Ricardo EITO-BRUN
Universidad Carlos III de Madrid, Spain
reito@bib.uc3m.es

INFORMATION MANAGEMENT TOOLS FOR INNOVATION ANALYSTS

Key words

Innovation, scientometrics, text mining, opinion mining; text visualization.

Abstract

Innovation management is a knowledge-intensive process that requires dealing with different sources of data to identify relationships between the concepts, techniques, and tools that may led to innovations. Innovation analysts need to handle huge amounts of unstructured information: ideas gathered from internal staff and external partners, research papers and technical reports, patents and applications, etc. All these sources constitute valid inputs to assess the innovativeness of ideas, the feasibility of their implementation, and their potential value in the market.

Innovation management discipline has widely used techniques and methods developed in the context of Information Science to support the identification of research trends, assess the outputs of innovation efforts and investments, and monitor the market and the activities made by competitors. The fruitful relationship between Information Science techniques and Innovation management needs to be regularly reviewed as new techniques and tools are designed and made available to the community. In the last years, significant progress has been achieved in areas like scientometrics, text visualization, and opinion mining.

This paper provides an overview of these techniques and discusses how they can help professionals involved in innovation programs.

Introduction

Innovation management is a sub-discipline of management that studies the rules that govern the generation, diffusion, and adoption of innovation, and the relationships between innovation inputs and outputs. Professionals involved in innovation management and assessments need to deal with a complex ecosystem of information sources. The number of information sources and databases they need to monitor makes their daily work difficult. The identification and regular monitoring of data sources is time-consuming and requires dedicated staff. Due to this reason, companies of different sizes and areas of activity opt for getting the support from external parties and Knowledge & Technology Transfer (K&TT) offices. Different experiences show that collaborating with K&TT offices is a feasible, affordable way to ensure that the information and knowledge needed to support, leverage, and optimize the innovation efforts are properly acquired and analysed.

The growing importance of the Open Innovation paradigm has led to a significant revision of the approaches traditionally applied. Innovation management moved from a linear model that comprised a sequence of activities completed by a single entity, to a collaborative model based on feedback loops and interactions between different partners [2, 15]. In the linear model, the achievement of innovations was determined by the planning, financing, and execution of internal R&D activities or external technology acquisition. Today, innovation is achieved thanks to the close collaboration between the company and its environment [16]. This evolution culminated in the Open Innovation model [4], which states that valuable ideas may come from both inside and outside the company, and it is the result of factors like knowledge specialization, the availability of highly skilled workers, the increasing capabilities of suppliers, and the difficulties of having a complete domain of all the aspects that need to be mastered in a successful innovation life cycle. Different entities or agents have a different level of participation in the generation of the knowledge streams that provide the inputs to create innovations (market, scientific and technical knowledge, and social knowledge) and the complex interfaces between them. The popularity of Open Innovation has led companies of any size to consider the need of making a systematic planning of innovation.

To which extent have these changes affected information management practices supporting innovation management? Until recently, information management practices focused on the surveillance of the activities done by competitors, market research, and monitoring. Information professionals' main tasks consisted in checking what was published by or about the company's competitors and summarize this information in the form of market intelligence reports. The search of patents was part of this continuous monitoring of the external environment.

With Open Innovation, a change was necessary. For example, new sources of information need to be considered, like *web-based innovation platforms*, defined as "*technical infrastructures for knowledge sharing, discovering, and social interaction*" [3, 11]. Examples of these popular platforms include, among others, InnoCentive, NineSigma, Brightidea, InnovationExchange, Atizo, YourEncore, Battle of Concepts or Yet2.com, just to name a few. Using these platforms, companies have the possibility of posting and forecasting *business challenges* and collecting ideas from outsiders. These ideas will be later analysed to identify potential improvements in existing products or requirements that may lead to the development of new products or services. Similar tools to support co-creation, collect ideas, and complete their assessment have been launched by companies. These tools embody the spirit of the Web 2.0 paradigm in the area of innovation, opening a door to outsiders to improve products and services with the collaboration of the users' community and potential partners.

Within this context, professionals dealing with the management of information to support innovation processes should consider the possibilities offered by tools and techniques developed in the context of Information Science. These range from scientometric indicators to complex visualization techniques.

