An analysis of IEEE publications

Michel Kerbaol, Jean-Yves Bansard, Jean-Louis Coatrieux

AN ANALYSIS OF IEEE PUBLICATIONS

Michel Kerbaol, Jean-Yves Bansard, Jean Louis Coatrieux, IEEE Fellow

Laboratoire Traitement du Signal et de l'Image, INSERM, Université de Rennes 1, Campus de Beaulieu, 35000 Rennes, France

Corresponding author :
Jean Louis Coatrieux
E-mail : jean-louis.coatrieux@univ-rennes1.fr

Many studies have recently been conducted in a number of disciplines that were aimed at tracking their evolution, detecting the emerging areas, etc. These studies have been made possible thanks to the availability of important databases (i.e Medline, ISI) and sophisticated tools based on statistical methods wich allow one to automatically analyse the contents of documents (a recent example on Genomics carried out by the authors can be found in [1]). Unfortunately, the above-mentioned popular databases do not bring enough references for engineering sciences. They are well suited to basic sciences like life sciences, physics and chemistry to mention a few, but most of the engineering data are not included [2]. This is why INSPEC has been selected here as the bibliometric corpus. Publishing in IEEE journals remains the most prestigious way to be scientifically recognized and it is this reason that leads us to limit our survey to IEEE publications. The IEEE covers most, if not all, the engineering areas from computer science to automatic control, signal and image processing, electronics and power, biomedical engineering and nuclear science.
The textual analysis that we applied to the abstracts included in INSPEC makes use of a method known as correspondence analysis [3], [4]. The frequency of words permits the selection of the most salient terms without using the keywords indicated by the authors or database indexers (which might bring some biases). These words are filtered in such a way that verbs and articles are eliminated. From there, the association of groups of words provides a link to the similarity between documents or groups of documents (journals, sub-fields, etc.) and allows one to to cross-analyse them. There are, of course, other methods that could be used for such analysis [5].

Materials

The IEEE subset consists of 354,356 documents published between 1967 and 2003. They include both papers in journals and in conferences or workshops. However, it must be emphasized that many major IEEE conferences are still not available in INSPEC.
The type (identified as record-type in the database) and number of documents that have been examined in our analysis are as follows:

1) papers in journals: 234,613

2) conference-paper; journal-paper: 52,494
3) conference-proceedings; journal-paper: 397
4) conference-papers: 66,564

The contents of these categories have been reviewed in order to verify potential classification errors. It appears that the subset 2 and 3 mainly refer to journals (they have been grouped together with 1). The number of papers in conferences is very low (category 4). The few conferences that we know like ICASSP or ICIP in signal and image processing gather individually more than 1,000 papers every year, and they play an important role in bringing new clues in research. They should be included in the near future to provide a full view of the field. In contrast, most of the IEEE journals are archived and their analysis is therefore more reliable.

Method

The basic concepts behind lexicon analysis rely on the correspondence analysis. They are derived from the observation that documents that use the same words with similar association frequencies have closely related contents. Therefore, the first step consists of estimating the frequencies of word occurrences within the whole set of documents. Then, the frequencies of word co-occurences per document are estimated and analysed. This analysis allows the construction of a space of words and a space of documents such that the words will be closer as they will be more often associated (a notion of "neighborhood" based on co-occurrence) and that the documents will be less distant as far as they contain the same word co-occurences (a "neighborhood" of documents due to characteristic co-occuring "constellations" of close words).
We define then, without any prior considerations :

1. Groups of words, named "metakeys", describing the contents of close document sets (a metakey is only valid for the corpus under study)
2. Sets of close documents, the proximity being based on their contents.

It can be seen that, in this method, the definition of metakeys results from the statistical analysis of the whole set of documents. Moreover, these metakeys are associated with the contents of the documents: they emerge a posteriori (the statistical analysis defines a classification through the contents of the document database).
The relevance of the document grouping is later submitted to experts who can interpret the contents of documents and the words that compose the metakeys. This expertise allows one to name and provide meaning to the sets of associated words, the metakeys.
The interests of the correspondence analysis can be found in its capability of addressing textual data, its underlying barycenter interpretation, the duality of spaces (which leads to an interpretation within the same space of rows and columns of the table under study or, better said, respective words and documents). The principle of distributional equivalence and the adjunction of supplementary elements (year of publication, for instance) are also some of the other advantages that can be highlighted (any new document belonging to the same corpus can be projected onto the computed space).

Results

From the datasets described above (with titles, authors, addresses, record-types, source with years of publication), a full dictionary of words was built. Note that the years 1967, 1968,and 1969 have been fused (the number of documents being low). The most frequent words (ranked by decreasing order) were then selected and a table was constructed with words on rows and years (e.g 35) in columns. The correspondence analysis was applied. This process was reiterated by reducing the number of most frequent words. The criterion used to control the appropriate number of words was based on the correlation between factors. It has been observed that this correlation was up to 0.99 for the three first factors when the number of words was varied from 10,000 to 3,000. The final selected subset of words was then reduced from 3,000 to 843 by selecting the words with a total inertia over the first 20 factors above the mean.

