Text mining on large corpora using Taltac4: An explorative analysis of the USPTO patents database

Text mining su corpora di grandi dimensioni utilizzando Taltac4: Un'analisi esplorativa del database dei brevetti USPTO

Pasquale Pavone, Arianna Martinelli and Federico Tamagni[1]

Abstract. This paper aims to make a brief presentation of the main features and potential of the Taltac4 freeware software through an exploratory analysis of a large corpus (more than 600 million of occurrences) which includes all the abstracts of the USPTO patent documents. Patents have been extensively used as a source of information on innovative activity but the textual content of patent documents has not been fully exploited in existing research. Our preliminary results are promising and suggest that text analysis of patent abstracts can help developing new classification of innovative activities, overcoming the shortcomings of existing classifications of technologies.

Abstract. *In questo lavoro vengono brevemente illustrate le principali caratteristiche e potenzialità del software freeware Taltac4 attraverso un'analisi esplorativa di un corpus di grandi dimensioni (più di 600 milioni di occorrenze) che include tutti gli abstract dei documenti dei brevetti USPTO. I brevetti sono stati ampiamente utilizzati come fonte di informazioni sull'attività innovativa, ma il loro contenuto testuale non è stato pienamente sfruttato nella ricerca esistente. I nostri risultati preliminari sono promettenti e suggeriscono che l'analisi testuale degli abstract dei brevetti può aiutare a sviluppare una nuova classificazione delle attività innovative, superando le carenze delle classificazioni esistenti delle tecnologie.*

Key words: Text mining, large corpus, textual Big Data, patents

[1] Pasquale Pavone, Sant'Anna School of Advanced Studies; pasquale.pavone@santannapisa.it:
Arianna Martinelli, Sant'Anna School of Advanced Studies; arianna.martinelli@santannapisa.it
Federico Tamagni, Sant'Anna School of Advanced Studies; federico.tamagni@santannapisa.it

1 Introduction

The enormous availability of textual data produced by the mass digitization of documents has generated considerable empirical data for scientific investigation. In the social sciences, the use of text mining tools and statistical methods to analyze textual Big-Data has become unavoidable. In this context, TaLTtaC4 represents an open-access tool offering great potential to analyze large collection of textual data.

TaLTaC is the acronym for "Trattamento automatico Lessico-Testuale per l'analisi del Contenuto" (lexical-textual automatic treatment for content analysis). It has been under development since 1999 within the research group coordinated by Prof. Bolasco, and has been designed for automatic text analysis in the dual logic of Text Analysis and Text Mining [4]. The previously released freeware version of TaLTaC, named TaLTaC2.11.3, has a limit on the size of the corpus it can analyze; in particular, it can analyze Corpora in text file format, with a maximum size of 150GB and 100,000 documents. The newly released TaLTtaC4 represents a substantial step forward, as it does not face limits on the size of the corpora's size to be analyzed, other than the storing limits those imposed by the machine on which TaLTtaC4 is working.

Technically, TaLTaC4 (T4) represents a multi-platform software that maximizes the exploitation of the hardware's computational capabilities. T4's architecture is divided between Graphical User Interface and computing core, communicating with each other via HTTP protocol. The computing core is capable of processing textual data in multi-process mode and thus exploits the host machine multi-core capabilities [3].

The aim of this paper is to provide a presentation of the T4 potential, analysing the large corpus of the United States Patent and Trademark (USPTO) patent documents. In the economic and innovation literatures, patent data are widely employed to measure innovative activities [6] and, over the years, scholars have been very active in exploiting information in patent documents to develop indicators highlighting patent intensity as well as different characteristics of the inventions disclosed in patents. For instance, as patent documents do not have any direct indication of the value of the inventions, some 'indirect' measures such as patent citations [14], patent renewals, patent families [8] and patent scope have been developed and validated to know more about the characteristics and qualities of inventive outputs. The number of patent claims (i.e. the list of the subject-matters protected by the patent) or the number technological domains covered in the patents have been used to measure the scope of the patents both from the technological [11] and legal point of views [9,7].

