networking, Information science, Signal and communications theory…
The corpus has been extracted for the publication years period 2007-2011 and contains about 222,000 bibliographic records.

A second corpus (AC) has been extracted from the E-CORDA database, collecting the information related to the project proposals answering the calls for projects launched under the aegis of the European Commission in relation to the Seventh Framework Programme (FP7). The query

extracted all the projects, successful or unsuccessful, submitted to 23 ICT-related calls launched during the same period (2007-2011). A selection of fields has been operated, to keep only those which describe the project scientific content, namely the title, disciplinary subcategory, abstract and author keywords when available. This second corpus contains around 8,600 records.

Methodology

The two corpora have been processed in order to homogenize their structure. For reasons of confidentiality, data extracted from the E-CORDA database have been anonymized in order to display neither personal data about the applicants, nor information that could allow their identification by cross-checking.

A data mining step is operated on both corpora. It is based on NLP (natural language processing) techniques used to produce an assisted indexing of the records by assigning them keywords, with the support of existing terminological resources. The consistency of these keywords is contextually validated by a scientific expert, who discards stop-words and words considered as too generic (for instance, trans-disciplinary terms) or not enough consistent.

Then, a non-supervised and non-hierarchical clustering algorithm, the axial K-means, inspired by Kohonen's self-organizing maps formalism ([Lelu, 1993], [Lelu and François, 1992]), is applied to these two enriched corpora. This method considers the keywords as indicators of the content of the records, which in their turn are considered as indicators of the research topics. This step is followed by a principal component analysis leading to a 2D-mapping of the clusters. Thematic networks emerge from the relations between clusters and, according to a geographical metaphor, build a mapping of the corpus scientific landscape. This step is realized by employing an INIST-CNRS in-house software tool, Stanalyst [Polanco *et al.*, 2001], devoted to the scientific and technical information analysis. In the maps presented in "Results" section, each dot corresponds to a cluster and each line gives the connection level between pairs of clusters. The connection level is numbered by decreasing strength, from 1 to 10, and code-colored, as showed in table 1.

Table 1. Scale of the connection strength between clusters (1= strongest, 10= weakest)

colour	black	green	white	blue	red	purple	pink	orange	yellow	grey
strength	1	2	3	4	5	6	7	8	9	10

At this stage, a scientific expert performs an analysis of the clusters in terms of the scientific matter which they deal with, and of their relative position and relations in the map. To perform this analysis, the expert must adopt a particular point of view, in order to evaluate whether the content of a corpus is essentially applied or essentially fundamental: he looks therefore at the content of each cluster by considering the bibliographic records' title, abstract and keywords, in order to assess how much the cluster content can be considered as applied. We call this property the ratio "applied-fundamental" (RAF) of the cluster. Pragmatically, for each cluster, the expert:

- determines the share of applied topics (P_a) and calculates P_f, the share of fundamental topics, as: $1 - P_a$;
- determines the RAF value of each cluster, equal to ($P_a - P_f$), which is included in the interval [-1, 1]. A value equals to -1, 1 or 0, respectively, represents a totally fundamental, fully applied, or well-balanced cluster. All intermediate values are allowed.

Subsenquently, we calculate the WRAF (weighted RAF) of each cluster with the help of a parameter coming from the clustering results, in order to reinforce the contextual impact of the map of thematic clusters considered as a whole. We define this parameter, ST ratio, as the ratio between

the number of bibliographic records specific to a cluster - namely exclusively associated to this cluster - and the total number of bibliographic records contributing to the clustering production.

The sum of these weighted values, after normalization with respect to the number of clusters, gives the corpus applicateness score. This value is included in the interval [-1, 1]. If it is negative, the corpus is mainly related to fundamental research, and conversely if it is positive, the corpus is essentially dealing with applied research.

Finally, the results obtained from the analysis performed on our two corpora allows us to compare the applicateness degree of the ICT scientific production, as published in ST literature, and as represented by the scientific content of the proposals submitted to the FP7's calls for projects.

