FACULTEIT ECONOMIE EN
BEDRIJFSWETENSCHAPPEN

KATHOLIEKE
UNIVERSITEIT
LEUVEN

IMPACT AND CONSEQUENCES OF SCIENCE-INTENSIVE PATENTING:

IN SEARCH OF ANTI-COMMONS EVIDENCE USING LATENT SEMANTIC ANALYSIS TEXT MINING TECHNIQUES

Proefschrift voorgedragen
tot het behalen van de graad
van Doctor in de Toegepaste
Economische Wetenschappen

door

Tom Magerman

Nummer 375 2011

Doctoral Committee

If you consider the contribution of plumbing to human life,
the other sciences fade into insignificance.
James Gorman

Advisor:

Prof. Dr. Ir. Koenraad Debackere	Katholieke Universiteit Leuven

Doctoral Committee:

Prof. Dr. Bart Van Looy	Katholieke Universiteit Leuven
Prof. Dr. Bart Baesens	Katholieke Universiteit Leuven
Prof. Dr. Sien Moens	Katholieke Universiteit Leuven
Prof. Dr. Francesco Lissoni	University of Brescia

Chairman:

Prof. Dr. Marleen Willekens	Katholieke Universiteit Leuven

Acknowledgements

Every day I remind myself that my inner and outer life are based on the labors of other men, living and dead, and that I must exert myself in order to give in the same measure as I have received and am still receiving.
Albert Einstein

This dissertation is the end of a pleasant – but sometimes rough – journey, starting fifteen years ago with obtaining my first university degree. In line with the subject of this dissertation – science-technology linkages – I wandered through different levels of university linkage, going back and forth from academic oriented work to more 'normal work' – for whatever reason family and friends do not regard academic work as 'work', let alone 'normal' – to eventually end up with this dissertation. When I took the decision a few year ago to 'go for it', many suspected moderate to substantial levels of madness, predicting dullness, loneliness, lack of women, poverty, disappointment and inevitable misery. Although they were right about some aspects – and there is probably something to say about my mental state – I'm very happy I did it. I have no clue where it will lead me to, and only the future will reveal whether I made the right choice, but so far I'm most happy with my journey, even if it sometimes looks as a random walk.

In that respect I want to thank Bart Van Looy to encourage me to take the jump and helping to create the circumstances to enable this endeavour. Although we have a long history of collaboration, the road was sometimes bumpy, so this opportunity was not for granted. But luckily his very careful planning, punctuality, pursuit for perfection, focus on efficiency and sense of organization makes us perfectly complementary. Special thanks to my advisor, Koenraad Debackere, to enable this endeavour. We also have a very long relationship – I think I was amongst his first employees in Leuven – and we kept collaborating, at least from a distance, but I realize that this next step was straightforward nor for granted. I'm very grateful for all opportunities I got from Koen, and most proud being part of his team. I hardly know anyone with such great professionalism mixed with outstanding skills, extraordinary performance and unlimited energy. Koen is the proof that the planet Krypton exists – and he loves good wine and good food too. Many thanks to my other committee members, Bart Baesens, Sien Moens and Francesco Lissoni, for all comments and advice. I'm fortunate to stand on the shoulders of giants.

Special thanks to professor Lambert Vanthienen and Cynthia Van Hulle, who blindly and unconditionally believed in my capacities and skills and hired me with great enthusiasm for my first job. I have great memories of those early days; the going was sometimes though, but we had awesome fun. Thank you very much to all those early day colleagues: Alain, Inge, Peter and Nick on the first floor, and later Edwin and Arnold on the third floor. One of the unpleasant aspects of getting older is the sense that things were much better in the past and will not come back. I dare to say that the collegiality and striving to excellence, combined with lots of fun, of those early days indeed complies with that feeling.

Thanks also to the more recent colleagues after my re-entry into academics. As working hours represent a big share of our available time, doing so in a pleasant environment makes it extra

stimulating. Petra, Julie, Bart, Cathy, Jan, Annelies, Mila, Cindy, Anneleen, Susanne, Sam, Maikel, Wim, and all the others I missed, thank you for the nice times throughout all those years, and special thanks to the Czarnitzki girls and boys for extending the stimulating environment outside the office hours. Very special thanks in that respect to Dirk for continuously challenging my absorptive capacity at Stapletons. All those brainstorming session at the Revue, Bierkelder and Seven Oaks kept me young at heart. And thanks to you all, the Belgians are very well respected at international summer schools, both for intellectual and less intellectual contributions. And thank you Susanne for not throwing an atomic bomb on Belgium.

Many thanks to Frizo Janssens, who was so kind to make his Matlab-code available to me to give me a jump start into text mining applications. Without his gesture I would have needed many additional weeks of programming.

Special thanks for all people who contributed to the validation of my datasets: Bart, Julie, Caro, Joris, Mariette and especially Hilde and Isabelle. Science is 20% excitement and 80% dull and patient labour. Without your contributions, no results.

Thanks for my friends of Leuven, who enable me to balance work and life, challenge the relevance of what I do and convince me that things are not better in the private sector, except the cheque at the end of the month.

Special thanks for my family. My mother and father who allowed me to study – taken for granted, but it is not – and supporting me in what I do (although my mother asks me from time to time when I will take a decent job – I'm afraid I will have to disappoint her once again in that respect). My brothers and sister, who always stood behind me and helped me in various ways; Bob for making me familiar with Pink Floyd, cars and beers; Jo for making me familiar with Genesis, wine and women, and especially Eva who always had to take care of me – gender issues you know – although she partially revenged: I have slight notions of being part of challenging experiments on the effects of suffocation with a pillow, the effects of poking a stick in a wasp's nest – I loved that one, my parents too – and last but not least a thrilling experiment on the influence of centrifugal force on the track of my favourite trolley – with me on top of it - towed by a bike at high velocity – to give a hint, it's fun as long as it is straight on, but far less fun when the trolley does not perfectly follows the track of the bike anymore in the turns. But what doesn't kill you makes you stronger; her contribution in the development of my scientific career and hence this dissertation is substantial: being part of her experiments made me sensitive to empirical hypothesis testing.

Last but not least many thanks to my partner, Veerle. Obtaining a doctorate is not a nine-to-five job and involved many evenings, weekends and nights. The complaint whether I was married with my computer was ubiquitous, and there was not always time left for her. Amazing how the dearest things in live gets the least attention - because you expect them to be there forever? Final thanks to Ella and Lola, my beautiful, cute, lovely and beloved – but challenging – daughters, a gift of nature, who make you realize there is so much more in life (and who were so kind to give me – mostly – quiet evenings and nights so the work could go on, as if they realized already at their age that science does not go well together with chaos, although currently it seems they are rapidly make up for the loss – time for another doctorate?).

Tom Magerman, October 2011

Table of contents

List of tables

List of figures

INTRODUCTION, BACKGROUND AND OVERVIEW OF THE DISSERTATION

1 Introduction and background.

If I have seen a little further [than you and Descartes]
it is by standing on the shoulders of Giants.
Isaac Newton, letter to Robert Hooke (originated from John of Salisbury)

1.1 Science-technology interactions, entrepreneurial universities and academic patenting

Starting from the land grant colleges (Morrill Acts of 1862 and 1890), university-industry links evolved from a pragmatic tool for knowledge transfer in an evolving society to a key concept in innovation dynamics in an evolving world. Land grant colleges were established to focus on the teaching of agriculture, science and engineering as a response to the industrial revolution and changing social class rather than higher education's historic core of the Liberal Arts[1], and are an example of the nature of early science-technology interactions with the separation between both spheres as we knew them for a long time: universities conduct research that is freely shared ('open science'), supply qualified scientist and engineers to industry and are largely public funded - industry funding is rather low and mostly related to open gifts rather than project contracts with clear directions and goals - while industry creates new products and processes, building on scientific evolutions whenever relevant.

From the eighties of the previous century onwards, a combination of factors changes the nature of university-industry links: change in legislation (Bayh-Dole act in the US and abolition of Professor's Privilege in other countries); institutionalized university-industry relationships (establishment of Technology Transfer Offices); development of the Industry/University Cooperative Research Centers Program by the US National Science Foundation; and large partnership deals between pharmaceutical companies and universities (Monsanto; Hoechst). Nowadays, more and more demand-inspired basic

[1] US Code Title 7, Chapter 13, Subchapter I, § 304

research is being conducted at universities, tearing down the traditional boundaries between science and industry - see e.g. Pasteur's Quadrant (Stokes, 1997). This phenomenon goes hand in hand with increasing industry funding for public R&D (see e.g. OECD 2009).

The traditional view of industrial research as an engine of economic growth (see e.g. Grossman & Helpman, 1991, for an overview) – derived from the notion of Creative Destruction of Shumpeter (1942) – is more and more combined with the view that the public stock of knowledge that accumulates from the spillovers of previous inventions is a fundamental input. The public-good aspects of knowledge create economy-wide increasing returns and success is partly achieved by 'standing upon the shoulders of giants': "If I have seen a little further [than you and Descartes] it is by standing upon the shoulders of Giants"[2] (Caballero & Jaffe, 1996).

Meanwhile, many scholars stress the role of science-technology interactions for the development of technological performance and international competitiveness and hence economic development and growth and welfare creation (see e.g. Freeman, 1987 and 1994; Lundvall, 1992; Nelson, 1993; Nelson & Rosenberg, 1993; Mansfield, 1995; Mansfield & Lee, 1996; Mowery & Nelson, 1999; Dosi, 2000), and the importance of the institutional framework (see e.g. the Triple Helix model: Leydesdorff & Etzkowitz, 1996 and 1998; Etzkowitz & Leydesdorff, 1997). They stress the role of science and the importance of interaction between a variety of institutional actors underlying the innovative capacity and consequent economic performance of an economical system. This more encompassing view on innovation dynamics has resulted in a growing popularity of the 'innovation system' concept which gained acceptance by scholars and policy makers alike as a guiding framework to understand innovation dynamics on an aggregated level (European Innovation Scoreboard, 2002).

In these models, knowledge generating institutions such as universities, research laboratories, industrial research centres and, more recently, government institutions are acknowledged - besides firms and entrepreneurs - as important and complementary

[2] Sir Isaac Newton, letter to Robert Hook, February 5, 1675. Newton's aphorism was popularized by Robert Merton, On the Shoulders of Giants (1965), but originated from John of Salisbury, 1159.

players in developing and stimulating the innovative capacity of a particular region or country.

There are multiple reasons why universities are relevant actors within innovation systems and can contribute to the national innovative capacity. First, research institutions produce information and ideas upon which the development of new products, processes and services can build. Secondly, research institutions can work on certain research agendas for a longer period of time, which can lead to the creation of new scientific insights. The latter can over time lead to economic applications. Notice in this respect that universities are well placed to address market failures that occur in the field of innovation (Arrow, 1962; Freeman, 1994; Baumol, 2002). Such market failures arise especially in relation to basic research, characterized not only by high levels of uncertainty both in terms of technical and commercial success, but also spanning long time frames to bear fruit (often decades). In addition, the nature of the outcomes of innovative activity - i.e. knowledge or information - complicates investment decisions even further (Foray, 2004). All these phenomena pose specific challenges for private investors, who tend to refrain from becoming involved in basic research activities. In order to avoid a loss of social welfare – due to non-investment behaviour of private firms – most national innovation systems nowadays invest considerably in basic research performed at universities and public research institutes.

As such, knowledge institutions like universities can play a specific role related directly to the potential these institutions possess to avoid technological lock-in phenomena. In order to continuously stimulate economic growth within a particular region or nation, based on knowledge intensive entrepreneurship, its technology portfolio should strike a balance between routine technological activities on the one hand (these are focused on process and incremental development in the more mature phases of the technology life cycle) and non-routine technological activities on the other hand (these are more focused on new technology platforms and fundamental developments). Local / regional knowledge centres, especially universities and research centres, can play a significant part in this respect. As they participate in high level scientific research, they contribute to the generation of new knowledge. Such research takes place in international

research communities. The exploration of new fields of knowledge – that can often not yet be categorized as routine activities – and the continued diffusion of this knowledge among regional actors can be considered an essential task of knowledge centres and especially universities. This double dynamic allows knowledge centres to play a fundamental role in regional innovation networks. These institutions are best placed to offer support in regard to the dual challenge of local and global knowledge development (Debackere, 2000; Van Looy, Debackere & Andries, 2003; Lester & Piore, 2004; Debackere & Veugelers, 2005). If a particular region fails to include this dual task as a priority in their regional innovation policy, there is a long-term risk of regression and growth stagnation due to the life cycle phenomenon. It is in this context that the significance of knowledge centres should be seen: they also develop non-routine activities in research communities which participate in knowledge exchange on an international scale. As such, universities offer regions exploration possibilities that are essential for mid to long-term innovation potential. Lester points in this respect to the importance for innovation of 'interpretative', problem defining activities, besides analytical, problem solving ones. When enterprises focus on the latter, it is essential that sufficient attention is paid to creating an environment for exploration. In this sense, universities, as fora where new ideas can be explored and studied, are indispensable.

These reflections also imply that universities are more effective in this respect as they are more active in scientific research. Recent research in the US as well as in Europe confirms this relation: an explicit research focus coincides with a larger number of enterprising activities (patents, spin-offs, contract research) (Di Gregorio & Shane, 2003; Van Looy, Ranga et al., 2004; O'Shea, Allen et al., 2005; Van Looy, Callaert & Debackere, 2006).

At the same time, contributing effectively to the innovative capacity of an innovation system requires a willingness of universities to become more 'entrepreneurial'. The notion of 'entrepreneurial universities' (Etzkowitz, Webster & Healy, 1998; Branscomb, Kodama & Florida, 1999) refers to the development of the following spectrum of activities: more intense commercialization of research results, patent and license activities, spin-off activities, collaboration projects with the industry, and greater

involvement in economic and social development. As such, one observes a 'second academic revolution' whereby education and research become complemented with service and valorisation activities aimed at transferring new scientific knowledge to economic activity realms.

Indeed, nowadays an increasing activity of academic researchers in exploiting their discoveries can be observed (Henderson, Jaffe & Trajtenberg, 1998; Thursby & Thursby, 2002; Meyer, Sinilainen & Utecht, 2003; Lissoni, Llerena et al., 2008) and university patents become an important – and visible - method of technology transfer (Basberg, 1987; Boitani & Ciciotti, 1990; Trajtenberg, 1990; Archibugi, 1992).

Interaction and exchange between academia and industry can result in positive aspects, both for the business partner (e.g. Zucker & Darby, 2001; Hall, Link & Scott, 2001; Faems, Van Looy & Debackere, 2005) and for the academic sector (e.g. realization of complementarities between applied and basic research – Azoulay, Ding & Stuart, 2009; generation of new research ideas – Rosenberg, 1998; attracting additional resources for (basic) research - Agrawal & Henderson, 2002). Additional benefits – when introducing intellectual property in scientific activities - can be found in the facilitation of the creation of a market for ideas and the ability of society to realize the commercial and social benefits of a given discovery (Kitch, 1977; Merges & Nelson, 1990; Gans & Stern, 2000; Arora, Fosfuri & Gambardella, 2004; Hellman, 2007; Murray & Stern, 2007).

At the same time some concerns arise due to the increasing commercialization of scientific activities undertaken by universities. Too much emphasis on (market) exploitation might negatively impact the quantity and quality of scientific research and change research orientation because of changing incentives (skewing problem: research topic decisions follow market demand and money). But also indirect effects get attention: shift of career choices of promising young graduate students and post-doctorals away from academia; increasing secrecy or delay of publication (demanded by industrial partners); and presence of an anti-commons effect (to many owners blocking the use of inventions). These concerns get particular attention because they might slow down the rate of innovation and long-term scientific and technological advancement might be traded in for short term benefits.