All these indicators exploit information easily retrieved from a limited section of the patent document, which is the first page. However, patents are granted over a complete disclosure of the protected invention which is described in detail in the abstract and in the remainder of the document. Increasingly easy access to the entire patent text (e.g., via the EPO-PATSTAT Database, Google patent database, web-scraping), together with advances in text mining techniques, brought about new research attempts, exploiting various parts of the text to unfold a number of invention's characteristics such as patent similarity [1], patent novelty, or the degree of basicness of a patent (i.e. relation to basic vs. applied science).

New techniques and tools as T4 allowing to better exploit the information content of the patent documents can provide the basis for further and more sophisticated analysis of the innovative process at different levels (e.g. firm, region, country).

The paper is organized as follows. In Sec. 2 we provide an outline of the logic work in T4, while in Sec. 3 we present the preliminary results that only concern an exploratory analysis applied to the Corpus of abstracts of the USPTO patent. Finally, conclusions are drawn in Sec. 4.

2 Methodology and logic of work on T4

Through T4, textual information - unstructured by nature - is structured in two main databases, defining the two analysis domains: the Vocabulary DB, for the lexical analysis and the Fragments DB for textual analysis.

In lexical analysis, the study object is the lexicon, and the single word represents the elementary unit of analysis. However, depending on the corpus characteristics and the research question, multi-word expressions, lemmas or word stems can be considered units of lexical analysis, instead of words. In natural processing languages (NLP), particular attention is devoted to recognizing the nominal multiword expressions in a corpus [13]. These expressions represent the specialized terminology of a sector and their recognition makes it possible to work with semantic unambiguous lexical units.

In the Vocabulary DB, each unit of analysis can be associated with annotations of grammatical, semantic and statistical nature. Each of these properties constitutes an example of meta-information attributed to the lexical units, which can be retrieved by querying the Vocabulary database's corresponding fields in which this information is stored. Additionally, T4 produces several vocabularies for a multi-level lexical analysis, in which every layer corresponds to a vocabulary with the different lexical units defined. The extraction/selection of the vocabulary parts serve to "tell" the lexical characteristics of the corpus by highlighting the significant elements of each "part of speech", or to "illustrate" certain subsets of units and the relations existing between them.

In text analysis, the object of study is the corpus, and the unit of analysis is the context unit, i.e., a fragment of text, whether it is a sentence, a section of a document, an entire document, or a group of documents. In analogy to Lexical Analysis, each context unit constitutes an entry in the Fragments database to which are associated both the modes of the a priori coded variables and the textual annotations (categorizations) resulting from Textual Analysis. These annotations can be of various kinds: i) syntactic, obtained through the categorization of documents in which certain syntactic structures or groups of variable elements are present; ii) semantic, concerning automatic categorizations on the basis of certain lexicons, and iii) quantitative. Strings of text, which can be the occurrences of lexical units, both of their classes and relationships between classes or between individual units and classes, are searched through Regular Expressions (RE). The result of such elaboration is to recover the fragments that verify the textual query; to inventory the list of the extracted strings; to annotate eventually the fragments.

Based on different lexicon fragments obtained through the analysis, both lexical matrices (words x categories) and textual matrices (fragments x words) can be extracted. These matrixes can be used to represent the extracted information using infographic tools or can be further analyzed using other statistical tools.

3 Explorative analysis of USPTO Corpus

The corpus under analysis includes 5,573,936 abstracts of patents granted by the USPTO between 1980 and 2015. Each patent is assigned to at least one IPC (International Patent Classification) code, indicating the subject to which the invention relates. The IPC classification is a hierarchical classification system consisting in 5 levels of different granularity to which correspond a different number of digits[2].