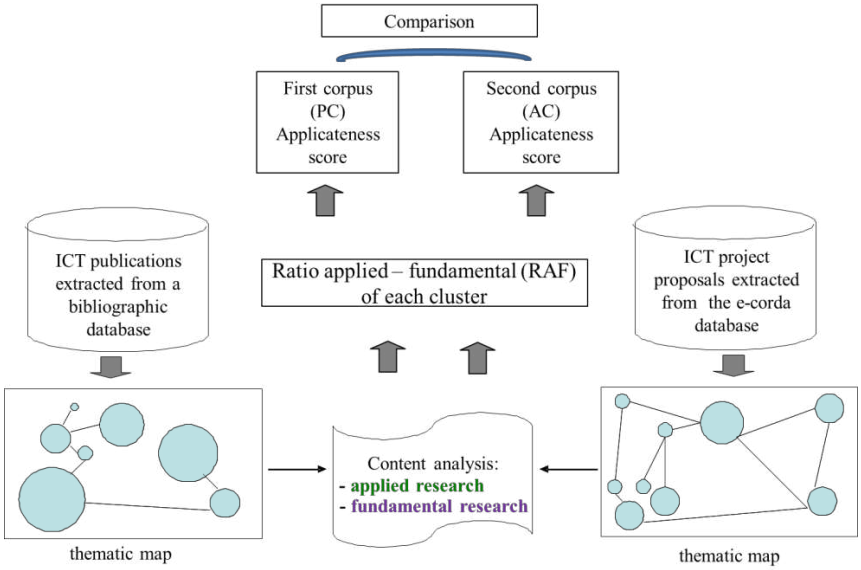

Figure 2. Methodological schema of a comparative evaluation process of the potential applicability of ICT scientific production extracted from a bibliographic database and from a funding agency repository of submitted proposals

This comparison intends to help us to evaluate the impact of the criterion "potential applicability", interesting the funding agencies, on the ICT researchers' scientific production. An exploratory study using the same methodology could be to investigate the eventual difference of applicateness score between the sub-corpora of accepted and rejected projects, obtained by splitting of AC.

Results

The cluster map obtained for the PC corpus, consisting of bibliographic records resulting from the analysis of the scientific and technical international literature and extracted from the PASCAL database, is shown in figure 3. The study of this map led the expert to validate the corpus, as all ICT sub-topics have been identified by the clustering process. These 5 sub-topics are: Automation, Computer science, Electronics, Information science, Telecommunications and signal processing. The sub-networks of the clusters related to these sub-topics are highlighted in figure 3 by blue dotted ellipses. Figure 4 shows the same cluster map, with an indication of P_a and P_f of each of the 50 clusters, in the form of colored, respectively green and purple, rectangles.
The main characteristics of theses clusters, employed in the calculation of the applicateness score of corpus PC, are shown in Table 2. For instance, the clusters "*Computer theory*", "*Signal classification*", "*Information system*", "*Fading channels*" and "*User interface*" of Figure 3 get, respectively, the RAF value -1, -0.5, 0, 0.5 and 1, which corresponds to content characteristics completely fundamental, mostly fundamental, balanced, mostly applied and definitively applied.

