Intellectual property semantic search methods and visualization tools

Damien Lapray* and Serge Rebouillat

*damien.lapray@epfl.ch
BMI, EPFL, Lausanne, 1005 (Switzerland)

Introduction

Today's economic model depends for a big part of it on innovation. Industries and universities have entered a race fuelled by the vast amount of data one can access through multiple sources. "Big Data" carries on vast fields of exploration and promises for discoveries and innovation. For the past decade patents are at the center of a conceptual revolution, the largely appraised 'open-innovation' as well as the "disruptive innovation". Firms do not exclusively exploit internally generated ideas to market their products but incorporate external ones as well. Granted patents are growing every day following a sharp slope barely encountered before which requires the use of 'clever algorithms' to identify relevant material. Companies' whole innovation strategy and sustainable growth rely largely on documents analysis and it is crucial to extract in the most efficient ways the best of the available information. Despite the importance to develop high performing tools that match users' requirements, information retrieval in the intellectual property (IP) domain is by far not as satisfactory as sometime acclaimed.

Every actor in the patent bigger data extraction, dreams of a recipe book that would lead to the miracle of smart mining and discovery. That is not the objective of the present study, which is by far much more realistic and aim at efficiently bringing the relevant communities to appreciate the opportunity and the dimension of the promise that goes with this new era.

Keyword-based search and in particular Boolean search have dominated for decades free and commercial information retrieval tools including IP search (Manning, Prabhakar, & Schuetze, 2009). On the contrary to other technologies these methods give clear and fast results, e.g. a document matches or not a query. However, its apparent simplicity carries many pitfalls that semantic solutions could potentially overcome. Though semantic technology is not a novel concept and has been a main branch of artificial intelligence research since its beginning, it has been only recently integrated in IP search tools (Rebouillat S. & Lapray D., 2014). In the IP domain Latent Semantic Analysis is very often favored upon Natural Language Processing because of its ability to not rely on external ontologies and other dictionaries.

The adage "a picture is worth a thousand words" suits particularly well the IP domain. In a patent; also true for scientific publications; drawings, diagrams, plots convey an often critical amount of information that is largely ignored by IP search engines (Bhatti & Hanbury, 2012; Rebouillat S. & Lapray D., 2014a). When looking for images in patents one has to do without color information, one major element of image retrieval technologies, as well as highly normalized sketches or drawings, perfectly suitable for patent written description, and not especially suitable for state of the art image analysis. This is obviously not a trivial task but some other domains seem to have done better in this matter and we do believe that IP search cannot afford to not exploit the huge potential of images.

Let's base our propos on comparative illustrations rather than smart and overconfident interpretations of what remains excessively "black boxed". Information retrieval is not complete without

the visualization tools that allow to intuitively understand the large amount of extracted data and to be the foundation of further analysis. However, patent retrieval technologies are still in the pie chart era and do not propose much beyond that.

The aim of this study was to qualitatively compare two semantic retrieval systems using visualization tools borrowed from other domains and show how it can benefit to the whole IP analysis process.

Case studies

Two Latent Semantic Analysis based tools, here named system P and X, were used to retrieve patents associated to the biomimicry domain. A recent publication exploring this latter and in particular the 'natural polymers' and 'pharma' aspects of it was used as query material (Rebouillat S. & Lapray M., 2014). For each query in each system, the 1000 most relevant documents, as classified by the two respective retrieval tools, were extracted and further analyzed using visualization software. In order to perform this task, open-source software from other fields of research were used such as the Science of Science (Sci2) tool, Pajek, Cytoscape, Gephi and VOSViewer. Visual comparisons of the different sets of retrieved documents by system P and X were based on metadata information such as the patent filed date, the International Patent Classification number and the assignee's name (Lapray D. & Rebouillat S., *accepted*).

Results

We show here that the retrieved sets of patents can differ quite significantly in one or several of the visualized dimensions. Furthermore, we demonstrate that semantic based technologies are effective to retrieve relevant IP material. We also demonstrate that visualization tools can be borrowed from other domains and easily adapted for such data points adding an important instrument to the IP experts' analytical package.

Conclusion

This research aims at reassuring the IP and technical/business communities that despite the imperfection of the tools at hand, solutions exist inside and outside their field of research. This study has far reaching applicability in the field of data mining and may trigger the long waited disruptive innovation in the retrieval of valuable IP information.

References

Bhatti, N., & Hanbury, A. (2012). Image search in patents: a review. *International Journal on Document Analysis and Recognition (IJDAR)*. doi:10.1007/s10032-012-0197-5

Manning, C. D., Prabhakar, R., & Schuetze, H. (2009). *An introduction to information retrieval*.

Rebouillat, S., & Lapray, D. (2014). A Review assessing the "used in the art" Intellectual Property Search Methods and the Innovation Impact therewith. *International Journal of Innovation and Applied Studies*, *5*(3), 160–191.

Rebouillat, S., & Lapray, M. (2014). Bio-inspired and Bio-inspiration : a Disruptive Innovation Opportunity or a Matter of " Semantic "? A Review of a "stronger than logic" Creative Path based on Curiosity and Confidence (4C22C©). *International Journal of Innovation and Applied Studies*, *6*(3), 299–325.