A crowding out-effect is suspected because of conflicts of interest: time-related limitations in combinations with changes in incentives – remuneration based on patenting activities as compared to remuneration schemas based on contributions to the scientific community (disclosure and contribution to cumulative learning and innovation) - might drift scientists away from traditional academic activities as teaching and conducting basic research. But crowding out-effects can be relieved, e.g. by a well-organized Technology Transfer Office to put off the burden of patent filing from the shoulders of the academic inventor (Hellman, 2007), resolving time constraints – not to mention potential positive effects of additional funding to attract additional post-doctoral scientists to eventually increase the output

Qualitative evidence suggest that patenting activities are direct byproducts of scientific efforts; patents and publications may pertain to a nearly identical set of research findings and the decision of whether or not to patent is more an ex-post decision and not part of the selection process to engage into a particular research trajectory (Agrawal & Henderson, 2002; Murray, 2002; Thursby, Thursby & Gupta-Mukherjee, 2007), although incentives of patenting activities could change the behavior of scientists away from incidental patenting as byproduct of research programs towards pursuing projects with commercial potential (skewing problem, Florida & Cohen, 1999) – which does not necessarily imply a causal relationship between patenting incentives and a shift to more applied research or research with commercial interest; maybe academic inventors get more interested in practical issues because of their contacts with industry, and insights encountered through interaction with industry might help them in their basic research (the survey of Siegel, Waldman & Link, 1999, reveals that 65% of researchers reported that interaction with industry influenced their research in a positive way).

While a complete crowding out of scientific activities by commercialization endeavours is considered as highly unlikely (Merton, 1968; Scotchmer, 2004; Thursby, Thursby & Gupta-Mukherjee, 2007), some scholars however do signal a (moderate) negative impact on the quality of research (Henderson, Jaffe & Trajtenberg, 1996; Trajtenberg, Henderson & Jaffe, 1997; Czarnitzki, Glänzel & Hussinger, 2009). At the same time, a majority of reported empirical findings report a positive relationship between patenting

and publication outcomes of academic researchers (Agrawal & Henderson, 2002; Van Looy, Ranga et al., 2004; Van Looy, Callaert & Debackere, 2006; Buenstorf, 2006; Breschi, Lissoni & Montobbio, 2007; Czarnitzki, Glänzel & Hussinger, 2007; Stephan, Gurmu et al., 2007; Fabrizio & Di Minin, 2008; Azoulay, Ding & Stuart, 2009).

1.2 The tragedy of the anti-commons

While most empirical evidence – at the level of individual scientists – reports a positive relationship between patenting activities and publication outcomes (quantity as well as quality), the shift from public funding to private funding is not only a mere replacement of funding sources, but does impose additional restrictions. One such restrictions is the delay or even expel of the publication of research results to safeguard commercial opportunities from the side of the private sponsor. This threatens the system of open science where scientists can build upon previously diffused knowledge resulting in a cumulative knowledge stock (Mukherjee & Stern, 2009; Murray, Aghion et al., 2009; Czarnitzki, Grimpe & Toole, 2011).

In relation to the issue of delay and secrecy, expansion of IPR might result in 'privatizing' the scientific commons and potentially limiting scientific progress (Argyres & Liebeskind, 1998; David, 2000; Krimsky, 2004). This fear is nicely expressed by the metaphor of the 'Tragedy of the anti-commons', introduced by Heller (Heller, 1998) as opposed to the 'Tragedy of the commons' of Hardin (Hardin, 1968). Heller states that the presence of too many owners with blocking power can lead to the underutilization of scarce resources, or, translated to the world of IPR, more intellectual property rights may lead paradoxically to fewer useful products instead of being an incentive to invent and disclose (too many owners hold rights in previous discoveries creating obstacles for future research). Although this phenomenon is induced by high transaction costs and can be transitional (market players have to learn to deal with each other or changing market circumstances), it is clear that it is not favourable for long-term technological and scientific development, although presence and impact will very amongst sectors and fields. E.g. patent anti-commons could prove more intractable in biomedical research than in other settings because of the importance of patents for the

biotechnology industry, the lack of substitutes for certain biomedical discoveries (rivals may not be able to invent around) and the heterogeneity of interests and resources among public and private patent owners (Heller & Eisenberg, 1998). Further development of entrepreneurial universities in general and academic patenting in particular risks to trade in long-term scientific development for short-term benefits of science-technology interactions; the blocking power of patent holders on knowledge from a scientific nature might prune promising new scientific development paths based on existent knowledge, radically changing the way science developed over the last centuries.

Although anecdotal evidence exists of problematic impact of IPR on scientific findings (e.g. the 'OncoMouse' or 'Havard mouse' of Leder and Stewart; and patents on human genes associated with breast and ovarian cancer owned by Myriad Genetics), large scale evidence of the presence of an anti-commons effect in biotechnology patenting is rare and the magnitude of the phenomenon and the real thread of patent thickets to block access to knowledge and technology is unclear.

E.g. when it comes to the potential effect on research direction – researchers abandoning lines of research because of limitations imposed by IPR - Walsh, Cohen & Cho (2007) observe that access to knowledge inputs is largely unaffected by patents. More problematic is access to materials and/or data possessed by other researchers. Restrictions on access, however, do not appear to turn on whether the material is itself patented. There is, hence, little evidence that patent policy is the direct cause of restricted access to tangible research inputs. Jensen & Webster (2010) find that transaction costs and the culture of the workplace have the largest influence over whether or not patents affect the direction of research.

Besides these studies based on surveys, Murray & Stern (2007) conducted an empirical study, using a dataset based on 'patent-paper pairs' in biotechnology to investigate the use of scientific publications (measured by forward citations) that are part of a patent application. They suggest a modest anti-commons effect based on a decline in citation rate – after granting of the patent – by 10 to 20% for a set of 169 so called 'patent-paper pairs' or 'patent-publication pairs' published in Nature Biotechnology between

1997 and 1999, although these authors also clearly point to the interpretation limits inherent to their study.

Criticism raises that the patent system stemming from centuries ago is inappropriate to deal with current day evolutions in innovation and knowledge development and a reform is needed to balance private and public interests. The debate on a patent reform is going on, both in the U.S. and in Europe. However, change in legislation, if at all, will be slow. But e.g. Hopkins, Mahdi et al. (2007) show that – in the area of DNA patenting – patent offices have responded to criticism by raising thresholds that make it much less attractive for applicants to file speculative, broad claims in the hope of obtaining what many would view as undue rewards. But legislation is evolving and many countries defined research exemptions excluding research and tests from patent infringement, although the scope of the exemption vary largely amongst countries and patent systems and is not always well defined.

To the extent an anti-commons effect exists, one can wonder whether IPR and the exploitation of scientific research in se is the problem, or the enforcement in specific circumstances and the behaviour of licensors (Walsh, Arora & Cohen, 2003; Murray, 2006). In that respect Van Overwalle (2010) proposes rules of contract to turn patent 'swords' into commonly shared assets. Legal measures may assist in narrowing down freedom of operation of patentees, like compulsory licenses and extending research exemption to diagnostic-use exemption. As the pace of changing statutory patent law might be slow, private, collaborative efforts like patent pools and clearinghouses might help dealing with those issues to lower down transaction costs and help creating an efficient market of technology.

It is clear that the importance of the consequences of the presence of an anti-commons effect on national innovation systems and eventful policy implications makes this topic particularly interesting for further research.

1.3 The use of patent data and non-patent references as indicator to study science-technology interactions

There is a long history of using patent data to understand and assess invention and innovation dynamics in both economics and science and technology studies (e.g. Schmookler, 1966; Jaffe, 1986; Griliches, 1990). While patent based indicators indeed have limitations – not all inventions are being patented, while at the same time the propensity to patent differs among industries – addressing issues of (sources of) economic growth, the competitive position of countries or companies and the dynamism of industrial structures, requires a profound insight into patterns of (differential) inventiveness. In this respect patent databases, and the indicators derived from it, are still one of the prime information sources given their coverage, transparency and accessibility. Griliches' observations – when surveying the state of the art on patent statistics 20 years ago – still seem to hold: "in this desert of data, patent statistics loom up as a mirage of wonderful plentitude and objectivity" (Griliches, 1990, p. 1661).

Not only have patent indicators become more and more institutionalized (see for instance OECD, 2001; NSF; EC Science and Technology Indicator Report), over the last decade one has witnessed refinements in terms of the use of patent indicators to assess technological activities and to examine relationships with (economical) performance. Important evolutions relate to the use of (patent) citations to differentiate the value of patents (Trajtenberg, 1987) as well analysing and mapping non-patent references – most of which are references to the scientific literature – found within patent documents (Narin & Noma, 1985; Collins & Wyatt, 1988; Van Vianen, Moed & Van Raan, 1990; Narin & Olivastro, 1992; Schmoch, 1993; Narin, Hamilton & Olivastro, 1997; McMillan, Narin & Deeds, 2000). These latter approaches can be related directly to the rising popularity of integrative notions like (1) scientific networks (Steinmueller**,** 1994; David, Foray & Steinmueller, 1997; Pavitt, 1997), (2) strategy and its concomitant structural analysis of industries and competitors (Porter, 1995), (3) evolutionary economic thinking (Nelson, 1995) and (4) a new vision on industry, academia and government interactions as encompassed by the 'Triple Helix' model (Etzkowitz &

Leydesdorff, 1997 and 1998; Leydesdorff & Etzkowitz, 1996 and 1998), as described in the first section of this introduction. Innovation effectiveness is conceived to an ever larger extent as stemming from (networks of) interactions unfolding amongst a variety of actors. These include companies as well as knowledge generating institutes like universities, while government agencies enact framework conditions in which such interactions can evolve effectively.

These conceptions of innovation dynamics almost naturally led to efforts to delineate and develop indicators that – at least partially – grasp the complex set of interactions between both activity realms. Among the potential candidates, the nature of references found in patent documents has received considerable attention. Whereas the pioneering work of Jaffe, Tratjenberg and Henderson (for an overview, see Jaffe & Tratjenberg, 2002) focused on the role of patent references and citations (e.g. as an indication of the value of patents), Narin and his colleagues pioneered the role and possible contribution of non-patent references. Studies in this field have been investigating the nature of the science-technology interaction that is implied by a citation link (e.g. Narin & Noma, 1985), the role of public science for developing technology (Narin, Hamilton & Olivastro, 1997), or still the frequency and nature of occurrence of such interactions in new emerging technology domains (Van Viaenen, Moed & Van Raan, 1990; Meyer, 2000; McMillan, Narin & Deeds, 2000; Verbeek, Callaert et al., 2002).

1.4 In search of additional indicators for more direct science-technology interactions

Although non-patent references can be used to study science-technology interactions at the macro level, this indicator falls short to describe direct science-technology interactions at the micro level. Meyer (2000), based on a limited number of patent cases studies, concludes that non-patent references should not be interpreted as signalling a direct – and unidirectional – link or influence from science to technology as sometimes suggested by the rhetoric of advocates who depict non-patent references as an indicator of science-technology interactions.

Tijssen (2001) points in a similar direction: non-patent references should not be seen as reflecting scientific sources leading directly to the invention, but rather be considered a general indicator of 'interaction' between science and technology (Tijssen, 2001, p. 39).

This is of particular interest for studies on the presence of an anti-commons effect where the identification of direct relations between scientific developments and patent protection is crucial. For small samples, manual matching can be used (e.g. Murray & Stern, 2007), but direct matching of patents and scientific publications requires specific methods when large samples are desired. An example of a current approach is matching patent inventor names and patentees with publication author names and affiliations to identify individual science-related patents. However, this approach is not easy to implement on a large scale: patentee name matching requires name cleaning and addressing problems of name changes, name variants and organization entity resolution (from research groups to faculties/departments and institutions/universities); inventor name matching requires dealing with homonyms and first names and middle names and initials. On top of that, inventor/author matching does not do the job alone as scholars can be very active in both patenting and publishing and these activities do not necessarily take place in the same research or technology (sub)field. Additional information is needed to identify direct relationships between a single patent and publication.

Developing new methods for the identification of direct science-technology interactions for larger scale studies of the presence of an anti-commons effect is a particular challenge and deserves further attention.

1.5 References

Agrawal, A. & **Henderson**, H. (2002). "Putting patents in context: Exploring knowledge transfer from MIT." *Management Science*, 48 (1) : 44-60.

Archibugi, D. (1992). "Patenting as an indicator of technological innovation: a review." *Science and Public Policy*, 19 : 357-368.

Argyres, N. S. & **Liebeskind**, J. P. (1998). "Privatizing the intellectual commons: universities and the commercialization of biotechnology." *Journal of Economic Behavior and Organization*, 35 (4) : 427-454.

Arora, A., **Fosfuri**, A. & **Gambardella**, A. (2004). *Markets for Technology. The Economics of Innovation and Corporate Strategy.* Cambridge/London: MIT press.

Arrow K. J. (1962). *Economic welfare and the allocation of resources for invention. The rate and direction of inventive activity: economic and social factors*. Princeton (NJ): Princeton University Press.

Azoulay, P., **Ding**, W. & **Stuart**, T. (2009). "The impact of academic patenting on the rate, quality and direction of (public) research output." *Journal of industrial economics*, 57 (4) : 637-676.

Basberg, B. L. (1987). "Patents and the measurement of technological change: A survey of literature." *Research Policy*, 16 (2-4) : 131-141.

Baumol, W. J. (2002). *The Free-Market Innovation Machine.* Princeton (NJ): Princeton University Press.

Boitani A. & **Ciciotti**, E. (1990). "Patents as Indicators of Innovative Performances at the Regional Level." In R. **Cappellin** & P. **Nijkamp** (Eds.), *The Spatial Context of Technological Development*. Aldershot: Avebury.

Branscomb, L. M., **Kodama**, F. & **Florida**, R. (1999). *Industrializing Knowledge: University-Industry Linkages in Japan and the United States.* London: MIT Press.

Breschi, S., **Lissoni**, F. & **Montobbio**, F. (2007). "The scientific productivity of academic inventors: New evidence from Italian data." *Economics of Innovation and New Technology*, 16 (2) : 101-118.

Buenstorf, G. (2006). "Commercializing Basic Science as a Competitor or Complement of Academic Accomplishment? The Case of Max Planck Directors." Max Plan Institute of Economics, Mimeo.

Caballero, R. J. & **Jaffe**, A. B. (1993). "How High Are the Giants' Shoulders: An Empirical Assessment of Knowledge Spillovers and Creative Destruction in a Model of Economic Growth." *NBER Macroeconomics Annual*, 8 : 15-74.

Collins, P. & **Wyatt**, S. (1988). "Citations in patents to the basic research literature." *Research Policy*, 17 : 65–74.

Czarnitzki, D., **Glänzel**, W. & **Hussinger**, K. (2007). "Patent and publication activities of German professors: An empirical assessment of their co-activity." *Research Evaluation*, 164 : 311-319.

Czarnitzki, D., **Glänzel**, W. & **Hussinger**, K. (2009). "Heterogeneity of patenting activity and its implications for scientific research." *Research Policy*, 38 (1) : 26-34.

Czarnitzki, D., **Grimpe**, C. & **Toole**, A. A. (2011). "Delay and Secrecy: Does Industry Sponsorship Jeopardize disclosure of Academic Research?" ZEW Discussion Paper No. 11-009.

David, P. A. (2000). "The digital technology boomerang: new intellectual property rights threaten global open science." Working Papers 00016, Stanford University, Department of Economics.

David, P. A., **Foray**, D. & **Steinmueller**, W. E. (1997). "The research network and the new economics of science: From metaphors to organizational behavior." In: A. **Gambardella** & F. **Malerba** (Eds.), *The Organisation of Innovative Activities in Europe*. Cambridge University Press.

Debackere, K. (2000). "Academic R&D as a Business: Context, Structure and Processes." *R&D Management*, 30 (4) : 323–329.

Debackere, K. & **Veugelers**, R. (2005). "The role of academic technology transfer organizations in improving industry science links." Research policy, 34 (3) : 321-342.