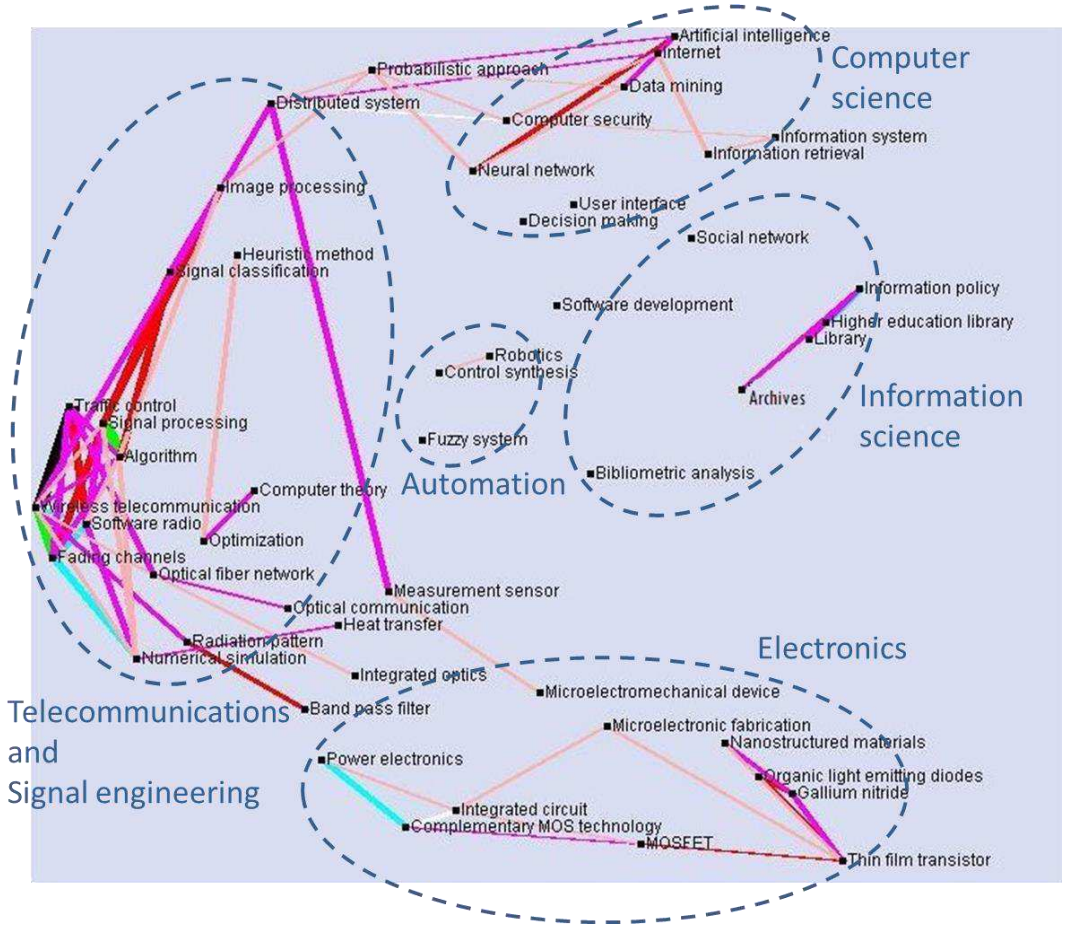

Figure 3. Cluster map of the corpus PC (PASCAL bibliographic records)