1. Scientometrics

Scientometric refers to the study of the rules that govern the generation of knowledge in scientific and technical domains. The term evolved from traditional *bibliometry*, which was developed in the second half of the last century to apply statistics in the analysis of the bibliographic production. Bibliometric studies were widely applied to study the productivity and impact of countries, journals, authors, or institutions in different disciplines, and have become an accepted way to measure the productivity of the investments in R&D policies. Classical methods developed in bibliometry are still applied. For example, Bradford's law [6, 12] which proposed by Samuel Bradford in 1948 to study the distribution of the scientific literature, is widely used to assess the relevance of journals and authors. This law's initial purpose was help librarians identify those journals libraries should subscribe, in order to make a better investment of the budget available for acquisitions. In the area of innovation, Bradford analysis has been used with collections of patents to identify companies' productivity [19] or to identify the impact that scientific and technical journals have on the development of inventions [20].

Additional bibliometric indicators have been developed more recently and have gained popularity in the assessment of research productivity and impact. Among others, we should include the Hirsch Index (h-index) and the g-index. The first one quantifies the cumulative impact and relevance of the scientific output of an individual [21, 1] and is used to compare researchers in different areas [5]. It takes as an input the number of papers published by the author and

the citations they have received: a researcher has an index h of his papers have received at least h citations each. The second indicator, the g-index [10], developed by Leo Egghe with the aim of improving some limitations of the h-index [9], is defined as the largest number such that the top g papers together receive g^2 or more citations. These indicators do not provide only a ranking of researchers; they can also be used to identify the core intellectual products (papers, articles, patents, etc.) produced by a person or organization. In the area of innovation-related studies, h-index has been used in the analysis of citations within patents [7, 13, 22].

There are more bibliometric indicators that may be consider when analysing innovation-related data: citation speed index, Publication Efficiency Index (PEI), etc. The first one measures how fast recently published items are incorporated as inputs (that is to say, as citations) in new documents; PEI is often applied to assess the misalignments between the output achieved by an innovation or research program and the effort invested on it.

In the area of innovation, bibliometric indicators have been used on the analysis of patent collections to identify the companies and institutions that act as leaders in different areas of research. The number of granted patents reflects the companies' productivity and their capability to develop innovations. The citations in the patents are the input to assess their impact on subsequent research and quantify the impact of the different actors working in a particular domain [18, 17].

Bibliometrics make an extensive use of visualization techniques to present graphically the results obtained with the analysis of the data. As an example, the image below shows a chart representing the productivity and impact of different companies in the Information Retrieval Software industry. The values for productivity and impact are obtained by analysing the patents granted in a particular period.

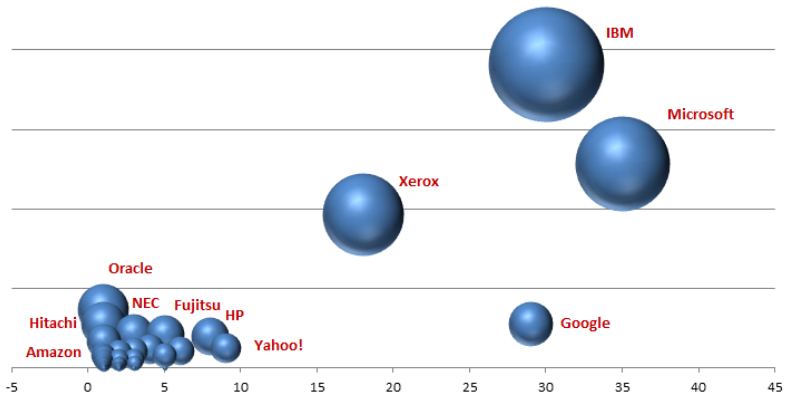

Fig. 1. Productivity and Impact Chart for a particular time period

Visual representations of bibliometric analysis are built on top of data that represents key variables of a document collection, like the number of items published by a company, the number of citations it receives, etc.

Innovation professional can benefit from these techniques to describe, analyse and assess the role that different persons, organisations, institutions, or geographical areas play in the innovation arena. Today, researchers have at their disposal different software tools to complete these bibliometric analyses. Most of them are freely distributed as open source and support visualisation features to display relationships between the data. Bibliometric analysis has usually been made with the help of general-purpose tools like Pajek, Cytoscape, or UCINET, which are mainly focused on network analysis and visualization. Other bibliometric tools support pre-processing, network creation, normalization, and analysis of data: BibExcel, CiteSpace, IN-SPIRE, Science of Science tool, VantagePoint, VOSviewer, and SciMAT [8]. Some of them are stronger in certain aspects of the data processing workflow, which leads to the possibility of combining different tools at different steps.