Di Gregorio D. & **Shane** S. (2003). "Why do some universities generate more start-ups than others?" *Research Policy*, 32 : 209-227.

Dosi, G. (2000). *Innovation, Organization and Economic Dynamics.* Cheltenham: Edward Elgar Publishers.

Etzkowitz, H. & **Leydesdorff**, L. (1997). "Introduction to special issue on science policy dimensions of the Triple Helix of university-industry-government relations." *Science and Public Policy*, 24 (1) : 2-5.

Etzkowitz, H., **Webster**, A. & **Healey**, P. (1998). *Capitalizing Knowledge: New Intersections of Industry and Academia*. State University of New York press.

European Innovation Scoreboard (2002). Cordis Focus, December 2002.

Fabrizio, K. R. & **Di Minin** A. (2008). "Commercialization the laboratory: Faculty patenting and the open science environment." *Research Policy*, 37 (5) : 914-931.

Faems, D., **Van Looy**, B. & **Debackere**, K. (2005). "Interorganizational collaboration and innovation: Toward a portfolio approach." *Journal of Product Innovation Management*, 223 : 238-250.

Florida, R. & **Cohen**, W. M. (1999). "Engine or infrastructure? The university role in economic development." In: **Branscomb**, L. M., **Kodama**, F. & **Florida**, R. (Eds.), *Industrializing Knowledge: University–Industry Linkages in Japan and the United States*: 589-610. London, MIT Press.

Foray, D. (2004). *Economics of Knowledge.* Cambridge, US / London, UK: MIT Press.

Freeman, C. (1987). *Technology Policy and Economic Performance.* London: Pinter.

Freeman, C. (1994). "The economics of technical change." *Cambridge Journal of Economics*, 18 : 463-514.

Gans, J. S. & **Stern**, S. (2000). "Incumbency and R&D incentives: Licensing the gale of creative destruction." *Journal of Economics and Management Strategy*, 9 (4) : 485-511.

Griliches, Z. (1990). "Patent statistics as economic indicators: A survey." *Journal of Economic Literature*, 28 : 1661–1707.

Grossman, G. M. & **Helpman**, E. (1991). *Innovation and Growth in the Global Economy.* Cambridge: The MIT Press.

Hall, B., **Link**, A. N. & **Scott**, J. T. (2001). "Barriers inhibiting industry from partnering with universities: Evidence from the advanced technology program." *Journal of Technology Transfer*, 26 (1-2) : 87-98.

Hardin, G. (1968). "Tragedy of commons." *Science*, 162 (3859) : 1243-1248.

Heller, M. A. (1998). "The Tragedy of the Anticommons." *Harvard Law Review*, 111 : 621-688.

Heller, M. A. & **Eisenberg**, R. S. (1998). "Can Patents Deter Innovation? The Anticommons in Biomedical Research." *Science*, 280 : 698-701.

Hellman, T. (2007). "The role of patents for bridging the science to market gap." *Journal of Economic Behavior and Organization*, 63 (4) : 624-647.

Henderson, R., **Jaffe**, A. B. & **Trajtenberg**, M. (1996). "The Bayh-Dole Act and trends in university patenting 1965-1988." *Proceedings of the Conference on University Goals, Institutional Mechanisms and the Industrial Transferabiliby of Research*.

Henderson, R., **Jaffe**, A. B. & **Trajtenberg**, M. (1998). "Universities as a source of commercial technology: A detailed analysis of university patenting." *Review of Economics and Statistics*, 80 (1) : 119-127.

Hopkins, M. M., **Mahdi**, S., **Patel**, P. & **Thomas**, S. M. (2007). "DNA patenting: the end of an era?" Nature Biotechnology, 25 (2) : 185-187.

Jaffe, A. B. (1986). "Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value." *American Economic Review*, 76 (5) : 984-1001.

Jaffe, A. B. & **Trajtenberg**, M. (Eds.) (2002). *Patents, citations and innovations*. Cambridge (MA): The MIT Press.

Jensen, P. H. & **Webster**, E. (2011). "Do patents influence academic scientists' choice of research projects?" University of Melbourne Working Paper No. 5/10.

Kitch, E. W. (1977). "The nature and function of the patent system." *Journal of Law and Economics*, 20 (2) : 265-290.

Krimsky, S. (2004). *Science and the Private Interest: Has the lure of profits corrupted biomedical research?* Rowman-Littlefield Publishing Co.

Lester, R. K. & **Piore**, M. J. (2004). *Innovation-The Missing Dimension.* Harvard University Press.

Leydesdorff, L. & **Etzkowitz**, H. (1996). "Emergence of a Triple Helix of University-Industry-Government Relations." *Science and Public Policy*, 23 (5) : 279-286.

Leydesdorff, L. & **Etzkowitz**, H. (1998). "Triple Helix of Innovation: Introduction." *Science and Public Policy*, 25 (6) : 358-364.

Lissoni, F., **Llerena**, P., **McKelvey**, M. & **Sanditov**, B. (2008). "Academic patenting in Europe: New evidence from the KEINS database." *Research Evaluation*, 17 (2) : 87-102.

Lundvall, B. A. (1992). *National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning.* London: Pinter Publishers.

Mansfield, E. (1995). "Academic research underlying industrial innovations: sources, characteristics, and financing." *The Review of Economics and Statistics*, 77 (1) : 55-56.

Mansfield, E. & **Lee**, J. Y. (1996). "The modern university: contributor to industrial innovation and recipient of industrial support." *Research Policy*, 25 : 1047-1058.

McMillan, S., **Narin**, F. & **Deeds**, D. (2000). "An analysis of the critical role of public science in innovation: The case of biotechnology." *Research Policy*, 29 : 1–8.

Merges, R. P. & **Nelson**, R. R. (1990). "On the complex economics of patent scope." *Columbia law Review*, 90 (4) : 839-916.

Merton, R. K. (1965). *On the Shoulders of Giants*. New York.

Merton, R. K. (1968). "The normative structure of science." In R. K. **Merton** (1979), *The sociology of science: Theoretical and Empirical Investigations.* Chicago (IL): University of Chicago Press.

Meyer, M. (2000). "Patent citations in a novel field of technology – What can they tell about interactions between emerging communities of science and technology." *Scientometrics*, 48 (2) : 151–178.

Meyer, M., **Sinilainen**, T. & **Utecht**, J. T. (2003). "Towards hybrid Triple Helix indicators: A study of university-related patents and a survey of academic inventors." *Scientometrics*, 58 (2) : 321-350.

Mowery, D. C. & **Nelson**, R. R. (1999). *Sources of Industrial Leadership*. Cambridge: Cambridge University Press.

Mukherjee, A. & **Stern**, S. (2009). "Disclosure or secrecy? The dynamics of open science." *International Journal of Industrial Organization,* 27 (3) : 449-462.

Murray, F. (2002). "Innovation as Co-evolution of Scientific and Technological Networks: Exploring Tissue Engineering." *Research Policy*, 31 : 1389–1403.

Murray, F. (2006). "The Oncomouse that roared: Resistance and accommodation to patenting in academic science." Sloan School of Management Working Paper.

Murray, F. E., **Aghion**, P., **Dewatripont**, M., **Kolve**, J. & **Stern**, S. (2009). "Of mice and academics: Examining the effect of openness on innovation." NBER Working Paper 14819. Cambridge (MA): National Bureau of Economic Research.

Murray, F. & **Stern**, S. (2007). "Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis." *Journal of Economic Behavior and Organization*, 63 : 648-687.

Narin, F., **Hamilton**, K. & **Olivastro**, D. (1997). "The increasing linkage between US technology and public science." *Research Policy*, 26 : 317–330.

Narin, F. & **Noma**, E. (1985). "Is technology becoming science?" *Scientometrics*, 7 : 369–381.

Narin, F. & **Olivastro**, D. (1992). "Status report: Linkage between technology and science." *Research Policy*, 21 : 237–249.

Nelson, R. R. (1993). *National Innovation Systems: A Comparative Analysis.* New York: Oxford University Press Inc.

Nelson, R. R. (1995). "Recent evolutionary theorizing about economic change." *Journal of Economic Literature,* 33 : 48-90.

Nelson, R. R. & **Rosenberg**, N. (1993). "Technical Innovation and National Systems." In R. R. **Nelson** (Ed.), *National Innovation Systems. A comparative Analysis.* New York: Oxford University Press, Inc.

OECD (2001). STI Scoreboard 2001. Paris: OECD Publishing.

OECD (2009). "Business-funded R&D in the higher education and government sectors." In *OECD Science, Technology and Industry Scoreboard 2009.* Paris: OECD Publishing.

O'Shea, R. P., **Allen**, T. J., **Chevalier,** A. & **Roche**, F. (2005). "Entrepreneurial Orientation, Technology Transfer and Spinoff Performance of U.S. Universities. " *Research Policy*, 34 (7) : 994-1009.

Pavitt, K. (1997). "Do patents reflect the useful research output of universities?" SPRU Electronic Working Papers Series, 6, November 1997.

Porter, M. (1995). *The Competitive Advantage of Nations*. New York: The Free Press.

Rosenberg, N. (1998). "Chemical engineering as a general purpose technology." In E. **Helpman** (Ed.), *General Purpose Technologies and Economic Growth.* MIT Press.

Schmoch, U. (1993). "Tracing the knowledge transfer from science to technology as reflected in patent indicators". *Scientometrics*, 26 (1) : 193–211.

Schmookler, J. (1966). *Invention and Economic Growth.* Cambridge (MA): Harvard University Press.

Schumpeter, J. (1942). *Capitalism, Socialism and Democracy*. New York: Harper.

Scotchmer, S. (2004). *Innovation and Incentives*. MIT Press.

Siegel, D. S., **Waldman**, D. & **Link**, A. (1999). "Assessing the impact of organizational practices on the productivity of university technology transfer offices: an exploratory study." *Research Policy,* 32 : 27–48.

Steinmueller, W. E. (1994). "Basic research and industrial innovation." In: M. **Dodgson** & R. **Rothwell** (Eds.), *The Handbook of Industrial Innovation*: 54-66. Aldershot: Edward Elgar.

Stephan, P., **Gurmu**, S., **Sumell**, A. & **Black**, G. (2007). "Who's patenting in the university? Evidence from the survey of doctorate recipients." *Economics of Innovation and New Technology*, 16 (2) : 71-99.

Stokes, D. (1997) Pasteur's Quadrant: Basic Science and Technological Innovation. Washington: The Brookings Institute.

Thursby, J. G. & **Thursby**, M. C. (2002). "Who is selling the Ivory Tower? Sources for growth in university licensing." *Management Science*, 48 (1) : 90-104.

Thursby, M., **Thursby**, J. & **Gupta-Mukherjee**, S. (2007). "Are there real effects of licensing on academic research? A life cycle view." *Journal of Economic Behavior and Organization*, 63 (4) : 577-598.

Tijssen, R. (2001). "Global and domestic utilization of industrial relevant science: Patent citation analysis of science-technology interactions and knowledge flows." *Research Policy*, 30 : 35-54.

Trajtenberg, M. (1987). "Patents, Citations and Innovations: Tracing the Links." NBER Working Paper 2457. Cambridge (MA): National Bureau of Economic Research.

Trajtenberg, M. (1990). "A Penny for your quotes: Patent citations and the value of innovations." *Rand Journal of Economics*, 21 (1) : 172-187.

Trajtenberg, M., **Henderson**, R. & **Jaffe**, A. (1997). "University versus corporate patents: A window on the basicness of invention." *Economics of Innovation and New Technologies*, 5 (1) : 19-50.

Van Looy, B., **Callaert**, J. & **Debackere**, K. (2006). "Publication and patent behavior of academic researchers: Conflicting, reinforcing or merely co-existing?" *Research policy, 35* (4) : 596-608.

Van Looy, B., **Debackere** K. & **Andries** P. (2003). "Policies to stimulate regional innovation capabilities via university-industry collaboration: an analysis and assessment." *R&D Management*, 33 (2) : 209-229.

Van Looy, B., **Ranga**, M., **Callaert**, J., **Debackere**, K. & **Zimmermann**, E. (2004). "Combining entrepreneurial and scientific performance in academia: towards a compounded and reciprocal Matthew effect?" *Research Policy*, 33 : 425–441.

Van Overwalle, G. (2010). "Turning Patent Swords into Shares." *Science*, 330 (6011) : 1630-1631.

Van Viaenen, B., **Moed**, H. & **Van Raan**, A. (1990). "An exploration of the science base of recent technology." *Research Policy*, 19 : 61–81.

Verbeek, A., **Callaert**, J., **Andries**, P., **Debackere**, K., **Luwel**, M. & **Veugelers**, R. (2002). "Science and Technology Interplay – A Modelling Approach on a Regional Level" Final Report to the EC DG Research, Brussels.

Walsh, J. P., **Arora**, A. & **Cohen**, V. M. (2003). "Science and the law: Working through the patent problem." *Science*, 299 (5609) : 1021-1021.

Walsh, J. P., **Cohen**, W. P. & **Cho**, C. (2007). "Where excludability matters: Material versus intellectual property in academic biomedical research." *Research Policy*, 36 : 1184–1203.

Zucker, L. G. & **Darby**, M. R. (2001). "Capturing technological opportunities via Japan's star scientists: Evidence from Japanese patents and products." *Journal of Technology Transfer*, 26 (1-2) : 37-58.

2 Research questions and overview of the dissertation.

What gets us into trouble is not what we don't know.
It's what we know for sure that just ain't so.
Mark Twain

2.1 Core topic : Importance and (potentially negative) consequences of science-intensive patents

In the introduction chapter we pointed already to the importance of science-technology interactions for the scientific and technological development and welfare creation. The interplay between technological and scientific realms is increasingly considered as essential for being effective in terms of knowledge creation, technology development and its translation into economic activity, especially for new emerging, knowledge intensive fields of economic activity (Gibbons, Limoges et al., 1994; Meyer-Krahmer, 2000; Tijssen, 2001).

In this dissertation we want to elaborate on the importance and especially the (potentially negative) consequences of the science-intensity of patents as one aspect of the phenomenon of science-technology interactions and 'scientification' of technological development. First we want to study the impact of the science-intensity of patents on the effectiveness of technological development to shed some light on the importance of science-intensity of patents. To the extent that there is a positive relationship between science-intensity of patents and technological development, strategies or policies to increase the science-intensity of patents makes sense and should be reinforced. One way to do so is to stimulate direct interactions between academia and industry and embrace the model of the 'entrepreneurial university' (see introduction chapter).

At the same time we must not be blind for potential pitfalls and recognize the threats of increasing commercialization of scientific activities and introduction of intellectual property rights in scientific activities. As described in the introduction chapter, most empirical evidence however reports a positive relationship between patenting activities of universities and publication outcomes, but the potential presence of an anti-commons effect, limiting scientific progress in the long-term, remains an open topic for discussion. Hence, in this dissertation we also want to check for the presence of an anti-commons effect as potential negative consequence of science-intensive and academic patenting.

The latter however requires the development of new techniques or methodologies to identify direct science-technology interactions as present indicators and techniques fall short to efficiently identify patent-publication matches on a large scale.

2.2 Data : Biotechnology patents and publications

We choose biotechnology as field under study throughout this dissertation because it is an active field creating big expectations in terms of development of new economic activities and welfare creation, and because it can be labelled as an industry in which the interplay between science and technology is important. From the seventies onwards scientific findings have been playing an important role within the industry (McMillan, Narin & Deeds, 2000) resulting as well in numerous studies focusing on the role of collaboration and networking (Deeds & Hill, 1996; Baum, Calabrese & Silverman, 2000; Rothaermel & Deeds, 2004), including science-technology linkage which is particularly strong in this field (e.g. Narin & Noma, 1985; Murray, 2002; Verbeek, Callaert et al., 2002).