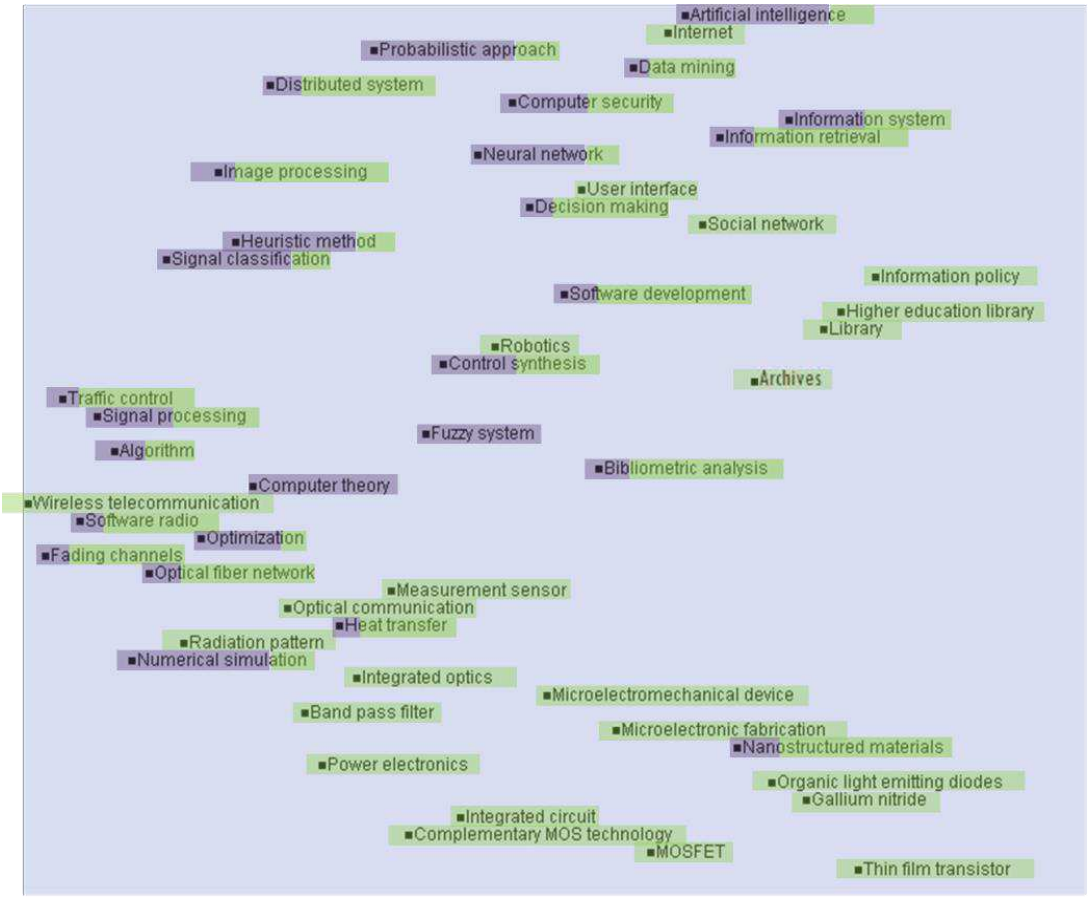

Figure 4. Cluster map of the corpus PC, with an indication of P_a and P_f of each cluster (green = P_a, purple = P_f)

Table 2. List of the 50 clusters of PC, characterized by their share of applied topics, by a clustering parameter showing their specificity in terms of bibliographic records and by their calculated WRAF

Cluster name	P_a Share of applied topics (from expert)	RAF $(2 \times P_a) - 1$	ST ratio (from clustering)	WRAF Weighted RAF (RAF x STR)
Algorithm	0.5	0	0.022	0
Archives	1	1	0.005	0.005
Artificial intelligence	0.25	-0.5	0.021	-0.011
Band pass filter	1	1	0.013	0.013
Bibliometric analysis	0.75	0.5	0.008	0.004
Complementary MOS technology	1	1	0.017	0.017
Computer security	0.5	0	0.013	0
Computer theory	0	-1	0.030	-0.030
Control synthesis	0.5	0	0.046	0
Data mining	0.25	-0.5	0.016	-0.008
Decision making	0.25	-0.5	0.020	-0.010
Distributed system	0.75	0.5	0.032	0.016
Fading channels	0.75	0.5	0.027	0.013
Fuzzy system	0	-1	0.006	-0.006
Gallium nitride	1	1	0.007	0.007
Heat transfer	0.75	0.5	0.007	0.004
Heuristic method	0.25	-0.5	0.034	-0.017
Higher education library	1	1	0.014	0.014
Image processing	0.75	0.5	0.035	0.017
Information policy	1	1	0.027	0.027
Information retrieval	0.75	0.5	0.013	0.007
Information system	0.5	0	0.013	0
Integrated circuit	1	1	0.025	0.025
Integrated optics	1	1	0.015	0.015
Internet	1	1	0.016	0.016
Library	1	1	0.005	0.005
Measurement sensor	1	1	0.013	0.013
Microelectromechanical device	1	1	0.018	0.018
Microelectronic fabrication	1	1	0.031	0.031
MOSFET	1	1	0.012	0.012
Nanostructured materials	0.75	0.5	0.020	0.010
Neural network	0.25	-0.5	0.018	-0.009
Numerical simulation	0.25	-0.5	0.013	-0.006
Optical communication	1	1	0.012	0.012
Optical fiber network	0.75	0.5	0.013	0.006
Optimization	0.25	-0.5	0.027	-0.013
Organic light emitting diodes	1	1	0.019	0.019
Power electronics	1	1	0.018	0.018
Probabilistic approach	0.25	-0.5	0.023	-0.012
Radiation pattern	1	1	0.011	0.011
Robotics	1	1	0.015	0.015
Signal classification	0.25	-0.5	0.010	-0.005
Signal processing	0.5	0	0.028	0
Social network	1	1	0.004	0.004
Software development	0.75	0.5	0.017	0.008
Software radio	0.75	0.5	0.004	0.002
Thin film transistor	1	1	0.025	0.025
Traffic control	0.75	0.5	0.010	0.005
User interface	1	1	0.013	0.013
Wireless telecommunication	1	1	0.031	0.031

The sum of the WRAF of each cluster is equal to 0.333 and, after normalization by the total number of clusters (50), the applicateness score for the PC corpus is calculated to be equal to 0.666E-02. This positive value means that the bibliographic references forming this corpus globally deal with mostly applied subjects.

The cluster map of the AC corpus, extracted from the E-CORDA database, collecting the information related to the ICT-related project proposals answering the calls for projects of the Seventh Framework Programme (FP7), is shown in figure 5. Taking the reduced size of AC corpus into account, in relation to PC, the AC classification is formed of only 20 clusters, compared to 50 for PC. By examining this map, the expert observes the presence of the same 5 ICT sub-topics as for corpus PC: Automation, Computer science, Electronics, Information science, Telecommunications and signal processing. The sub-networks of the clusters related to these sub-topics are highlighted in figure 5 by blue dotted ellipses. Furthermore, some non-directly ICT-related topics emerge, for instance in the clusters "*Public health*", "*Energy consumption*" or "*Firm management*". The analysis of their constituting proposals, as performed by the expert, concludes that these clusters concern ICT applications in the fields of respectively Health, Energy or Management.