Bibliometric techniques are valid tools to know, represent, and assess the role that different companies, institutions, clusters, researches, or even countries have on the Research and Development landscape. They also provide a valid tool to identify potential partners in either traditional or Open Innovation programs. However, there are other areas where innovation analysis can benefit of Information Science techniques. One of them is the analysis of large collections of documents and the extraction of meaning.

2. Dealing with Unstructured data

Innovation professionals need to process data from different sources to identify trends in the market and in the activity of partners and competitors. These analyses require going into the details of the documents within the collection in order to extract their meaning.

The extraction of meaning from text is the subject of the Text Mining discipline. This is an application of the Information Retrieval and the Computational Linguistics aimed to help users identify and extract new knowledge from large collections of documents. Text mining is to discover new, previously unknown information from existing written resources with the assistance of computers. Definitions of text mining in the professional literature refer to the capability of identifying new knowledge from unstructured data: [17, p. 324], talked about "... any operation related to gathering and analysing text from external sources for business intelligence purposes [...] discovery of previously unknown knowledge in text" and "the discovery of previously unknown knowledge in text." Another recognized expert in text mining, Marti A. Hearst said that the purpose of text mining is "discovering information and

knowledge previously unknown, not explicitly present in any of the analysed documents."

These definitions are ambitious. To try to give a more pragmatic definition, it is convenient to consider that there is an intermediate objective that should be achieved in order to extract new knowledge from existing data: information available in large document collections must be processed and presented in a way that makes their understanding and comprehension easier. We could then define text mining as "*the processing of collection of documents to present their data in a way that makes easier their analysis by the staff in charge of making decisions. The presentation of the data should help identify relationships between the facts described in the documents*."

Under the term "text mining," we group together several software-based techniques, which include document classification and clustering, automatic summarization, feature extraction, and text visualization. Document clustering and automatic text classification are valuable techniques to make the initial pre-processing of documents. Both are used to create groups of documents with similar meaning; thereby, documents within the same group or class are represented by a surrogate or centroid made up of several terms. End users' queries are compared against these surrogates, and end users can explore large collections of documents by browsing through the groups or classes. Other text mining technique, feature extraction, identifies the facts described in the documents. Its most popular use has been the identification of personal names, organizations, dates and events, as well as the relationships between them. For example, it would be possible to extract, from an unstructured text, the relationship between a manufacturing method, a company, and a particular technology. Text summarization is another interesting technique that extracts those fragments of the text that are likely to give the best representation of its meaning.

Recent achievements in text mining are related to the capability of visualizing the data in individual documents and in document collections. Text visualization relies on the methods applied for feature and concept extraction. The graphical representations of the text help users understand the content of the documents under analysis. These techniques range from simple, popular "wordnets" to more complex visualization schemas that not only highlight the most important concepts in the document(s), but the relationships between them. This is possible by (a) the possibility of representing documents, sections, or fragments within them by means of vectors where each term is given a value based on its frequency in the document and in the collection as a whole, and (b) the analysis of co-occurrences of terms. Once we have this representation of the texts, attractive visualizations showing the relationships between concepts based on syntactic or statistics patterns are feasible. Different visualizations methods for textual content may be found in the web site: http://www.manyeyes.com.