As stated by Heller and Eisenberg (1998), patent anti-commons could prove more intractable in biomedical research than in other settings because of the importance of patents for the biotechnology industry, the lack of substitutes for certain biomedical discoveries (rivals may not be able to invent around) and the heterogeneity of interests and resources among public and private patent owners. This makes biotechnology a highly relevant choice as data source for our research questions.

Datasets used across the distinct parts of this dissertation vary but start from a common selection process to compile a large and exhaustive set of biotechnology patents and publications for a broad range of countries and years. We select EPO and USPTO biotechnology patents from *PATSTAT* (EPO Worldwide Patent Statistical Database) based on the *WIPO International Patent Classification* system using the OECD biotechnology classification (OECD 2005 and 2009) and we select biotechnology publications from the *WOS* database (Thomson Reuters ISI Web of Science) based on the subject areas as defined by the Web of Science

2.3 Research question 1 : Relation between science-intensity of patents and technological development

One aspect of science-technology interactions is the 'scientification' of technological development and increasing science-intensity of patents. One way to observe this phenomenon is by looking at the number of non-patent references pointing to scientific literature: the number of non-patent references doubled for USPTO biotechnology patents from 1991 to 2005. As mentioned in the introduction chapter, these references must not be seen as a direct link from science to technology, but are nevertheless an indication of science-technology interaction. The presence of scientific research in the 'prior art' description of a patented invention is seen as an indicator of the 'distance' between scientific findings on the one hand and technology development on the other hand. As references to be found in patents are a reflection of prior art, more references towards science fields signal more relevant prior art derived from scientific sources. While this does not equal a unidirectional, influencing or contributing, link from the cited paper towards the citing patent, it is clear that the more scientific references are considered relevant for assessing and contextualizing the claims made within the patent, the closer the technology is situated to scientific activity.

Studies addressing the relationship between science-intensity of patents – as measured by the amount of non-patent references – on the one hand and the effectiveness of technology development on the other hand are not frequent to be found. At a country level, Van Looy, Zimmermann et al. (2003) examined the relationship between the science-intensity of patents and technological performance (productivity, revealed

technological advantage). Positive relationships have been observed for so-called high tech domains like biotechnology, pharmaceuticals, organic fine chemistry and semiconductors, while for other domains – including chemistry, food chemistry, measurement and control technology but also telecommunications – no relationship was found. These observations led the authors to conclude that the relevancy of science-intensity when developing technology is a domain specific phenomenon (see in this respect also Narin & Olivastro, 1992, and Grupp & Schmoch, 1992).

At the same time it can be observed that within the aforementioned analysis, indicators pertaining to the scientific capabilities of a country have not been taken into account. As such, the observed positive relationships might stem from the presence of scientific capabilities; in this case one would merely be counting 'spillover' effects that could be assessed equally by established bibliometric indicators pertaining to scientific publications.

In this dissertation we want to contribute to the research on the effects of science-intensive patents and want to address following research question: does the science-intensity of patents – as measured by the amount of non-patent references – relates to technological performance at the country level when scientific capabilities are brought into the question?

Based on a large set of biotechnology patents and publications of 20 countries, we investigate the relationship between science-intensity of patents (measured by the amount of non-patent references) and technological productivity (measured by the number of patent grants divided by the total population) controlling for scientific productivity (measured by the number of scientific publications divided by the total population).

2.4 Methodological part : Assessment of Latent Semantic Analysis (LSA) text mining techniques to map patent and scientific publication documents

To the extent that there is a positive relationship between science-intensity of patents and technological development, and hence general welfare creation, we want to check

for the presence of an anti-commons effect due to the 'scientification' of technological development.

As mentioned in the introduction chapter, a major challenge for the study of the presence of an anti-commons effect, and in studies on science-technology interactions in general, is the identification of science-related patents in general and the identification of scientific results protected by intellectual property rights (IPR) in particular to understand the magnitude and characteristics of the phenomenon. For broader or high-level studies at the level of countries, sectors or technologies, the matching of non-patent references with databases with bibliographic data or scientific publications might yield valuable insights. For more low-level studies or direct science-technology interactions, current approaches involve the use of the number of non-patent references on patent documents, or, as described in the previous introduction chapter, matching patent inventor names and patentees with publication authors and affiliations. The former approach based on the number of non-patent references is easy to conduct on a large scale but suffers from the vagueness of the value of a non-patent reference as an indicator of science-intensity or science-relatedness (see introduction chapter and e.g. Callaert, Van Looy et al., 2006). The latter approach based on patentee and inventor name matching allows identification of direct interactions, but is not easy to implement on a large scale as described in the introduction chapter.

A promising new approach involves text mining to directly match text documents based on their contents to find patent and publication documents that are related by the topics they address, the methods they use, the results they obtain and the inventions or discoveries they address, as this might allow (semi)-automated compilation of large datasets based on content similarity. In general this could be instrumental for, amongst others, domain studies, trend detection/emerging field detection and science-technology linkage and thus contribute to technology and innovation research. At this moment, we are particularly interested in this text mining approach to identify patents related to scientific publications based on their shared contents and especially to check for documents with identical contents to identify scientific publications protected by

patents, allowing to compile large datasets to check for the presence of an anti-commons effect.

In this dissertation we want to contribute to the development of new techniques and indicators to detect links between patents and scientific publications. We conduct a thorough assessment of the Latent Semantic Analysis (LSA) text mining method and its options (pre-processing, weighting, …) to grasp similarities between patent documents and scientific publications. We want to assess effectiveness (in terms of precision and recall) and derive best practices on weighting and dimensionality reduction for application on patent data, given the technical and juridical nature and hence different linguistic context of patent and scientific publication documents.

The results of this methodological part will be used for the identification of patent-publication pairs as input for the part devoted to our second research question about the presence of an anti-commons effect.

2.5 Research question 2 : In search of anti-commons evidence

As stated in the introduction chapter, the presence of an anti-commons effect is not favourable for long-term technological and scientific development. Further development of entrepreneurial universities risks to trade in long-term scientific development for short-term benefits of science-technology interactions. The blocking power of patent holders on knowledge from a scientific nature might prune promising new scientific development paths based on existent knowledge, radically changing the way science developed over the last centuries.

In this dissertation we want to contribute to the research on an anti-commons effect and want to address following research question: are scientific publications from which the contents is part of a patent application less used, on average, in further technological and scientific developments compared to scientific disclosure not protected by intellectual property rights?

Again based on a large set of biotechnology patents and publications we first identify patent-publication pairs, i.e. scientific publications from which the content (subject,

methodology, findings, discovery) is subject of a patent application. Next we compare forward citation rates of those patents and publications involved in patent-publication pairs with patents and publications not involved in patent-publication pairs.

2.6 Overview of the dissertation

After this introduction and overview part we continue with a first empirical part devoted to our first research question: chapter 3 contains the starting point of our journey and examines the impact of the science-intensity of patents on the technological performance of countries in the field of biotechnology using non-patent references as indicator of science-intensity.

Because of limitations of current indicators and methods like non-patent references and inventor name matching, we continue with a methodological part devoted to the development of new techniques and indicators to detect (direct) links between patents and scientific publications: chapter 4 contains a brief introduction to text mining and the Latent Semantic Analysis (LSA) method. Next we present the results of an explorative study of the use of LSA to detect similarity between patent documents and scientific publications for small datasets (individual academic inventors) in chapter 5 and continue with the results of large-scale assessment of LSA to map patent and scientific publication documents and derive a method for the identification of patent-publication pairs in chapter 6. This methodological part is instrumental for our final goal to detect anti-commons evidence.

Next we move to the second empirical part devoted to our final research question about potential pitfalls: chapter 7 contains the analysis of citation flows of patent-publication pairs to search for evidence of an anti-commons effect.

We finalize this dissertation with a part with summary and conclusions and directions for further research: chapter 8 contains conclusions on the methodological part and chapter 9 and on the empirical part.

2.7 References

Baum, J. A. C., **Calabrese**, T. & **Silverman**, B. S. (2000). "Don't go it alone: Alliance network composition and startups' performance in Canadian biotechnology." *Strategic Management Journal*, 21 (3) : 267–294.

Callaert, J., **Van Looy**, B., **Verbeek**, A., **Debackere**, K. & **Thijs**, B. (2006). "Traces of prior art: A systematic analysis of other references found within the USPTO and EPO patent system." *Scientometrics*, 69 (1) : 3-20.

Deeds, D. L. & **Hill**, C. W. (1996). "Strategic alliances and the rate of new product development: An empirical study of entrepreneurial biotechnology firms. " *Journal of Business Venturing*, 11 : 41–55.

Gibbons, M., **Limoges**, C., **Nowotny**, H., **Schwartzman**, S., **Scott**, P., & **Trow**, M. (1994). *The new production of knowledge: The dynamics of science and research in contemporary societies*. London: Sage.

Grupp, H. & **Schmoch**, U. (1992). "Perception of scientification of innovation as measured by referencing between patents and papers". In: H. **Grupp** (Ed.), *Dynamics of Science-Based Innovations*: 73-128. Berlin/Heidelberg: Springer Publishers.

Heller, M. A. & **Eisenberg**, R. S. (1998). "Can Patents Deter Innovation? The Anticommons in Biomedical Research." *Science*, 280 : 698-701.

McMillan, S., **Narin**, F. & **Deeds**, D. (2000). "An analysis of the critical role of public science in innovation: The case of biotechnology." *Research Policy*, 29 : 1–8.

Meyer-Krahmer, F. (2000). New Modes of Knowledge Production: Science-based Technologies and Interdisciplinarity - Consequences for R&D Infrastructure and STI Policy. Europolis Workshop, Lisbon, 5-6 Juni.

Murray, F. (2002). "Innovation as Co-evolution of Scientific and Technological Networks: Exploring Tissue Engineering." *Research Policy*, 31 : 1389–1403.

Narin, F. & **Noma**, E. (1985). "Is technology becoming science?" *Scientometrics*, 7 : 369–381.

Narin, F. & **Olivastro**, D. (1992). "Status report: Linkage between technology and science." *Research Policy*, 21 : 237–249.

OECD (2005). A framework for biotechnology statistics. Paris: OECD publishing.

OECD (2009). OECD Biotechnology Statistics. Paris: OECD publishing.

Rothaermel, F. T. & **Deeds**, D. L. (2004). "Exploration and exploitation alliances in biotechnology: A system of new product development." *Strategic Management Journal*, 25 : 201–221.

Tijssen, R. (2001). "Global and domestic utilization of industrial relevant science: Patent citation analysis of science-technology interactions and knowledge flows." *Research Policy*, 30 : 35–54.

Van Looy, B., **Zimmermann**, E., **Veugelers**, R., **Mello**, J. & **Debackere**, K. (2003). "Do science-technology interactions pay off? An exploratory investigation of 10 science intensive fields." *Scientometrics*, 57 : 335-367.

Verbeek, A., **Callaert**, J., **Andries**, P., **Debackere**, K., **Luwel**, M. & **Veugelers**, R. (2002). "Science and Technology Interplay – A Modelling Approach on a Regional Level" Final Report to the EC DG Research, Brussels (also forthcoming in the EC Indicators report 2003).

3 Developing technology in the vicinity of science : An examination of the relationship between science-intensity of patents and technological productivity within the field of biotechnology[3].

There are no such things as applied sciences, only applications of science.
Louis Pasteur

3.1 Introduction

The starting point of this dissertation is the question about the impact of the science-intensity of patents on technological development. Studies addressing the relationship between science-intensity of patents – as measured by the amount of non-patent references – on the one hand and the effectiveness of technology development on the other hand are less frequent to be found. At a country level, Van Looy, Zimmermann et al. (2003) examined the relationship between the science-intensity of patents and technological performance (productivity, revealed technological advantage). Positive relationships have been observed for so-called 'high tech' domains like biotechnology, pharmaceuticals, organic fine chemistry and semiconductors, while for other domains – including chemistry, food chemistry, measurement and control technology but also telecommunications – no relationship was found. These observations led the authors to conclude that the relevancy of science-intensity when developing technology is a domain specific phenomenon (see in this respect also Grupp & Schmoch, 1992; and Narin & Olivastro, 1992).

At the same time it can be observed that within the aforementioned analysis, indicators pertaining to the scientific capabilities of a country have not been taken into account. As such, the observed positive relationships might stem from the presence of scientific

[3] The study as described in this chapter has been published in Scientometrics (Van Looy, Magerman & Debackere, 2007).

capabilities; in this case one would merely be counting 'spillover' effects that could be assessed equally by established bibliometric indicators pertaining to scientific publications. Hence, in order to assess the relevancy of using non-patent references as an (additional) indicator to explain differences in technological performance, further analysis – whereby scientific capabilities are taken into account – is required. It is in this area that we want to situate this contribution. Within the biotechnology domain – identified as a field in which the relationship between science and technology is intimate (see for instance, McMillan, Narin & Deeds, 2000; and Van Looy, Zimmermann et al., 2003) – the following research questions are being addressed:

- Does the science-intensity of patents – as measured by the amount of non-patent references – relates to technological performance (country level) when scientific capabilities are brought into the equation?
- If the relationship between scientific capabilities and technological activity over time is to be conceived as bi-directional, what role does science-intensity play in this respect?

Within this chapter we analyse biotechnology patents granted within the USPTO patent system covering the time period 1992–2001. In line with the concept of national innovation systems, countries are acting as the unit of analysis (Nelson, 1993). In the next section we discuss the concepts and indicators used within this analysis, which will allow us to present and discuss the findings obtained. However, before turning to the empirical analysis, we first discuss in more detail the nature of non-patent references found within patents as these references play a central role in the analysis undertaken.

3.2 A closer look at non-patent references

As became clear in the introduction chapter, considerable attention has been paid recently to the analysis of non-patent references. At the same time, some concerns about the exact meaning of these references have been uttered. Meyer (2000a), based on a limited number of patent cases studies, concludes that non-patent references should not be interpreted as signalling a direct – and unidirectional – link or influence

from science to technology, as sometimes suggested by the rhetoric of advocates who depict non-patent references as an indicator of science-technology interactions.

Tijssen (2001) points in a similar direction: non-patent references should not be seen as reflecting scientific sources leading directly to the invention, but rather be considered a general indicator of 'interaction' between science and technology (Tijssen, 2001, p. 39). A closer look at the specific role references to prior art play within the patent application is in this respect highly informative. Within the next paragraphs we focus on the USPTO legislation and procedures as the data used within this analysis pertain to USPTO. For an exhaustive overview with an emphasis on EPO, we refer to Michel & Bettels (2001). At the same time, it can be noticed that the main difference between both systems with respect to citing prior art relates to the amount of references to be found - due to different disclosure obligations imposed on applicants - rather than its nature.

Patents are documents issued by an authorized governmental agency which grants the applicant the right to exclude others to produce or use a specific new device, apparatus or process for a limited time period. Patents are granted to the applicant/assignee after an examination that focuses on the novelty, inventive activity and industrial applicability. During the granting process, patent examiners review the prior art pertaining to the invention. While applicants are obliged – within the USPTO examination process – to provide an overview of all known relevant prior art - which can be either patents or other written documents - patent examiners do not limit themselves to the prior art signalled by inventors and/or applicants. Based on information, archives and databases available, patent examiners in the end decide which references are relevant to assess the claims made. The examiners references are used to decide on granting, including restricting claims, and are to be found on the front page of patent documents, besides information pertaining to the invention, assignee(s) and inventor(s). These references do not necessarily coincide with references provided by the applicant; references provided by the applicant can be omitted while examiners might add additional references as well.

Stated otherwise, front page references as found in patent documents are being introduced during the examination process for the purpose of evaluating the novelty and inventiveness of the claims and their applicability, including contextualizing the claims that are being made. As can be read in the patent examining procedure manual, "The basic purpose for citing prior art in patent files is to inform the patent owner and the public in general that such patents or printed publications are in existence *and should be considered when evaluating the validity of the patent claims*. Placement of citations in the patent file along with copies of the cited prior art will also ensure consideration thereof during any subsequent reissue or re-examination proceeding." (*Manual of Patent Examining Procedure*, USPTO, italics added; see also, 35 U.S.C. 301 and 37 CFR 1.501).