The main characteristics of theses clusters, employed in the calculation of the applicateness score of corpus AC, are shown in Table 3.

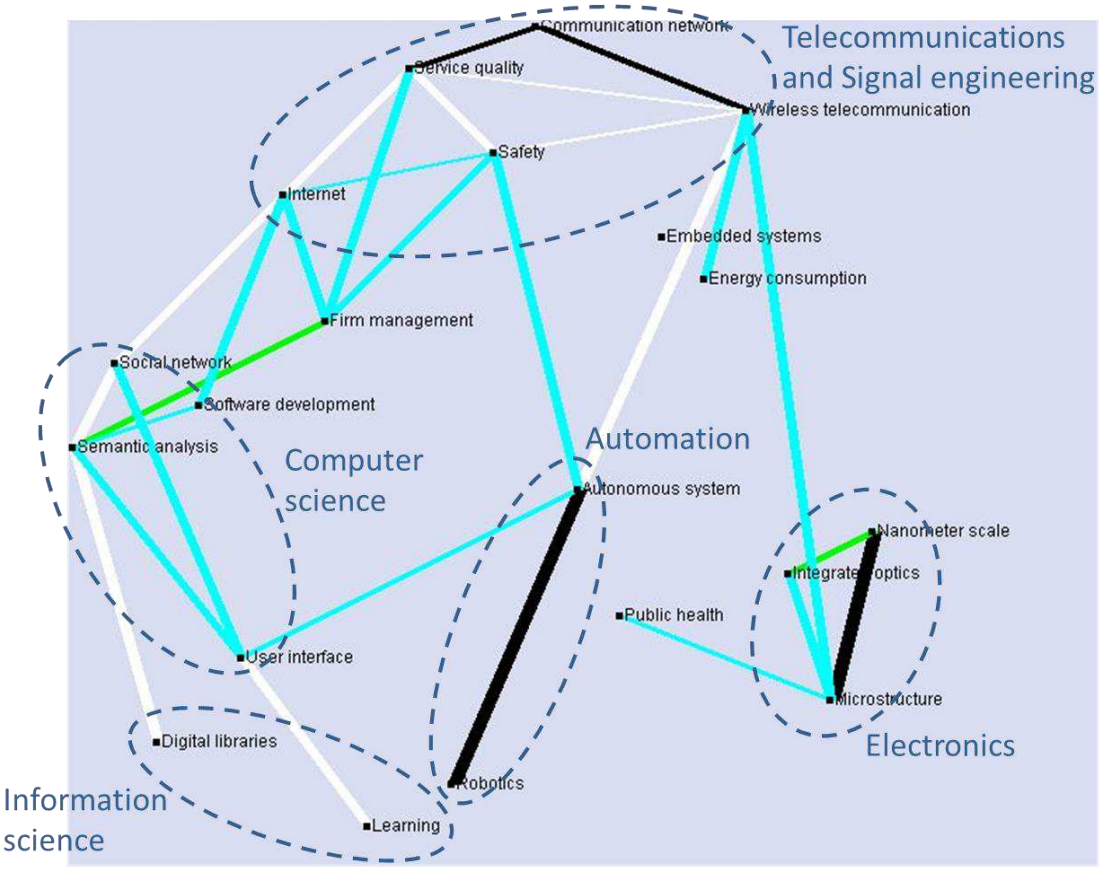

Figure 5. Cluster map of the corpus AC (submitted proposals)

The sum of the WRAF of each cluster is equal to 0.359 and, after normalization by the total number of clusters (20), the applicateness score for the AC corpus is calculated to be equal to 1.794E-02. As for PC, the positive value of WRAF means that the bibliographic references constituting corpus AC globally deal with mostly applied subjects. But the applicateness score of the AC corpus is more than 171% higher than that of the PC corpus. As in both cases the authors of the contributions

are scientific research teams, we hypothesize that a given team will more emphasize the potential applicability of its work when asking for funds than when presenting it to peers.