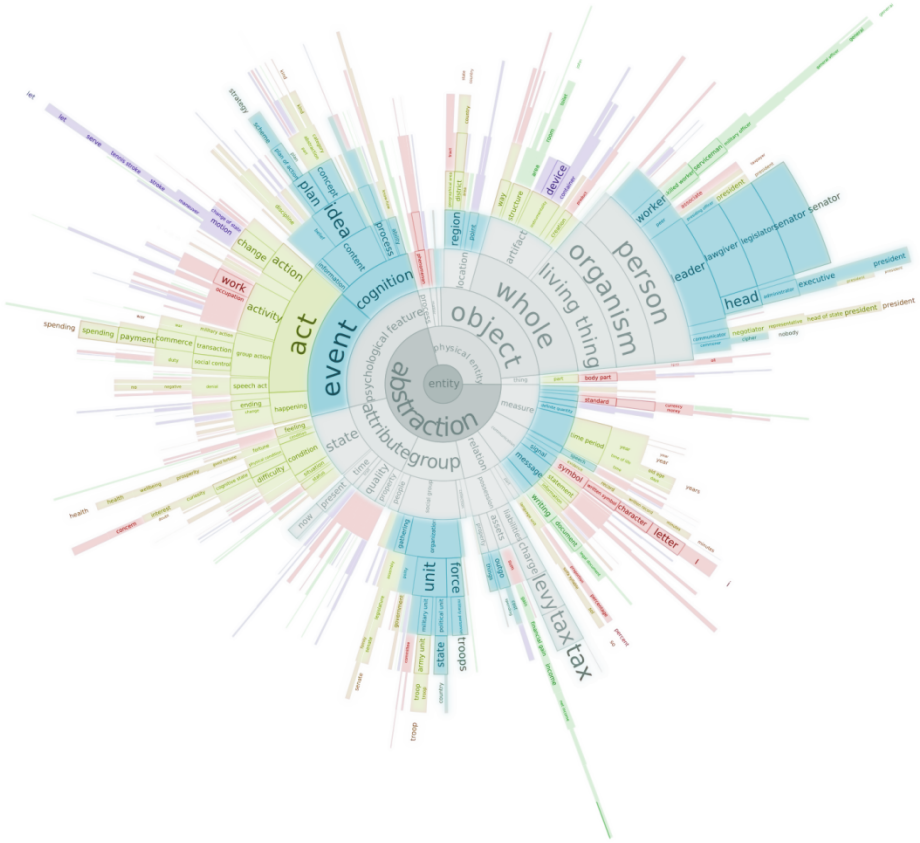

Fig. 2. Conceptual graph showing term relationships
Source: http://vialab.science.uoit.ca.

From the perspective of the analysts involved in innovation activities, these visualization schemas may be useful when approaching large sets of unstructured data. Potential usages include the identification of (a) research trends, (b) relationships between institutions and subjects of research, and (c) between research concepts, methods, and tools leading to innovative solutions through combination, etc.

Conclusions

Continuous monitoring of the environment is an activity that professionals involved in the planning, management and support of innovation programs cannot obviate. Data about the external environment is usually collected in the form of unstructured data, document sets, or textual inputs collected from web-based innovation platforms supporting idea collection or co-creation processes.

The assessment of the results achieved by innovation efforts is also an integral component of the activities under innovation professionals' responsibilities.

Techniques developed in the context of Information and Documentation Science may help innovation professionals in the analysts these data. The combined use of scientometric techniques and text analysis and visualization is a promising solution to deal with the inputs needed to plan, develop, and monitor innovation strategies. Scientometric provides tools to identify research trends and gain knowledge about the agents that are doing R&D and producing innovations. The impact of these innovations can be assessed through citation analysis and studies of their impact. Text analysis and visualization can be used to gain an understanding of the knowledge embodied in document collections and get a preliminary view of the main trends and concepts under discussion. Once recent, relevant application of text mining techniques, called "opinion mining," focuses on the analysis of opinions posted by users of Web 2.0 tools like Facebook or Twitter, and these same techniques might be applied in the analysis of web-based innovation platforms where users post and discuss potential improvements in existing products.

The combined use of bibliometric and text mining techniques is not new. For example, Hearst described a project where the impact of public funded research in the development of patents was studied. The study required the analysis of citations in a wide collection of patents and the identification of the funding agencies and sponsor organizations that made possible the research. Additional opportunities can be derived from the combined use of these techniques: bibliometric studies help gain a top-down view of the data, identifying agents and main research trends, and text mining and visualization techniques provide a bottom-up view of the data leading to the identification of the concepts under discussion and the complexity of their relationships.

References

1. Bornmann L., Werner M. (2011): The h index as a research performance indicator. European Science Editing; 37, pp. 77–80.
2. Busse M. et al. (2013): Innovation mechanisms in German precision farming. Precision Agriculture, pp. 1–24.
3. Bygstad B., Aanby H. (2010): ICT infrastructure for innovation: A case study of the enterprise service bus approach, Information Systems Frontiers, 12, pp. 257–265.
4. Chesbrough H. (2003): Open innovation: how companies actually do it. Hardware Business Review, 81(7), pp. 12–14.
5. Ciriminna R., Pagliaro M. (2013): On the use of the h-index in evaluating chemical research. Chemistry Central Journal; 7 (1), pp. 132.