It is clear that the specific role of references within patent application procedures is to some extent different from the role references or citations play within scientific publications. Within articles, references indicate sources of influence or serve as reference points to delineate differences (novelty). At the same time, references to previous work in scientific publications are introduced by the authors (sometimes with some support of reviewers), implying in general that the cited references are known to the author(s) and hence have had a certain influence on the genesis of the ideas and insights developed within the article or paper. This clearly is not necessarily the case for the front page references to be found within patent documents. References might be added by examiners without the applicants being aware of their presence or without this knowledge having influenced in any way the creation of the invention, as documented recently by Meyer (2000a) and Tijssen, Buter & Van Leeuwen (2000).

Against this background, a citation is perceived here as a bit of information linking two different documents. The presence of scientific research in the 'prior art' description of a patented invention is seen as an indicator of the 'distance' between scientific findings on the one hand and technology development on the other hand. As references to be found in patents are a reflection of prior art, more references towards science fields signal more relevant prior art derived from scientific sources. While this does not equal a unidirectional, influencing or contributing, link from the cited paper towards the citing

patent, it is clear that the more scientific references are considered relevant for assessing and contextualizing the claims made within the patent, the closer the technology is situated to scientific activity. As such it can be noticed that some of the debate around the nature and meaning of non-patent references arises from neglecting the precise role non-patent references fulfil within the patent procedure. Rather than equalling non-patent references as signalling direct or unidirectional influences - contributing to the genesis and development of the invention at hand - they are part of the context in which the patent and its claims are to be situated. Hence, the more scientific references are to be found within patents, the more technology development is considered here as situated in the neighbourhood or vicinity of scientific developments.

Indicators reflecting the amount of non-patent references can be grasped through directly available and accessible data sources. More specifically, we use the amount of non-patent references as found within published USPTO patent documents (so called 'other references') as an indicator of the science-intensity or science proximity of patents.[4]

3.3 Data sources and indicators used

Delineating the biotechnology domain

As stated before we take biotechnology as science and technology field under study because of the importance of the field and the presence of a strong interplay between science and technology. In order to select relevant patents we build further on previous research efforts focusing on the biotechnology industry. To retrieve all relevant biotechnology patents, we use the search strategy developed in the biotechnology domain study of the Steunpunt O&O Statistieken (Glänzel, Meyer et al., 2003). This search strategy takes the OECD definition of the biotechnology area - which draws on five different patent classes of the International Patent Classification - as a starting point

[4] A detailed content analysis of 10,000 non-patent references reveals that about 60% of these references are references towards scientific journals (see Callaert, Van Looy et al., 2006). The remainder relates to books, company reports, databases and the like. At the same time, the correlation between the number of non-patent references and the number of journal references nearly equals 1, allowing to use the frequency of occurrence of non-patent references to address the research questions outlined.

(OECD, *STI Scoreboard 2001*: 32; see also Van Beuzekom, 2001). This search is extended with two IPC subclasses which the Fraunhofer Classification Scheme includes as biotechnology-relevant: C07G (Compounds of unknown constitution) and C12R (indexing scheme related to subclasses C12C to C12Q or C12S, related to micro-organisms). In addition to the WIPO International Patent Classification scheme, the US Patent and Trademark Office (USPTO) uses its own classification scheme. Based on the US classification, Jaffe and his colleagues at NBER set up an alternative classification scheme to the IPC-based Fraunhofer classification (e.g. Hall, Jaffe & Trajtenberg, 2001) which also allows identifying biotechnology related patents. In accordance with this NBER classification scheme, the US patent classes 435 and 800 were added to the search strategy to delineate biotechnology from other US patents. Equipped with these datasets, validation interviews were carried out with a number of field experts confirming the overall validity of the approach. Appendix 3-1 provides more details about the search keys used for patents.

As for publications, data retrieval is based on 9 relevant subject categories provided within the framework of the *WOS SCI Expanded* database (Thomson Reuters ISI Web of Science): DE 'PLANT SCIENCES'; CO 'BIOCHEMICAL RESEARCH METHODS'; CQ 'BIOCHEMISTRY and MOLECULAR BIOLOGY'; DA 'BIOPHYSICS'; DB 'BIOTECHNOLOGY and APPLIED MICROBIOLOGY'; QU 'MICROBIOLOGY'; DR 'CELL BIOLOGY'; KM 'GENETICS and HEREDITY'; and HY 'DEVELOPMENTAL BIOLOGY'. All articles, letters, notes and reviews published in journals classified in those subject categories are retrieved.

For the time period 1992–2001, this resulted in a total number of 51,460 USPTO granted patents and 967,188 *WOS* publications.

Technological and scientific productivity

As an indicator of technological performance we use technological productivity. In order to obtain this indicator, the yearly number of granted patents by country applied for during the time period 1992–1999 is divided by the total population.[5] As data were

[5] At the same time, we conducted a parallel analysis using technological performance as measured by the total number of patents (logarithmically transformed) and using population figures as independent

retrieved in 2002, patents granted during the time period 2000–2001 have been omitted from the analysis as for these years the diminishing number of granted patents – due to the time delays implied by the granting process – might considerably affect the relationships examined.

It can be noted that the concept of technological productivity would be more accurately depicted by dividing the number of patents by the total amount of R&D expenditures within the field, or other input-related indicators (e.g. number of engineers or researchers). While this is undoubtedly the case, the variety of disciplines and industries involved as well as the lack of reliable data – at the country level – on R&D expenditures and other input-related indicators that are to be attributed unambiguously to biotechnology activities prevented such an approach. For the analysis reported in the following section, logarithmic transformation of the technological productivity variable has been applied in order to obtain a normal distribution.

All patents are grouped by country based on the nationality of the patent assignee.[6] In the case of co-patenting involving multiple countries, full counts are applied for all countries involved. Whereas patent data retrieved relate to patents granted during the time period 1992–1999, we will use the application year of the patents within the analysis. Given that the time period between applying and granting averages between 2 and 3 years, application dates are preferred to assess the relationship with scientific capabilities. With respect to this latter, the total number of scientific publications – as retrieved from *WOS* – within biotechnology is aggregated by country. Also here, in the case of co-authored papers implying different countries, full counts are been applied. In order to examine the relationship between technological performance and scientific capabilities, patent indicators reflecting technology activity at moment T0 (application date) are related to scientific capabilities as measured at T1 to account for the delays

variables. Such an analysis reveals similar results as obtained here with respect to the relationships between technological productivity, scientific productivity and science-intensity (see Appendix 3-2).

[6] As outlined by Dernis, Guellec & Van Pottelsberghe (2001) one could opt for allocating patents to countries based on the nationality of the assignees and/or on the nationality of the inventors. While both approaches have their advantages and disadvantages, we opted for assignees in order to be congruent with the allocation process for publications which is based on the nationality of the affiliated institution. In addition it can be noted that data obtained by adopting both approaches correlate high (e.g. $r = 0.99$, $p < 0.001$ for the country level data pertaining to 1994 reported by Dernis, Guellec & Van Pottelsberghe).

observed in publications.[7] Also this variable has been transformed in order to obtain a normal distribution.

Science-intensity

In order to assess the science proximity of patents, the number of non-patent citations retrieved in patent documents is used here as a measure of the distance (or closeness) between science and technology development (see also Schmoch, 1997). For each year and country, average values are calculated.[8] As became clear above and similar to the majority of studies that investigate the link between science and technology, we make use of so-called USPTO front-page references. While it could be argued that the analysis undertaken might benefit from including applicant given references (see for instance Meyer, 2000b), their present unavailability within the database does not allow for inclusion.[9] Hence, the analysis includes all examiner given references – to be found on the front page – which include in most cases a considerable amount of applicant references.

3.4 Results

Applying the search keys for patents and publications resulted in a dataset covering 20 countries[10] demonstrating technology and scientific activities on a yearly base during

[7] An analysis with a time lag of two years yielded similar results as the ones reported here. As pointed out by one of the reviewers, both for Canada and the USA, the time lag might be less due to the fact that applicants tend to apply first in their country of origin. In order to assess whether this phenomenon would affect the results obtained we also performed an analysis implying a differentiated time lag (USA and Canada versus all other countries) of one year. These analysis differed from the ones presented here only marginally in absolute terms and did not result in any change of the nature of the relationships found nor the significance levels obtained.

[8] This variable has not been transformed as it is normally distributed. We also performed an analysis with a logarithmical transformed version of this variable; this yielded similar results as the ones reported further on in the study.

[9] Whether or not big differences would result from using either source remains to a large extent unclear. While the social processes in which applicant and examiner's roles are embedded might justify expected differences, empirical work that demonstrates these differences remains scarce. A detailed analysis of citations made in a sample of 366 patents in the genetic field (time period 1980–1985) by Collins & Wyatt (1988) revealed no major differences with respect to citations given by examiners and applicants. Also the recent analysis of Meyer (2000a) indicates that the majority of applicant given references tend to be included in the references assigned by examiners.

[10] In total biotechnology patents were found for over 50 countries; within this analysis we only used country data if patent activity was to be observed throughout the whole time period considered (1992–1999).

the time period covered (1992–1999). Table 3-1 provides an overview of summary statistics related to the key indicators by country. Science-intensity reflects the average number of non-patent references found within patents; scientific and technological productivity is expressed in terms of number of publications and patents per million inhabitants. For technological and scientific productivity a logarithmic transformation is performed in order to obtain a normal distribution.

Table 3-1 : Descriptive statistics. Science-intensity, Technological productivity and Scientific productivity by country

	SCIENCE-INTENSITY			SCIENTIFIC PRODUCTIVITY			TECHNOLOGICAL PRODUCTIVITY		
Country	Mean	Std dev	N	Mean	Std dev	N	Mean	Std dev	N
Austria	14.81	4.16	8	101.82	21.17	8	1.13	0.42	8
Australia	14.72	5.42	8	138.18	18.39	8	1.94	1.11	8
Belgium	11.32	3.56	8	145.30	19.13	8	2.10	1.78	8
Canada	24.12	4.90	8	150.59	5.09	8	3.08	1.23	8
Switzerland	11.10	3.39	8	277.56	28.51	8	5.24	1.74	8
Germany	11.32	2.38	8	101.70	13.90	8	1.95	0.71	8
Denmark	10.25	3.12	8	232.76	26.06	8	10.80	4.90	8
Spain	8.17	5.30	8	64.55	27.93	8	0.14	0.08	8
Finland	22.47	8.31	8	149.40	63.08	8	2.34	0.98	8
France	11.44	2.99	8	116.10	10.45	8	2.02	1.11	8
Great Britain	15.14	5.31	8	159.03	12.65	8	2.25	0.88	8
Israel	18.81	8.19	8	197.35	13.13	8	3.90	1.19	8
Italy	5.98	2.70	8	67.30	8.01	8	0.38	0.19	8
Japan	6.65	1.06	8	78.30	8.01	8	2.82	0.63	8
Korea	6.23	2.46	8	23.28	12.51	8	0.39	0.15	8
Netherlands	13.60	4.10	8	179.79	11.88	8	4.08	1.46	8
Norway	12.23	4.83	8	135.53	17.91	8	0.99	0.41	8
Sweden	17.22	8.66	8	255.94	24.30	8	2.31	1.03	8
Taiwan	12.65	7.99	8	31.43	6.43	8	0.45	0.29	8
USA	23.53	4.07	8	123.59	7.40	8	9.08	3.47	8
Total	**13.59**	**7.15**	**160**	**136.48**	**70.39**	**160**	**2.87**	**3.12**	**160**

Science-intensity: Average amount of non-patent references found within patent documents;
Scientific Productivity: Average amount of scientific publications (within the field of biotechnology) normalized by population (divided by million inhabitants);
Technological Productivity: Average amount of patents (within the field of biotechnology) normalized by population (divided by million inhabitants).

As Table 3-1 makes clear one observes considerable differences between countries for all three variables. Science-intensity is highest for Canada, the United States of America,

Finland, Sweden and Israel. Considerable lower levels are observed for Spain, Italy, Korea and Japan. In terms of scientific productivity, Switzerland, Sweden and Denmark are ranked in the top 3; Taiwan, Korea, Italy, Japan and Spain are characterized by productivity levels which are three to five times lower. These differences are remarkably as they do not coincide systematically with English being a native language within these countries. In terms of technological productivity, Denmark and the USA display the highest figures, followed – although at a distance (50%) – by Switzerland and the Netherlands. Low levels of technological productivity can be observed for Spain, Italy, Korea, Taiwan and Norway.

Table 3-2 complements these data by providing an overview of the correlations observed between the variables. One observes positive, significant relationships between technological and scientific productivity; at the same time science-intensity correlates positively with both productivity indicators.

Table 3-2 : Correlations between Scientific productivity, Technological productivity and Science-intensity

		Application year	Scientific productivity	Technological productivity	Science-intensity
Application year	Pearson Corr	1	0.147	0.11	–0.041
	Sig (2-tailed)		0.065	0.168	0.609
	N	160	160	160	160
Scientific productivity	Pearson Corr	0.147	1	0.683**	0.255**
	Sig (2-tailed)	0.065		0	0.001
	N	160	160	160	160
Technological productivity	Pearson Corr	0.11	0.683**	1	0.399**
	Sig (2-tailed)	0.168	0		0
	N	160	160	160	160
Science-intensity	Pearson Corr	–0.041	0.255**	0.399**	1
	Sig (2-tailed)	0.609	0.001	0	
	N	160	160	160	160

*** Correlation is significant at the 0.01 level (2-tailed).*

Within a next step, partial correlations have been calculated relating the different key constructs one by one controlling for application year and the third key variable. Table 3-3 summarizes the results. While one observes again a positive relationship between scientific and technological productivity as well as science-intensity and technological productivity, the positive relationship between science-intensity and scientific

productivity found without controlling for technological productivity disappears. Stated otherwise, the correlation observed between scientific productivity and science-intensity can be attributed to the relationship between both variables and technological productivity.

Table 3-3 : Partial correlations between Scientific productivity, Technological productivity and Science-intensity

		Scientific productivity	Technological productivity	Science-intensity
Scientific productivity	Pearson Corr	1	0.647**	–0.0173
	Sig (2-tailed)		0	0.829
	N	160	160	160
Technological productivity	Pearson Corr	0.647**	1	0.320**
	Sig (2-tailed)	0		0
	N	160	160	160
Science-intensity	Pearson Corr	–0.0173	0.320**	1
	Sig (2-tailed)	0.829	0	
	N	160	160	160

*** Correlation is significant at the 0.01 level (2-tailed).*

The positive relationship between technological productivity on the one hand and scientific productivity and science-intensity can be observed as well when performing a regression analysis as Table 3-4 makes clear. Both scientific productivity and science-intensity turn out to be positively related to technological productivity.

Table 3-4 : Regression model. Dependent variable: Technological productivity. Independent variables: Scientific productivity, Science-intensity and Application year.

Model summary

R	R^2	Adjusted R^2	Std. error of the estimate
0.722	0.521	0.521	0.35633

Coefficients

	Unstandardized coefficients		Standardized coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-13.579	24.880		-0.546	0.586
Science-intensity	173.7	0.004	0.243	4.232	0.000
Application Year	6490	0.012	0.029	0.502	0.604
Scientific Productivity	4468	0.000	0.617	10.619	0.000

These findings allow answering the two research questions addressed in this chapter in an affirmative way: science-intensity is positively associated with technological performance. This positive relationship is to be found as well when bringing scientific capabilities into the equation. Stated otherwise, countries in which patents include more non-patent references display higher levels of technological activity. Notice that the same relationship holds for scientific capabilities: higher (lower) levels of scientific productivity coincide with higher (lower) levels of technological productivity.