After the first parsing of texts, the Vocabulary (lexical DB) includes 1,469,138 different words referred to 641,666,177 total occurrences. Through grammatical tagging of vocabulary entries, it was possible to define word lemmas as lexical analysis units. Based on grammatical annotations, we apply a hybrid multiword expression (MWEs) recognition system [5,12], based on string search according to syntactic structures. Through this technique, we identify 3,570 MWEs with at least 3,000 occurrences in the corpus[3].

In order to explore patents' content, all lexical units classified as adjectives and nouns (lemmas and MWEs) were selected to build a series of matrices for the graphical representation. To observe and study the general relationships between the elements of the matrices, we use the correspondence analysis [2]. As our first objective of the study is the analysis of the temporal evolution of innovative activities, we construct yearly matrices of the type <Lexicon x Year>. The correspondence analysis (CA) on the yearly matrices highlight similarity and differences of lexical profiles over time.

Figure 2 shows the distribution of the years and the selected lexical units., on the bi-dimensional plane spanned by the first two factors from the CA. The figure unfolds

[2] An IPC class has the form of H04J 1/10. The first letter represents the "section", combined with a two digits' number, it represents the "class" (H04), and the final letter indicates the "subclass". The digits after the subclass indicates the "group" and after the oblique stroke the least two digits indicates the "main group". A three-digit IPC class is at the level of subclass. For a complete overview see: https://www.wipo.int/classifications/ipc/en/

[3] The 20 most recurring MWEs we have found are: *mobile device, control device, control system, computer program, control circuit, computer system, virtual machine, mobile terminal, network device, light guide, management system, optical system, medical device, gas turbine, processing system, video data, film transistor, data processing, bit line, solar cell.*

a chronological development of the patent content and through cluster analysis we identify four temporal clusters represented with the different colours.

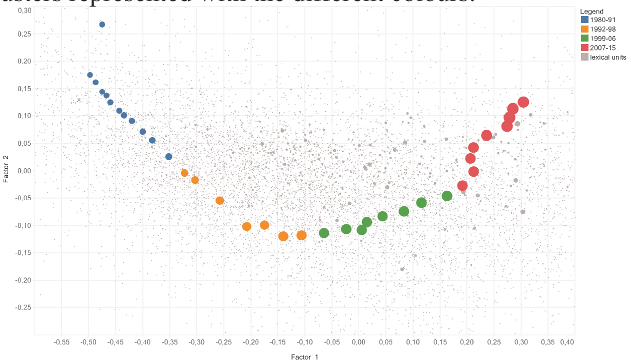

Figure 2: Distribution on the factorial plane *f1f2* of the lexicon and the years grouped in four clusters.

While this result is interesting, it would be compelling to observe in detail the elements characterizing these different temporal moments. We undertake this further step of the analysis using a different matrix of the type <Lexicon x IPC_3_DIGIT>, where IPC_3_DIGIT indicates the three-digit level of the primary IPC classification of each USPTO patents in our corpus. The CA on this matrix highlights the similarity of patent groups defined through their three-digit IPC codes. In this case the cluster analysis applied on the CA results, allows us to identify groups of patents with high technological semantic similarity.

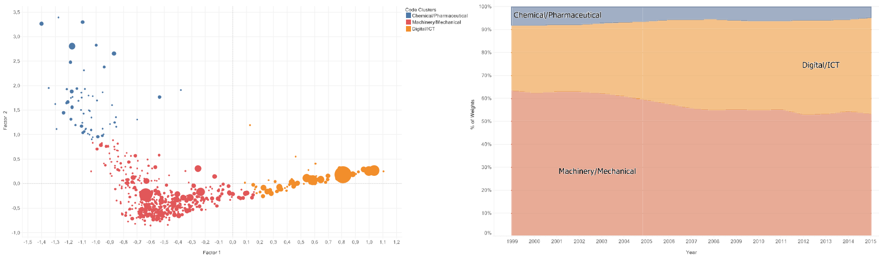

Figure 3: Distribution on the factorial plane of the codes (left) and of the percentage weight of each detected Industry over time (right)