Table 3. List of the 20 clusters of AC, characterized by their share of applied topics, by a clustering parameter showing their specificity in terms of bibliographic records and by their calculated WRAF

Cluster name	P_a Share of applied topics (from expert)	RAF $(2 \times P_a) - 1$	ST ratio (from clustering)	WRAF Weighted RAF (RAF x STR)
Autonomous system	0.75	0.5	0.059	0.029
Communication network	0.75	0.5	0.036	0.018
Digital libraries	0.75	0.5	0.020	0.010
Embedded systems	0.5	0	0.028	0.000
Energy consumption	0.5	0	0.052	0.000
Firm management	0.5	0	0.052	0.000
Integrated optics	1	1	0.038	0.038
Internet	0.5	0	0.068	0.000
Learning	0.75	0.5	0.051	0.026
Microstructure	1	1	0.033	0.033
Nanometer scale	0.75	0.5	0.055	0.028
Public health	1	1	0.078	0.078
Robotics	0.5	0	0.036	0.000
Safety	0.75	0.5	0.060	0.030
Semantic analysis	0.75	0.5	0.054	0.027
Service quality	0.5	0	0.028	0.000
Social network	0.5	0	0.037	0.000
Software development	0.5	0	0.026	0.000
User interface	0.75	0.5	0.044	0.022
Wireless telecommunication	0.75	0.5	0.040	0.020

After having answered in a positive way the question asked in the title of this work, by observing this huge increase of applicateness score between PC and AC corpora, we tried to go ahead by investigating the applicateness score of the two corpora formed, respectively, by the successful and the rejected project proposals, using the same methodological approach.
To this end, the AC corpus was split into two sub-sets: AC-S containing the 1,295 (15 %) successful proposals and AC-R the 7,288 (85 %) rejected ones. Both sub-sets have been processed as described in the "Methodology" section, to obtain the cluster map and the applicateness score of each one. Figure 6 and 7 show the cluster maps of respectively AC-S and AC-R. The applicateness score of these two corpora is presented in table 4.

Table 4. Applicateness score of each studied corpus

Corpus	PC	AC	AC-S	AC-R
Applicateness score	0.66E-02	1.79E-02	1.47E-02	2.49E-02

The sub-network structure of the AC-S map is very different from that of the AC one. At first, all the five main ICT topics identified in the AC map are not clearly visible in the AC-S map. Indeed, the topics Electronics, Automation and Information science are not highlighted. Furthermore, Telecommunications and Computer science are not only present but strongly associated into a very dense sub-network emphasizing clusters dealing with more specific topics like connectivity (clusters "*Access network*", "*Interactive system*" or "*Interoperability*") or security (clusters "*Computer security*" or "*Identity management*"). Beside this dense sub-network, in the lower part of the map, we can observe three isolated clusters ("*Public administration*", "*Conceptual analysis*", "*Sustainable development*") dealing with applied, but usually considered as out of the ICT-classical perimeter, topics.

The AC-R map displays several sparse and widely-dispersed sub-networks among which the three ICT topics absent from the AC-S map, namely, Electronics, Automation and Information science. Other sub-networks deal with topics as Learning, Economics or Energy, also considered out of the classical ICT perimeter.

This structural difference between AC-S and AC-R maps is probably due to the difference of homogeneity level of the factors behind the acceptance or the rejection of a proposal. We hypothesize that the reasons that lead to the rejection of a proposal are much more heterogeneous than those explaining its acceptance.