Within a final analysis, we explore the relationship over time between the three variables under study. In a first step, additional variables where introduced reflecting technological and scientific productivity as well as science-intensity at T+2.[11] Given the limited number of variables, we calculated the partial correlation coefficients between the variables in line with the path analysis logic outlined by Blalock (1961) (see also Davis, 1985). Two observations are emerging: first of all evidence is found for the mutual or bi-directional influence of scientific and technological productivity over time. Whereas technological productivity at T+2 is largely associated with past technological productivity (T), a positive and significant relationship with scientific productivity is observed and vice versa. Second, the science-intensity of patents not only seems to be a 'path dependent' phenomenon – as the positive relationship between science-intensity at T and T+2 indicates – a distinctive and positive association with technological productivity (T+2) can be observed as well. Again, these findings reveal that the amount of non-patent references coincides with varying levels of technological productivity. Stated otherwise; the more technology development is situated in the vicinity of scientific developments, the higher technological productivity. This relationship holds within the field of biotechnology – at a country level – even after introducing scientific productivity into the analysis.

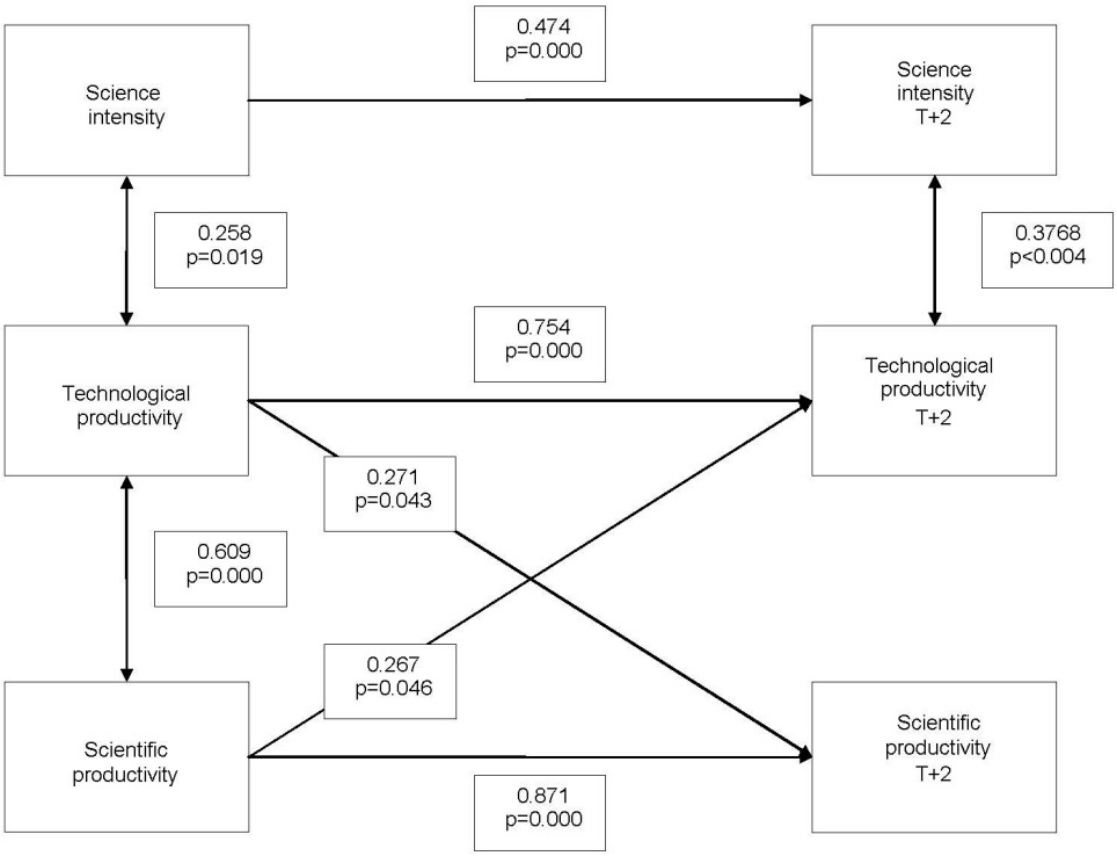
Figure 3-1 summarizes the obtained partial correlation coefficients whereby only significant relationships are depicted.[12] Two observations are emerging: first of all evidence is found for the mutual or bi-directional influence of scientific and

[11] An analogous analysis has been conducted with T+3 variables, yielding similar results.

[12] For the variables at T+2 all other variables (T and T+2) have been controlled for; while for variables at moment T, partial correlations imply a correction for the third variable (at moment T) only.

technological productivity over time. Whereas technological productivity at T+2 is largely associated with past technological productivity (T), a positive and significant relationship with scientific productivity is observed and vice versa. Second, the science-intensity of patents not only seems to be a 'path dependent' phenomenon – as the positive relationship between science-intensity at T and T+2 indicates – a distinctive and positive association with technological productivity (T+2) can be observed as well. Again, these findings reveal that the amount of non-patent references coincides with varying levels of technological productivity. Stated otherwise; the more technology development is situated in the vicinity of scientific developments, the higher technological productivity. This relationship holds within the field of biotechnology – at a country level – even after introducing scientific productivity into the analysis.

Figure 3-1 : Partial correlation coefficients Technological and Scientific productivity and Science-intensity T0 and T+2 (path analysis)



3.5 Conclusions, discussion and directions for further research

The relationship between science and technology within the field of biotechnology indeed reveals itself here as reciprocal and bi-directional rather than unidirectional or linear while at the same time both activity domains deploy their own 'internal' dynamics (see e.g. Rip, 1992).

At the same time a distinctive relationship between science-intensity or science proximity – as measured by the amount of non-patent references – and technological productivity has been observed. These findings corroborate the construct validity of indicators based on non-patent references found within patents. In addition, the positive relationship between science-intensity - or stated otherwise, the closeness between science and technology - and technological productivity, corroborates the relevancy of policy frameworks that foster interaction between knowledge/science generating institutions (universities, research centres) and technology producers (companies).

These findings also suggest interesting avenues for further research. While we focused on one specific field (biotechnology), refining the insights obtained in terms of their field specific nature requires extensions towards other fields. Likewise, introducing extended time frames would allow assessing whether differential effects are to be observed related to technological life cycle dynamics (Abernathy & Utterback, 1978; Andersen, 2001). Finally, extending the analysis to include other patent system and different counting methods (see Guellec & Van Pottelsberghe, 2001) seems more than worthwhile to pursue in order to assess the robustness or the peculiarities of the findings obtained.

3.6 Limitations of the use of non-patent references to study direct science-technology relationships and the need for additional methodological research

Although non-patent references can be used as indicator of science-technology linkages at an aggregate level, as used in this chapter, this indicator falls short to detect direct science-technology interactions – as described in the introduction of this dissertation.

Because of limitations of current indicators and methods, we continue with a methodological part devoted to the development of new techniques and indicators to detect (direct) links between patents and scientific publications. This methodological part is instrumental for our final goal to look for the presence of an anti-commons effect, an important potential drawback induced by increasing academic patenting – one aspect of increasing science-technology interactions. In this methodological section, we develop a new indicator for direct interactions based on text mining.

3.7 References

Abernathy, W. J. & **Utterback**, J. M. (1978). "Patterns of industrial innovation. " *Technology Review*, 80 (7) : 40-47.

Andersen, B. (2001). *Technological Change and the Evolution of Corporate Innovation*. Edward Elgar Publishing.

Blalock, H. (1961). *Causal Inferences in Nonexperimental Research*. Chapel Hill (NC): UNC Press.

Callaert, J., **Van Looy**, B., **Verbeek**, A., **Debackere**, K. & **Thijs**, B. (2006). "Traces of prior art: A systematic analysis of other references found within the USPTO and EPO patent system." *Scientometrics*, 69 (1) : 3-20.

Collins, P. & **Wyatt**, S. (1988). "Citations in patents to the basic research literature." *Research Policy*, 17 : 65–74.

Davis, J. A. (1985). "The Logic of Causal Order." *Quantitative applications in the social sciences series*, 55 : 38-44. Thousand Oaks (CA): Sage Publications.

Dernis H., **Guellec**, D., **Van Pottelsberghe de la Potterie**, B. (2001). "Using patent counts for crosscountry comparisons of technology output." *STI Review*, OECD, 27 : 129–146.

Glänzel, W., **Meyer**, M., **Schlemmer**, B., **du Plessis**, M., **Thijs**, B., **Magerman**, T., **Debackere**, K. & **Veugelers**, R. (2003). *Domain Study Biotechnology: An analysis based on publications and patents.* Leuven, Belgium: Steunpunt O&O Statistieken.

Grupp, H. & **Schmoch**, U. (1992). "Perception of scientification of innovation as measured by referencing between patents and papers." In: H. **Grupp** (Ed.), *Dynamics of Science-Based Innovations*: 73-128. Berlin/Heidelberg: Springer Publishers.

Guellec, D. & **Van Pottelsberghe de la Potterie**, B. (2001). "The internationalisation of technology analysed with patent data." *Research Policy*, 30 (8) : 1253–1266.

Hall, B. H., **Jaffe**, A. B. & **Trajtenberg**, M. (2001). "The NBER patent citations file: Lessons, insights and methodological tools". NBER Working Paper 8498. Cambridge (MA): National Bureau of Economic Research.

McMillan, S., **Narin**, F. & **Deeds**, D. (2000). "An analysis of the critical role of public science in innovation: The case of biotechnology." *Research Policy*, 29 : 1–8.

Meyer, M. (2000a). "Patent citations in a novel field of technology – What can they tell about interactions between emerging communities of science and technology." *Scientometrics*, 48 (2) : 151–178.

Meyer, M. (2000b). "Does science push technology? Patents citing scientific literature." *Research Policy*, 29 : 409–434.

Michel, J. & **Bettels**, B. (2001). "Patent citation analysis: A closer look at the basic input data from patent search reports." *Scientometrics*, 51 (1) : 185–201.

Narin, F. & **Olivastro**, D. (1992). "Status report: Linkage between technology and science." *Research Policy*, 21 : 237–249.

Nelson, R. R. (Ed.) (1993). *National Innovation Systems: A Comparative Analysis*. Oxford University Press.

OECD (2001). STI Scoreboard 2001. Paris: OECD Publications.

Rip, A. (1992). “Science and technology as dancing partners.” In: **Kroes**, A. & **Bakker**, M. (Eds), *Technological Development and Science in the Industrial Age*. Kluwer Publications.

Schmoch, U. (1997). “Indicators and the relations between science and technology.” *Scientometrics*, 38 : 103–116.

Tijssen, R. J. W., **Buter**, R. K., **Van Leeuwen**, T. N. (2000). “Technological relevance of science: Validation and analysis of citation linkages between patents and research papers.” *Scientometrics*, 47 : 389-412.

Tijssen, R. J. W. (2001). “ Global and domestic utilization of industrial relevant science: Patent citation analysis of science-technology interactions and knowledge flows.” *Research Policy*, 30 : 35–54.

Van Beuzekom, B. (2001). Biotechnology Statistics in the OECD Member Countries: Compendium of Existing National Statistics. OECD STI Working Paper 2001/6. OECD Publishing.

Van Looy, B., **Magerman**, T. & **Debackere**, K. (2007). “Developing technology in the vicinity of science: An examination of the relationship between science intensity (of patents) and technological productivity within the field of biotechnology.” *Scientometrics*, 70 (2) : 441-458

Van Looy, B., **Zimmermann**, E., **Veugelers**, R., **Mello**, J. & **Debackere**, K. (2003). “Do science-technology interactions pay off? An exploratory investigation of 10 science intensive fields.” Scientometrics, 57 : 335-367.

Appendix 3-1 : Search strategy for biotechnology patents.

The search strategy developed in the biotechnology domain study of the Steunpunt O&O Statistieken is used (Glänzel, Meyer et al., 2003). The starting point of that search key is the OECD definition of the biotechnology of 2001 (OECD, *STI Scoreboard 2001* : p. 32; see also Van Beuzekom, 2001) based on following 5 IPC subclasses:

- C12M : Apparatus for enzymology or microbiology;
- C12N : Micro-organisms or enzymes; propagating, preserving, or maintaining micro-organisms; mutation or genetic engineering; culture media;
- C12P : Fermentation or enzyme-using processes to synthesise a desired chemical compound or composition or to separate optical isomers from a racemic mixture;
- C12Q : Measuring or testing processes involving enzymes or microorganisms; compositions or test papers therefore; processes of preparing such compositions; condition-responsive control in microbiological or enzymological processes;
- C12S : Processes using enzymes or micro-organisms to liberate, separate or purify a pre-existing compound or composition; processes using enzymes or micro-organisms to treat textiles or to clean solid surfaces of materials.

Furthermore, two subclasses which the Fraunhofer Classification Scheme includes as biotechnology-relevant were added: C07G (Compounds of unknown constitution) and C12R (indexing scheme related to subclasses C12C to C12Q or C12S, related to micro-organisms).

In addition to the WIPO International Patent Classification scheme, the US Patent and Trademark Office (USPTO) uses its own classification scheme. Based on the US classification, Jaffe and his colleagues at NBER set up an alternative classification scheme to the IPC-based Fraunhofer classification (e.g. Hall, Jaffe & Trajtenberg, 2001). Also their classification allows to identify biotechnology related patents. In accordance with this NBER classification scheme, the US patent classes 435 and 800 were added to delineate biotechnology from US patents. There are few differences between USPTO patents retrieved through IPC classes and the set of USPTO patents that was defined based on the US classification. As initial tests indicated, the IPC classification identified 8% more patents than the US patent classification.

Based on these datasets, validation interviews were carried out with a number of field experts from the Belgian and Flemish biotechnology industry and research. The interviews confirmed the validity of the overall approach, in particular for fields of 'modern biotechnology'. All important actors in the biotechnology area were identified. However, the interviews also indicated certain fields were not covered to the extent the experts consulted would have expected. In collaboration with them, additional IPC subclasses were added to the search strategy, in particular in the area of health and food-biotechnology. These areas are covered mainly by the subclasses A61K (health) and A23C (food). In order to avoid the inclusion of non-biotechnology patents, the IPC was not only used at the subclass-level but also at the group-level.

Table 3-5 gives an overview of the different search strategies in the biotechnology domain. The broadest possible search strategy was used, encompassing patents after the OECD, Fraunhofer, and NBER classification as well as additional classes and groups as identified in the validation exercise.

Table 3-5 : Overview of search strategy for biotechnology patents

Source	Selected classes
OECD definition based on IPC classification	C12M; C12N; C12P; C12Q and C12S
Fraunhofer classification	C07G; C12M; C12N; C12P; C12Q; C12R and C12S
NBER based US patent classification	435 and 800 (OCL classes)
Additional IPC subclasses based on expert interviews	A23C*; A23J*; A61K*; C07H; C07J and C07K

** For these 3 subclasses only a selection of main groups and subgroups were included. See the Domain study of the Steunpunt O&O Statistieken for an exhaustive list of all IPC codes used (Glänzel, Meyer et al., 2003).*

Appendix 3-2 : Alternative regression model with technological performance measured by the number of patents as dependent variable and population as independent variable.

Table 3-6 : Alternative regression model. Dependent variable: Technological performance (log patents). Independent variables: Scientific productivity, Science-intensity, Application year and population

Model summary

R	R^2	Adjusted R^2	Std. error of the estimate
0.865	0.749	0.742	0.34363

Coefficients

	Unstandardized coefficients		Standardized coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-11.180	24.839		-0.450	0.653
Application year	4218	0.012	0.014	0.340	0.734
Science Intensity	127.7	0.004	0.134	3.096	0.002
Population	6643	0.101	0.005	0.066	0.948
Scientific Capabilities	1.206	0.112	0.833	10.798	0.000

.

4 Introduction to text mining, potential applications in the field of innovation studies, and the Latent Semantic Analysis (LSA) method.