This conceptual homogeneity emerges from the prevailing theme or semantic trait in each group, read through their characteristic dictionaries highlighted by test-values [10]. This procedure allowed us to define three clusters of Patent codes for each year, which single out specific industrial activities. Specifically, the following three industries were recognized: Chemical/Pharmaceutical; Machinery/Mechanical; Digital/ICT. Figure 3 shows the distribution on the factorial plane of the IPC_3_DIGIT (left) and the weight of each Industry detected over time (right). It is possible to clearly observe how Digital/ICT has gradually occupied a greater space in the world of patents, going from 28% in 1999 to 42% in 2015.

4 Conclusion

As working with textual Big Data is becoming increasingly common and relevant for research, in general, and for social sciences research in particular, T4 represents a freeware essential tool for text mining of very large corpora. Its initial application to USPTO patent data, as shown here, was particularly successful in identifying text patterns, in turn mapping into meaningful and well recognizable industry classes. This initial result indicates that text analysis can provide a viable way to overcome some shortcomings of the existing classification of innovation activities based on IPC codes. The further step of our research will exactly move in the direction to build a new taxonomy based on a fuzzy categorization of patents' membership within a system of industrial categories defined through text analysis. A key ingredient to this aim will be the integration in the software of the most recent text analysis tools, in particular those aimed at identifying the universe of topics in a corpus.

References

1. Arts, S., Cassiman, B., Gomez, J.C., 2018. Text matching to measure patent similarity. Strateg. Manag. J. 39, 62–84.
2. Benzécri, J.-P., 1992. Correspondence analysis handbook, Statistics, textbooks and monographs. Marcel Dekker, New York.
3. Bolasco, S., De Gasperis, G., 2017. TaLTaC 3.0. A Multi-level Web Platform for Textual Big Data in the Social Sciences, in: Data Science and Social Research. Springer, pp. 97–103.
4. Bolasco, S., Morrone, A., Baiocchi, F., 1999. A paradigmatic path for statistical content analysis using an integrated package of textual data treatment, in: Classification and Data Analysis. Springer, pp. 237–246.
5. Bolasco, S., Pavone, P., 2010. Automatic dictionary-and rule-based systems for extracting information from text, in: Data Analysis and Classification. Springer, pp. 189–198.
6. Griliches, Z. (1990). Patent Statistics as Economic Indicators: A Survey. Journal of economic literature, 28(4), 1661-1707. Retrieved from www.jstor.org/stable/2727442
7. Kuhn, J. M., & Thompson, N. (2017). The Ways We've Been Measuring Patent Scope are Wrong: How to Measure and Draw Causal Inferences with Patent Scope. Available at SSRN 2977273.
8. Lanjouw, J. O., Pakes, A., & Putnam, J. (1998). How to count patents and value intellectual property: The uses of patent renewal and application data. The journal of industrial economics, 46(4), 405-432.
9. Lanjouw, J. O., & Schankerman, M. (2001). Characteristics of patent litigation: a window on competition. RAND Journal of economics, 129-151.
10. Lebart, L., Salem, A., Berry, L., 1998. Exploring Textual Data, Text, Speech and Language Technology. Springer Netherlands. https://doi.org/10.1007/978-94-017-1525-6
11. Lerner, J., 1994. The importance of patent scope: an empirical analysis. RAND Journal of Economics, 319-333.
12. Pavone, P., 2018. Automatic Multiword Identification in a Specialist Corpus, in: Tuzzi, A. (Ed.), Tracing the Life Cycle of Ideas in the Humanities and Social Sciences. Springer International Publishing, Cham, pp. 151–166. https://doi.org/10.1007/978-3-319-97064-6_8
13. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D., 2002. Multiword expressions: A pain in the neck for NLP, in: International Conference on Intelligent Text Processing and Computational Linguistics. Springer, pp. 1–15.
14. Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. The RAND Journal of Economics, 172-187.