6. Chung Y-K. (1994): Core International Journals of Classification Systems: An Application of Bradford's Law. Knowledge Organization; 21 (2), pp. 75–83.
7. Chung H.K., Mu-Hsuan H., Dar-Zen C. (2011): Ranking patent assignee performance by h-index and shape descriptors. Journal of Informetrics; 5 (2), pp. 303–312.
8. Cobo M.J., et al. (2012): SciMAT: A New Science Mapping Analysis Software Tool, Journal of the American Society for Information Science and Technology, 63, 8, pp. 1609–1630.
9. Costas R., Bordons M. (2007): The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. Journal of Informetrics; 1 (3), pp. 193–203.
10. Egghe L. (2006): Theory and practise of the g-index. Scientometrics; 69 (1), pp. 131–152.
11. García-Barriocanal E., Sicilia M.A., Sánchez-Alonso, S. (2012): Social Network-Aware Interfaces as Facilitators of Innovation. Journal of Computer Science and Technology, 27(6), pp. 1211–1221.
12. Garg K.C., Karki M.M.S., Krishnan Marg K.S. (1988): Bibliometric study of world literature on patents. World Patent Information; 10 (4), pp. 237–242.
13. Guan J.Ch. (2009): Exploring the h-index at patent level. Journal of the American Society for Information Science and Technology; 60 (1), pp. 35–40.
14. Hearst M. (1999): Untangling Text Data Mining. In Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20–26, 1999.
15. Harmelen F. van et al. (2012): Theoretical and technological building blocks for an innovation accelerator, The European Physical Journal, 214, pp. 183–214.
16. Iordatii M., Venot A., Duclos C. (2013): Designing concept maps for a precise and objective description of pharmaceutical innovations, BMC Medical Informatics and Decision Making, 13(10), pp. 1–8.
17. Jaffe A.B., Trajtenberg M. (eds.) (2002): Patents, Citations and Innovations: A Window on the Knowledge Economy. Cambridge Ma: London: The MIT Press Sullivan D. (2001): Document Warehousing and Text Mining. New York [etc.]: Wiley Computer Publishing, XVIII, pp. 542.
18. Karki M. (1997): Patent citation analysis: A policy analysis tool. World Patent Information, 19 (4), pp. 269–272.
19. Lo S. (2008): Patent coupling analysis of primary organizations in genetic engineering research. Scientometrics; 74(1), pp. 143–151.
20. Lo S. (2010): Scientific linkage of science research and technology development: a case of genetic engineering research. J Scientometrics; 82(1), pp. 109–120.
21. Norris M, Oppenheim C. (2010): The h-index: a broad review of a new bibliometric indicator. Journal of Documentation; 66 (5), pp. 681–705.

22. Zhang S. (2012): Exploring the nonlinear effects of patent H index, patent citations, and essential technological strength on corporate performance by using artificial neural network. Journal of informetrics; 6 (4), pp. 485–495.

Narzędzia zarządzania informacją dla analityków innowacji

Słowa kluczowe

Innowacja, naukometria, eksploracja tekstu, badanie opinii, wizualizacja tekstu.

Streszczenie

Zarządzanie innowacjami to oparty na wiedzy proces, w którym definiowany jest poziom zależności pomiędzy pomysłami, technikami i narzędziami mogącymi skutkować opracowaniem innowacji. Analityk innowacji musi zarządzać treściami niestrukturalnymi: pomysłami zgromadzonymi od pracowników jak i partnerów, wiedzą pochodzącą z publikacji naukowych i raportów technicznych, patentami i zgłoszeniami patentowymi itp. Wszystkie te źródła stanowią istotny wkład w proces oceny innowacyjności pomysłu, możliwości jego realizacji oraz konkurencyjności rynkowej.

W zarządzaniu innowacjami powszechnie stosowane są techniki i metody informatyczne, które wspomagają proces identyfikacji trendów, oceny rezultatów, oszacowania niezbędnych nakładów finansowych czy monitorowania rynku. Oznacza to, że należy regularnie monitorować stan wiedzy i techniki w tym obszarze w celu zapewnienia jak najbardziej owocnej współpracy na styku nauk informatycznych i zarządzania innowacjami. W ostatnich latach znaczący postęp osiągnięto w takich dziedzinach jak naukometria, wizualizacja tekstu i badanie opinii.

W artykule dokonano przeglądu tych technik i omówiono sposób, w jaki mogą one wspomóc specjalistów zaangażowanych w realizację innowacyjnych programów.