The question of whether computers can think
is about as relevant as the question of whether submarines can swim.
Edsger W. Dijkstra

4.1 Text mining

Text mining refers to the automated extraction of knowledge and information from text by means of revealing relationships and patterns present, but not obvious, in a document collection. Text mining covers a broad field of tasks including text categorization, text clustering, information extraction, sentiment analysis, document summarization, named entity recognition and question answering and is an interdisciplinary field based on artificial intelligence, natural language processing, computational linguistics, information retrieval, data mining, machine learning and statistics.[13]

Technological advances and large scale availability of computing power attracted a lot of interest for text mining in recent years, together with the observation that the vast majority of (electronically available) information is stored as (unstructured) text and not in structured databases. Hence database technologies and knowledge discovery in structured databases (data mining) alone will fall short to disclose knowledge and information from available resources. Text mining techniques can help to reveal knowledge and information from large text collections, disclosing data not available in

[13] For more information on the application of text mining and its relation to other fields and techniques, see e.g. Hearst, 1999, or Fan, Wallace et al., 2006. For an overview of techniques, see Vidhya & Aghila, 2010.

structured databases. Given the overwhelming amount of unstructured data recorded as texts, text mining will become increasingly valuable for research.

It is important to stress that these text mining techniques go beyond information retrieval. Information retrieval helps in finding information based on a user request, and is it obvious that text mining techniques can help in improving this. As such, currently, information retrieval is probably the biggest area of text mining application and related techniques are widely used. Information retrieval in itself does however not discover new knowledge or insights, it just reveals what is already known to somebody (and also the user issuing the search request has to know what he is looking for)[14].

Text mining does go one step further and is about knowledge discovery, revealing new things that were not obvious to discover by humans. Nice illustrations are some cases of a literature-based approach to scientific discovery by Swanson: fish oil and Raynaud's syndrome; migraine and magnesium; and somatomedin C and arginine (Swanson, 1986, 1988 and 1990). The second case e.g. describes the discovery of the relationship between 'migraine' and 'spreading depression' on the one hand, and 'magnesium' and 'preventing depression' on the other hand after a thorough search into medical literature on migraine, suggesting magnesium deficiency as a factor in migraine. Prior to this remarkable discovery – Swanson is an information scientist, not a physician - there was no indication of this relationship whatsoever; his results triggered additional clinical research, confirming his suggestion[15]. These case studies can be regarded as pioneering cases of text mining – when text mining as such did not even exist – and were at the basis of formalized study to literature-based discovery – so called Swanson Linking (Swanson & Smalheiser, 1997).

4.2 History

Quantitative linguistics dates back to at least the middle of the 19th century (see Grzybek & Kelih, 2004). However, the classical theoretical work by Zipf (1949) is considered pioneering in quantitative linguistic (or text) analysis. Since the 1970s, a

[14] For an elaboration on this issue, see Hearst, 1999.
[15] Ramadan, Halvorson et al., 1989

remarkable increase in activity has been witnessed in this aspect of information science. As for its application to scientific literature, Wyllys's study (1975) is among the first. Co-word analysis, one of the most frequent techniques, was founded on the idea that the co-occurrence of words describes the contents of documents and was developed for purposes of evaluating research (Callon, Courtial et al., 1983). The extension of co-word analysis to the full texts of large sets of publications was possible as soon as large textual databases became available in electronic form; also the increasing availability of computational power allowed further emergence of text mining approaches. Manning & Schütze (2000) provide a comprehensive introduction to the statistical analysis of natural language; Berry (2003) provide a survey of text mining research; Leopold, May & Paaß (2004) give an overview of data and text mining fundamentals for science and technology research; and Porter & Newman (2004) introduced the term 'tech mining' to text mining of collections of patents on a specific topic. Other practical applications in the field of bibliometrics and technometrics are presented by Courtial (1994), Noyons, van Raan et al. (1994), Bassecoulard & Zitt (2004), Leydesdorff (2004), Glenisson, Glänzel et al. (2005) and Janssens, Leta et al. (2006).

4.3 Application in innovation studies

As described in the previous section, application of text mining techniques in innovation studies is not new and can provide the necessary input for a complex research setup that would be impossible or at least difficult (i.e. time consuming because of involved manual data treatment) to conduct without text mining techniques.

As mentioned earlier, a first and more traditional application of text mining is in the field of information retrieval (conducting patent or publication searches on bibliographic databases). But text mining techniques also allow for a new range of applications:

- Domain studies: starting from a set of 'seed patents' that are representative for a technological domain, concepts and topics can be extracted and used to match with concepts and topics of other patents to identify related patents and delineate technological domains by a set of patents;

- Trend detection / emerging field detection: Concepts and topics extracted from a set of patent documents can be clustered over time to identify new domains or to follow the evolution of a domain;
- Science-technology linkage: concepts and topics extracted of a set of patent documents can be compared to concepts and topics extracted from a set of scientific publications to reveal similarity between patents and publications.

In this dissertation we will focus on this latter application by comparing patents and scientific publications to identify patents originated from scientific disclosure.

4.4 Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) was developed late 1980s at BellCore/Bell Laboratories by Landauer and his team of Cognitive Science Research (Landauer & Dumais, 1997):

"Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the meaning of words. Meaning is estimated using statistical computations applied to a large corpus of text. The corpus embodies a set of mutual constraints that largely determine the semantic similarity of words and sets of words. These constraints can be solved using linear algebra methods, in particular, Singular Value Decomposition."[16]

LSA is a mathematical and statistical approach, claiming that semantic information can be derived from a word-document co-occurrence matrix and words and documents can be represented as points in a (high-dimensional) Euclidean space. Dimensionality reduction is an essential part of this derivation.

LSA is based on the Vector Space Model (VSM), an algebraic representation of text documents commonly used in information retrieval. This 'bag-of-words' approach can be seen as a simple yet powerful representation (Salton, 1968; Salton, Wong & Yang, 1975; Salton & McGill, 1983). The vector space of a collection of texts is constructed by representing each document as a vector containing the frequencies of the words or terms the document is composed of as elements. Altogether, these document vectors

[16] Landauer, McNamara et al. (Eds.), 2007, Handbook of Latent Semantic Analysis, Preface x.

add up to a term-by-document matrix representing the full text collection. Relatedness of documents can be derived from those vectors, e.g. by calculating the angle between document vectors by means of a cosine measure.

However, this numerical representation of text data does not solve typical issues of working with language. On the one hand there are morphological problems for the proper identification of terms and the fact that not all terms in a text are of equal importance. This can be solved by feature selection techniques (stemming, stop word removal, collocations, synonym lists, domain vocabulary, part-of-speech taggers, chi-square tests and information gain) and weighting schemes (TF-IDF, Log-Entropy). On the other hand, there are or more fundamental issues with homonymy/polysemy and synonymy. These issues require specific methods to (try to) understand the meaning of words, and that is what LSA claims to do:

"It was thus a major surprise to discover that a conceptually simple algorithm applied to bodies of ordinary text could learn to match literate humans on tasks that if done by people would be assumed to imply understanding of the meaning of words and passages. The model that first accomplishes this feat was LSA."[17]

LSA rests on the single conceptually simple constraint that the representation of any meaningful passage of text must be composed as a function of the representation of the words it contains. Thus, LSA models a passage as a simple linear equation, and a large corpus of text as a large set of simultaneous equations. The solution is in the form of a set of vectors, one for each word and passage, in a semantic space and is solved by Singular Value Decomposition (SVD).

Optimal dimension reduction is a common workhouse in analysis of complex problems in many fields of science and engineering. Over 99.9% of the cells in the word-by-paragraph or term-by-document matrix representing the documents in the vector space turn out to be empty. This makes the comparison of word or paragraph meanings quite chancy. However, after dimension reduction and reconstruction, every cell will be filled with an estimate that yields a similarity between any paragraph and any other and

[17] Landauer, McNamara et al. (Eds.), 2007, Handbook of Latent Semantic Analysis, page 5.

between any word and any other. This dimension reduction is crucial and is what accounts for LSA's advantage over most current methods of information retrieval, which rely on matching literal words. It is also what accounts for its ability to measure the similarity of two essays that use totally different words, and for all of the other properties of LSA that defy the intuition that learning language from language is impossible.

LSA builds upon semantic similarity and hence uses proximity models such as clustering, factor analysis and multidimensional scaling (see Carroll & Arabie, 1980, for a survey). Discovering latent proximity structure has previously been explored for automatic document indexing and retrieval, using term and document clustering (Sparck Jones, 1971; Salton, 1968; Jardin & van Rijsbergen, 1971) and factor analysis (Atherton & Borko, 1965; Borko & Bernick, 1963; Ossorio, 1966); LSA builds further on these factor analysis techniques and constructs a concept-by-document matrix using a low-rank approximation of the term-by-document matrix, combining dimensions or terms into 'concepts'.

Singular Value Decomposition (SVD) is used as a rank lowering method to truncate the original vector space to reveal the underlying or 'latent' semantic structure in the pattern of word usage to define documents in a collection. This truncation allows dealing with typical language issues like synonymy as different words expressing the same idea are supposed to be close to each other in the reduced *k*-dimensional vector space. SVD will decompose the original term-by-document matrix into orthogonal factors that represent both terms and documents:

$$A = U \,.\, \Sigma \,.\, V^T$$

with A the original term-by-document matrix, Σ a diagonal matrix with the square roots of singular values of $A.A^T$ and $A^T.A$ ($\sigma_1^2 > \sigma_2^2 > \ldots > \sigma_n^2$), and U and V containing left and right singular vectors.

Instead of working with the original vector space represented by the original term-by-document matrix A, one can work with the reduced vector space of lower rank, ignoring all but the first *k* singular values in Σ and all but the first *k* columns of U and V:

$$A = A^{m\times n} \cong A_k^{m\times n} = U^{m\times k}.\Sigma^{k\times k}.V^{k\times n}$$

This dimension reduction to *k* dimensions provided by SVD is the closest rank-*k* approximation available and allows eliminating noise and capturing the underlying latent structure. The *k* dimensions in the new space are no longer (stemmed) words or terms, but linear combinations of such linguistic terms, and the basic unit of analysis becomes not just a mere word but a word-and-its-context, a concept (hence the denomination of 'concept space').

Mind that the dimension reduction is not about computational simplification[18] but a fundamental aspect of the method to deal with language issues and reduce noise (terms in documents that do not contribute to the meaning of the document or parts of the document). As such, the choice of *k* is not arbitrary but needs to be chosen carefully to truly represent the underlying latent structure of the data.

The choice of the number of concepts to be retained is not straightforward. Current literature suggests to take 100 to 300 concepts for large datasets (Berry, Drmac & Jessup, 1999; Jessup & Martin, 2001; Lizza & Sartoretto, 2001). For some applications it might be better to use a subset of the first 100 or 300 dimensions (Landauer & Dumais, 1997).

4.5 Practical indexing and additional pre-processing steps

Indexing in practice

The encoding of the documents into vectors is called indexing. During indexing, a global vocabulary is built up, assigning a unique identifier to each word encountered in the entire document collection. With this global vocabulary, a vector is constructed for each

[18] At the contrary, SVD will convert the original sparse matrix into a full matrix. Even for low values of *k* – the number of retained dimension or concepts – this will result in a new document-by-term matrix of lower rank but occupying far more memory and in general taking more computational resources to process.

document with as many elements as the total number of words in the global vocabulary, generating the vector space. For words appearing in the document at hand, the value of the respective vector elements of the document vector is equal to the number of occurrences of that word in the document at hand. For words not appearing in the document, the respective vector elements obtain a zero value. Thus, each document is represented by a vector representing raw frequencies of occurrences in a high-dimensional vector space of terms. As each document uses only a small subset of words to describe its content, the resulting matrix is extremely sparse (containing mostly zeros)[19].

To improve the indexing process and achieve better grasp of the context of the documents, subsequent additional pre-processing actions are commonly used:

Stop word removal

All common words that do not contribute to the distinctive meaning and context of documents can be removed before indexing (e.g. “a”, “the”). Commonly used word lists are available containing a large set of so-called ’stop’ words (e.g. the SMART list of Buckley and Salton, Cornell University).

Stemming

Instead of indexing words as they appear in the documents, linguistic stems can be used for indexing. The basic idea is to reduce the number of words by introducing a common denominator, called a stem, for words that share a common meaning (e.g. ‘produc’ for “product”, “production”, “producing”, etc.). A well-known example is the Porter stemmer (see van Rijsbergen, Robertson & Porter, 1980, and Porter, 1980). This stemmer does not perform a linguistically correct lemmatization, but takes a pragmatic approach in stripping suffixes from words to combine word variants with shared meanings.

The idea of stemming is to improve the ability to detect similarity regardless of the use of word variants (stemming reduces the number of synonyms, since multiple terms

[19] About 99.99% zero values.

sharing the same stem are mapped onto the same concept or stem), but occasionally stemming will create new homonyms because of stemming errors[20].

Term reduction

According to Zipf's law a large number of terms only appear in one document. Such hapaxes can be removed from the vocabulary because they are of little value in finding communality between documents.

Weighting

Representing documents based on the occurrence and co-occurrence of terms (raw frequencies) can be refined by introducing a weighting scheme to better distinguish the distinctive nature of words and terms given the specific context under study (e.g. the word 'computer' does not reveal the distinctive nature of a certain contribution within a document set covering only papers on computer science). A commonly used weighting scheme is the TF-IDF weighting scheme (Salton & McGill, 1983), in which the raw term frequencies are multiplied by the inverse document frequency (IDF) for that term; this results in augmenting the impact of relatively rare terms when calculating distance measures:

$$Idf_i = \log \frac{n}{\sum_{j=1}^{n} \chi(f_{ij})},$$

with

$$\chi(t) = \begin{cases} 1 \text{ if } t > 0, \\ 0 \text{ if } t = 0 \end{cases}$$

i = term index
j = document index
f_{ij} = frequency of term i in document j
and n the number of documents.

[20] A more in-depth analysis of the performance and advantages and disadvantages of stemming (which are also language and corpus dependent) is outside the scope of this publication. The reader interested in this aspect is referred to Lennon, Pierce et al., 1981; Harman, 1991; Krovetz, 1995; and Porter, 2001.

Weighting has a similar effect as stop-word removal, since words commonly used across all documents in the document set will be down-weighted compared to medium frequency words, which carry the most significant information (Salton & Wu, 1981) – as can be expected according to Zipf's law. On the other hand, TF-IDF weighting attributes might introduce extreme weights to words with very low frequencies. Also, TF-IDF will not grasp synonyms; hence, weights of commonly used synonyms will be over-rated, as the weights of the individual (synonym) terms will be higher than the weight of the underlying common concept. Despite these shortcomings, TF-IDF weighting is one of the most popular weighting schemes, but other weighting schemes can also be used (see Manning & Schütze, 2000, for an overview).

Additional, more advanced, pre-processing tasks can be performed to further optimize the indexing process (proper name recognition; word sense disambiguation; acronym recognition; compound term and collocation detection; feature selection using application-specific domain vocabulary or ontology, information gain, entropy or Bayesian techniques)[21].

4.6 Similarity or distance calculation

The similarity measure typically used in information retrieval applications is the cosine similarity measure (Berry & Browne, 1999). It is an expression for the angle between vectors, formulated as an inner product of two vectors, divided by the product of their Euclidean norms.

If the vectors are normalized beforehand, this formula reduces to the simple inner product. Since, in the original vector space, all vector elements are positive (a word will appear zero times or more in a document), the results are values between 1 (for similar vectors, i.e. pointing in the same direction) and 0 (for mutually orthogonal, entirely unrelated vectors), even after application of a weighting scheme like TF-IDF. This yields distances between 0 and 1 (1 – cosine α). This no longer holds for vectors in the reduced concept space after SVD, since vector elements may become negative because of the SVD, resulting in a concept-by-document space V^{kxn} that is no longer positive semi-

[21] A more detailed description of these topics can be found in Moens, 2006.

definite, and cosine values that might be negative, hence distances between 0 and (theoretically) 2, although values larger than 1.3 are quite rare in practice. While other similarity measures are possible (e.g. Jaccard, Dice, Euclidean distance – see Baeza-Yates & Ribeiro-Neta, 1999), the cosine measure is amongst the most commonly used when using LSA and seems superior as a similarity measure in LSA applications (Harman, 1986).