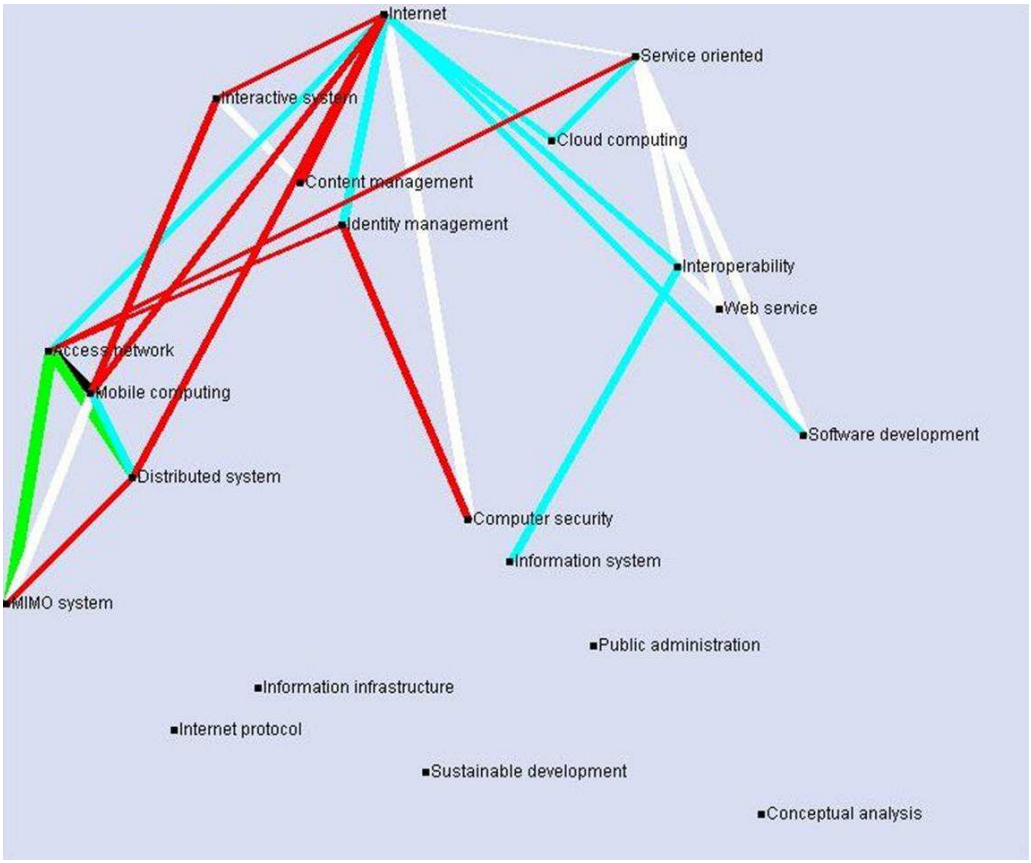

Figure 6. Cluster map of the corpus AC-S (successful proposals)

Conclusions

The initial working hypothesis was verified: indeed, our results show that the applicateness score of a corpus of project proposals submitted for funding is notably higher than that of a corpus of bibliographic records extracted from the ST literature. One inference of this result could be that researchers submitting a grant application will naturally tend to explicitly highlight the potential applicability of their proposals, while conversely the same researchers publishing in the ST literature will consent less efforts to emphasize the potential applicability of their works.

The interpretation of the comparison results of the corpora formed by successful and rejected proposals is less clear. Indeed, surprisingly, the applicateness score obtained for corpus AC-R is higher than that of AC-S (see table 4). We can however put forward some hypotheses to explain this outcome:

- We focused exclusively on the applicability criterion, being aware that in the whole selection process, the experts have to consider a bunch of criteria (applicability but also

innovation, interdisciplinarity…), which combination can sometimes contradict the result obtained with a single criterion;

- The applicateness score, as defined in this work, is not designed to determine whether a proposal is located or not within the ICT perimeter. We are unable to model the intellectual reasoning process developed by the experts to set out the frontiers of ICT-field, neither to determine a "subjective" threshold beyond which he considers a proposal as off-topic. We suppose that, among the rejected proposals, some have a high applicateness score but are considered as off-topic. Maybe, these proposals would have benefit from having been submitted to a FP7-call devoted to their core discipline.