The first factorial plane resulting from the correspondence analysis, F1-F2 represents about 60% of the total inertia and will be used as reference. When looking at the words projected at the most distant positions over the axes (figure 1), it appears that F1 opposes the nuclear science metakey (neutron, accelerator, synchrotron), let us say A, to the computer science and telecommunication metakeys (Internet, web, multimedia, CDMA), B. F2 separates these two domains from the electronic area: submicrometer, VLSI, Mesfets (C).

The area at the centre of F1-F2 represents the transitions between A, B and C. On the positive coordinates of F1, and without any correlation with F2, we find words like radiation, gamma, and electron as well as microprocessor, MOS and diodes. The other group of terms centered on the vertical axis F2 displays mainly electronics (transistor, bipolar, device) related to C, and a few first computer words (parallel, language, database). When moving toward the negative side of F1 (B), the group of words is dominated by implementation, images, knowledge, algorithm, simulation, modelling, etc. in short making the transition between hardware and software on one hand and, on the other hand, components versus computer (routing, traffic, network) and telecommunication (antennas, fading, mobile).

This analysis is confirmed by looking at the projection of years on the F1-F2 plane. The parabola-like shape (or horseshoe) from 1970 - 2003 is depicted Figure 2 and shows a continuum over time, going from the right side of F1 to the left side. It exemplifies the major trend towards software engineering and, at the same time, the reduced contribution of basic sciences (nuclear science, power). However, if the 1970's were relatively stable in terms of contents, the 1980s and the 1990s have brought quick evolutions. The most recent years (1999 - 2003) appear to be less subject to this trend (they have almost constant coordinates on the F1 axis). The F2 axis, separating A-B from C, shows the opposition between the earliest and the latest publications with the end of the 1980s and beginning of the 1990s.

This path onto the first factorial plane is well matched by journals and conferences, gathering all disciplines of engineering sciences. When projecting journals and conferences onto the same space, the major feature is related to the difference over time between journals and conferences. It can be estimated about 3 years which can be considered to be large. This problem is not new, of course, and has been emphasized for several years

inside and outside the IEEE.
This situation, however, can mask major differences between journals regarding both their capacity to track emerging areas (with a short delay after conferences) and to evolve in contents over time. The path of the Proceedings of IEEE, for instance, fits very well with the trajectory displayed over years (figure 2) in magnitude and shape. Such a result means that it efficiently captures the trends. In contrast, IEEE transactions on Automatic Control seems to remain focused on the same topics over years: its trajectory is much shorter and stays concentrated at the centre of the plane. Most of the early published IEEE journals depict this parabola-like shape, illustrating the influence of electronics or devices and later of microprocessors and computers. However, IEEE Transactions on Signal Processing as well as IEEE transactions on Medical Imaging, for instance, project onto the bottom-left part of the plane, demonstrating that they are mainly concerned by algorithms and software (the latter being more stable over the last decade). They evolve very fast after their launch. This first glance demonstrates that the engineering field, as displayed by IEEE, is more and more attracted to computer science and less rooted to electronics and electrotechnical engineering.

Conclusion

This study has reported a few elements on the scientific production of the IEEE. Other features could be displayed that would be interesting for a better understanding of the trajectories of the Societies, the journals, etc. It should be extended to more specific datasets. Biomedical engineering, our area of interest, will be explored in a future article. What conclusions can be derived from such studies? It is always important to know where we are coming from, to understand the relations between subfields, journals, and conferences, but it is also important to know where we are going. The possibility of projecting new data onto the current spaces allow us to see if journals are static (the concepts and methods remaining stable) or dynamic (evolutions, ruptures can be tracked). In other words, this type of analysis can be used as a strategic tool to follow the impact and trends in engineering sciences. The fact that we concentrate on the IEEE publications prohibits any comparison with other societies publishing engineering papers. Such insights are fully feasible through the analysis of the INSPEC database (and perhaps others). They could bring other clues on the coverage, the competition, and the reaction to new areas.

References

[1] G. Filliatreau, S. Ramanana-Rahary, V. Blanchard, N. Teixeira, M. Kerbaol, J.Y. Bansard, Bibliometric analysis of research in genomics during the 90's, *OST Thematic Studies*, Oct 2003 [Online] Available: http://www.obs-ost.fr

[2] J.L. Coatrieux , J.Y. Bansard, M. Kerbaol, About the use of bibliometry for evaluation, *Innov.Technol.Biol.Med*, vol 25, n°1, pp. 61-66, 2004.

[3] J.P Benzecri, « Description des textes et analyse documentaire », *Cahiers de l'Analyse des Données*, vol 9, n°2, pp. 205-211, 1984.

[4] M.J. Greenacre, *Theory and Applications of Correspondence Analysis*, London: Academic Press, 1984

[5] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R.A. Harshman, *Indexing by latent semantic analysis*, J.Amer.Soc.Information Science, vol. 41, n°6, pp. 391-407, 1990.

CAPTIONS

Figure 1. The first factorial plane F1-F2 with the most significant words (they have been sampled in order to make the graph readable) highlighting the main clusters (A, B, and C). The other points, not explicity defined, are also representing words contributing to the formation of the two first axes.

Figure 2. The evolution over time of the contents of IEEE publications. The study was performed per year ans has shown a continuum over years. The evolution is captured here by subsampling the years for readability.