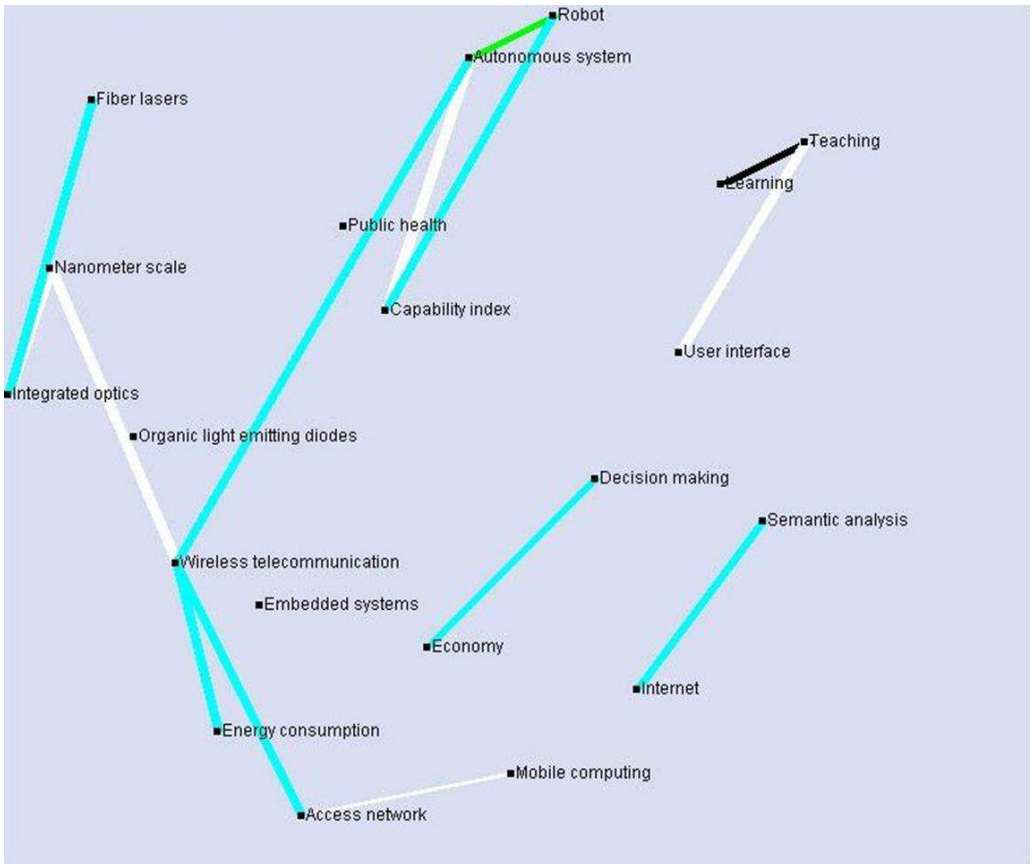

Figure 7. Cluster map of the corpus AC-R (rejected proposals)

Perspectives for future works

In future works, we could improve the developed methodology by introducing a decision-aided tool for use by the expert, making a guess for the share of applied and fundamental topics (P_a and P_f) of each cluster, based on a calculation taking into account an existing semantic categorization of its keywords.

Furthermore, as the operated data mining approach, producing the indexing keywords used in the clustering step, is hugely time-consuming, we intend to develop a computer-aided tool based on NLP recent techniques, aiming at generating the terminological extraction (CATEX) from the textual information available in the bibliographic records. CATEX is expected to drastically reduce the expert workload by decreasing the number of keywords to be validated.

Finally, taking into account the rapid evolution of the ICT field over time, we could obtain more accurate results by considering shorter time periods and by analyzing them diachronically.

Acknowledgements
This work is partially accomplished in the context of the DBF (Development and Verification of a Bibliometric Model for the Identification of Frontier Research - http://www.ait.ac.at/dbf) project within the Coordination and Support Actions (CSAs) of the IDEAS specific programme of the EU's 7th Framework Programme for Research and Technological Development (project reference n° 240765). The authors wish to acknowledge this support.
This work is made possible with the provision of data coming from E-CORDA (External COmmon Research DAta Warehouse) produced by EC (https://webgate.ec.europa.eu/e-corda/).

References

EC – European Commission (2005). Frontier research: The European Challenge. *High Level Expert Group Report*, EUR 21619.

Holste D, Scherngell T, Roche I, Hörlesberger, Besagni D, Züger M-E, Cuxac P, Schiebel E, François C (2012). Capturing frontier research in grant proposals and initial analysis of the comparison between model vs. peer review, *submitted to Research Evaluation*.

Lelu A (1993). Modèles neuronaux pour l'analyse de données documentaires et textuelles. PhD Dissertation, Université de Paris 6

Lelu A, François C (1992). Hypertext paradigm in the field of information retrieval: A neural approach. *4th ACM Conference on Hypertext*, Milano, November 30th–December 4th

Polanco X, François C, Royauté J, Besagni D, Roche I (2001). Stanalyst: An integrated environment for clustering and mapping analysis on science and technology, In *Proceedings of the 8th ISSI*, Sydney, July 16th-20th.

Roche I, Vedovotto N, François C, Besagni D, Hörlesberger M, Holste D, Schiebel E (2012). Towards a content-analysis-based methodology to estimate the potential applicability of a research project, In *17th International Conference on Science and Technology Indicators*, Montréal.

Stokes D (1997). Pasteur's Quadrant - Basic Science and Technological Innovation, Brookings Institution Press.