4.7 Other text mining methods

Before moving to the next chapters with our practical results of applying LSA on patent and publication data, we wish to stress that the proposed LSA methodology is only one available method for text content based similarity deduction. Other methods e.g. do not rely on semantic representation but use semantic topic models based on generative models (e.g. probabilistic inference models like probabilistic latent semantic modelling and latent Dirichlet allocation - see e.g. Wong & Yao, 1995; Hofmann, 1999; Blei, Ng & Jordan, 2003). These models do not rely on a spatial representation and do not suffer from limitations to Euclidean geometry as imposed by the assumption of LSA that documents can be represented as vectors in a vector space[22].

[22] In an Euclidean space, similarity should be symmetric and not violate triangle inequality - $d(x,z) \leq d(x,y) + d(y,z)$ - placing strong constraints on the location of points in a space given a set of distances (Griffiths, Steyvers & Tenenbaum, 2007).

4.8 References

Atherton, P. & **Borko**, H. (1965). "A test of factor-analytically derived automated classification methods." AIP Report AIP-DRP 65-l.

Baeza-Yates, R. & **Ribeiro-Neto**, B. (1999). *Modern information retrieval.* New York: ACM Press.

Bassecoulard, E. & **Zitt**, M. (2004). "Patents and publications: The lexical connection." In H. F. **Moed**, W. **Glänzel** & U. **Schmoch** (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*: 665–694. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Berry, M. W. (Ed.). (2003). *Survey of text mining*. New York: Springer.

Berry, M. W. & **Browne**, M. (1999). *Understanding search engines: Mathematical modelling and text retrieval*. Philadelphia: Society for Industrial and Applied Mathematics.

Berry, M. W., **Drmac**, Z. & **Jessup**, E. (1999). "Matrices, vector spaces, and information retrieval." *SIAM Review*, 41 : 335-362.

Blei, D. M., **Ng**, A. Y. & **Jordan**, M. I. (2003). "Latent Dirichlet allocation." *Journal of Machine Learning Research*, 3 : 993-1022.

Borko, H. & **Bemick**, M. D. (1963). "Automatic document classification." *Journal of the ACM*, 10 : 151–162.

Callon, M., **Courtial**, J. P., **Turner**, W. A. & **Bauin**, S. (1983). "From translations to problematic networks—an introduction to co-word analysis." *Social Science Information - Sur Les Sciences Sociales*, 22 (2) : 191–235.

Carroll, J. D. & **Arabie**, P. (1980). "Multidimensional scaling." In M. R. **Rosenzweig** & L. W. **Porter** (Eds.), *Annual review of psychology*: 31 : 607-649. Palo Alto, CA: Annual Reviews, Inc.

Courtial, J. P. (1994). "A coword analysis of Scientometrics." *Scientometrics*, 31 (3) : 251–260.

Fan, W., **Wallace**, L., **Rich**, S. & **Zhang**, Z. (2006). "Tapping the power of text mining." *Communications of the ACM*, 49 (9) : 77-82.

Glenisson, P., **Glänzel**, W., **Janssens**, F. & **De Moor**, B. (2005). "Combining full-text and bibliometric informationin mapping scientific disciplines." *Information Processing & Management*, 41 (6) : 1548–1572.

Griffiths, T. L., **Steyvers**, M. & **Tenenbaum**, J. B. (2007). "Topics in Semantic Representation." *Psychological Review*, 114 (2) : 211-244.

Grzybek, P. & **Kelih**, E. (2004). "Anton S. Budilovic (1846–1908): A forerunner of quantitative linguistics in Russia?" *Glottometrics*, 7 (9) : 4–97.

Harman, D. (1986). "An experimental study of the factors important in document ranking." In F. **Rabbit** (Ed.), *Association for computing machine's ninth conference on research and development in information retrieval*. New York: Association for Computing Machines.

Harman, D. (1991). "How effective is suffixing?" *Journal of the American Society for Information Science*, 42 : 7–15.

Hearst, M. A. (1999). "Untangling Text Data Mining." Proceedings of the 37th annual meeting of the Association for Computational Linguistics: 3-10. College Park, Maryland.

Hofmann, T. (1999). "Probabilistic latent semantic indexing." *Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval* : 50-57.

Janssens, F., **Leta**, J., **Glänzel**, W. & **De Moor**, B. (2006). "Towards mapping library and information science." *Information Processing and Management*, 42 (6) : 1614–1642.

Jardin, N. & **van Rijsbergen**, C. J. (1971). "The use of hierarchic clustering in information retrieval." *Information Storage and Retrieval*, 7 : 217–240.

Jessup, E. & **Martin**, J. (2001). "Taking a new look at the latent semantic analysis approach to information retrieval." In M. W. **Berry** (Ed.), *Computational information retrieval*: 121-144. Philadelphia: SIAM.

Krovets, B. (1995). "Word sense disambiguation for large text databases." Ph. D. Thesis. Department of Computer Science, University of Massachusetts Amherst.

Landauer, T. K., & **Dumais**, S. T. (1997). "A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge." *Psychological Review*, 104 (2) : 211-240.

Landauer, T. K., **McNamara**, D. S., **Dennis**, S. & **Kintsch**, W. (Eds.) (2007). *Handbook of Latent Semantic Analysis*. Mahwah (NJ): Lawrence Erlbaum Associates.

Lennon, M., **Pierce**, D. S., **Tarry**, B. D. & **Willett**, P. (1981). "An evaluation of some conflation algorithms for information retrieval." *Journal of Information Science*, 3 : 177–183.

Leopold, E., **May**, M., & **Paaß**, G. (2004). "Data mining and text mining for science & technology research." In H. F. **Moed**, W. **Glänzel**, & U. **Schmoch** (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*: 187–213. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Leydesdorff, L. (2004). "The university-industry knowledge relationship: analyzing patents and the science base of technologies." *Journal of the American Society for Information Science and Technology*, 55 (11) : 991–1001.

Lizza, M. & **Sartoretto**, F. (2001). "A comparative analysis of LSI strategies." In M. W. **Berry** (Ed.), *Computational information retrieval*: 121-144. Philadelphia: SIAM.

Manning, C. D. & **Schütze**, H. (2000). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Moens, M. F. (2006). *Information extraction: Algorithms and prospects in a retrieval context* (The Information Retrieval Series 21). New York: Springer.

Noyons, E. C. M., **van Raan**, A. F. J., **Grupp**, H. & **Schmoch**, U. (1994). "Exploring the science and technology interface–inventor author relations in laser medicine." *Research Policy*, 23 (4) : 443–457.

Ossorio, P. G. (1966). "Classification space: A multivariate procedure for automatic document indexing and retrieval." *Multivariate Behavior Research*, 1 : 479–524.

Porter, M. F. (1980). "An algorithm for suffix stripping." *Program*, 14 (3) : 130–137.

Porter, M. F. (2001). Snowball: A language for stemming algorithms. (www.snowball.tartarus.org/texts/introduction.html).

Porter, A. L. & **Newman**, N. C. (2004). "Patent profiling for competitive advantage. "In H. F. **Moed**, W. **Glänzel** & U. **Schmoch** (Eds.), *Handbook of quantitative science and technology*

research. The use of publication and patent statistics in studies of S&T systems: 587–612. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ramadan, N. M., **Halvorson**, H., **Vandelinde**, **A**. & **Levine,** S.R. (1989). "Low brain magnesium in migraine." *Headache*, 29 (7) : 416-419.

Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw-Hill.

Salton, G. & **McGill**, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw Hill.

Salton, G., **Wong**, A. & **Yang**, C.S. (1975). "A vector space model for information retrieval." *Journal of the American Society for Information Science*, 18 (11) : 613-620.

Salton, G. & **Wu**, H. (1981). "A term weighting model based on utility theory." In R. N. **Oddy**, S. E. **Robertson**, C. J. **van Rijsbergen** & R. W. **Williams** (Eds.), *Information retrieval research*: 9–22. Boston: Butterworths.

Sparck Jones, K. (1971). *Automatic keyword classification for information Retrieval*. London: Buttersworth.

Swanson, D. R. (1986). "Fish Oil, Raynaud's syndrome, and undiscovered public knowledge." *Perspectives in Biology and Medicine,* 30 : 7-18.

Swanson, D. R. (1988). "Migraine and magnesium: Eleven neglected connections." *Perspectives in Biology and Medicine,* 31 : 526-557.

Swanson, D. R. (1990). "Somatomedin C and arginine: Implicit connections between mutually-isolated literatures." *Perspectives in Biology and Medicine,* 33 : 157-186.

Swanson, D. R. & **Smalheiser**, N. R. (1997). "An interactive system for finding complementary literatures: a stimulus to scientific discovery." *Artificial Intelligence*, 91 : 183-203.

van Rijsbergen, C. J., **Robertson**, S. E. & **Porter**, M. F. (1980). *New models in probabilistic information retrieval*. British Library Research and Development Report, No. 5587. London: British Library.

Vidhya, K. A. & **Aghila**, G. (2010). "Text Mining Process, Techniques and Tools: an Overview." *International Journal of Information Technology and Management*, 2 (2) : 613-622.

Wong, S. K. M. & **Yao**, Y. Y. (1995). "On modeling information retrieval with probabilistic inference." *ACM Transactions on Information Systems*, 13 (1) : 69–99.

Wyllys, R. E. (1975). "Measuring scientific prose with rank-frequency (''Zipf'') curves: A new use for an old phenomenon." *Proceedings of the American Society for Information Science*, 12 : 30–31.

Zipf, G. K. (1949). Human behavior and the principle of least effort: An introduction to human ecology. Cambridge: Addison-Wesley.

5 Exploring the feasibility and accuracy of text mining techniques based on Latent Semantic Analysis to detect similarity between patent documents and scientific publications[23].

I can't understand a word you say. And you're poorly dressed.
You must be some sort of technology expert. Or a rodeo clown.
Dilbert's pointy haired boss

5.1 Introduction

In this chapter, we investigate the feasibility and relevancy of content (lexical) analysis implying both patent and publication documents. Text analysis is already being used in efforts to delineate specific domains or subfields. Until now, such demarcation has relied heavily on expert opinions to identify appropriate sets of terms and/or classes in available classification schemes (e.g. Hicks, Martin & Irvine, 1986; Hinze & Grupp, 1996; Glenisson, Glänzel et al., 2005; Glenisson, Glänzel & Persson, 2005; Rabeharisoa, 1992, in fuel cells; Noyons, van Raan et al., 1994, in laser medicine; Schmoch, 2004, in genetics; Glänzel, Meyer et al., 2003, in biotechnology and Meyer, 2000, in nano-science and nanotechnology). Clearly, experts involved in domain studies would benefit from automated results that indicate similarity and hence enable mapping, categorization or classification.

But not only domain studies would profit from methodologies that permit the identification of content similarity across different sets of documents. The current debate on the relevance and appropriateness of academic patenting and entrepreneurship reveals that, under certain conditions, combining scientific and technological activities yields certain beneficial effects, including scientific productivity

[23] The study as described in this chapter has been published in Scientometrics (Magerman, Van Looy & Song, 2010).

(see introduction chapter). At the same time, it can be noted that the occurrence of such beneficial effects may be partly dependent on the (topic) relatedness of both activities. Further analysis of whether and to what extent knowledge spillover dynamics – between scientific and technological activity realms – are present and result in positive 'reinforcing' rather than 'jeopardizing' dynamics might benefit from the ability to assess content-relatedness between sets of documents – in this case, patents and publications. Indeed the identification of science-related patents in general and the identification of scientific results protected by intellectual property rights (IPR) in particular is a major challenge. As described in the introduction chapter, current approaches based on non-patent references or matching of patent inventors and patentees with publication authors and affiliations has limitations.

We propose a new approach involving text mining to directly match text documents based on their contents to find patent and publication documents that are related by the topics they address, the methods they use, the results they obtain and the inventions or discoveries they address, as this might allow (semi)-automated compilation of large datasets based on robust constructs (content similarity). At this moment, we are particularly interested in this text mining approach to check for documents with identical contents to identify scientific publications protected by patents, allowing to compile large datasets to check for the presence of an anti-commons effect.

However, applicability of off-the-shelve text mining solutions is not straightforward at this stage. Multiple methods are available but existing experience for patent data is limited and more research is needed concerning effectiveness and best practices (methods, pre-processing, source data, indexing options, number of concepts to be retained, …). In this chapter we present our study on a try-out of Latent Semantic Analysis (LSA) based lexical text analysis techniques to construct distance measures that are well suited to grasp similarities between patent and publication text documents. This might allow identification of patents originated from scientific publications.

In this study, we limit ourselves to patents and publications of the same academic inventor. Contrary to domain studies, which often involve thousands of documents, the

number of relevant documents under consideration is much smaller. The choice of this small scale set-up will simplify computational challenges. Indeed, in advance we do not know where related patent-publication pairs can be found, forcing us to select a very large number of patents and publications to be sure we have related patent-publication pairs in our sample, resulting in computational challenges. By sticking to datasets of patents and publication of the same academic inventor, we can use far smaller samples and concentrate on validation issues to get first insights in the feasibility and accuracy of this text mining based approach.

So, a first question that arises is related to whether traditional assumptions applied in large-scale text mining applications are as relevant for small-scale applications - such as the one envisaged here. In addition, combining different types of documents – i.e. scientific publications and patent documents – introduces an additional level of complexity, which justifies further analysis to assess the relevance and accuracy of text mining algorithms.

As became apparent from the previous chapter on text mining and the LSA method, different options and methods are available to arrive at similarity measures. This variety of possible approaches is translated into following research design: for the six academic inventors under study, we calculate a set of 23 distance measure variants and use these variants to derive similarity scores for all scientific puplications and patent documents based on the content of the documents (title and abstract). It will become clear that different options and calculation methods indeed yield different outcomes. Hence, in a next step of the analysis, we compare the accuracy of the measures obtained by comparing them with independently obtained assessments of similarity. This will not only allow us to draw conclusions on the feasibility of the overall approach; our findings also suggest tentative propositions on the methods and options that are best suited for small-scale applications, implying documents of a heterogeneous nature.

5.2 Research design

The ability to automatically match large numbers of patent and publication documents opens interesting perspectives for search and retrieval applications, clustering

applications, discriminate analysis, domains studies, emerging fields detection, science and technology linkage, and so on. Although text mining applications have proven to be useful in some areas, there is still limited proof of its ability to actually identify relevant similarities for patent and publication documents, especially at the micro level (see e.g. Engelsman & van Raan, 1994, and Bassecoulard & Zitt, 2004, for some meso and macro level applications of lexical patent and publication coupling).

When it comes to patents and publications, only titles and abstracts are widely and easily available. Large sets of full-text documents are difficult or expensive to obtain. At the same time, while text mining may be relevant for natural language documents, publication and especially patent abstracts rarely read as natural language. Moreover, as the previous chapter has shown, implementing a text mining procedure requires many options and parameters to be set. Together, all these options and parameters generate a broad spectrum of possibilities to represent the documents in a vector space, and hence to arrive at distance measures. Although some generally accepted practices exist, there is still a lack of clarity about which options yield better results and under what circumstances.

This study aims at a systematic comparison between variants of distance measures resulting from a set of procedural options based on LSA. First, we seek to verify whether different options yield different similarity outcomes when applied to small sets of patents and publications. Next, we wish to determine if these differences also coincide with differences in accuracy by comparing the obtained similarity measures with independently obtained similarity ratings. This comparison will also allow us to draw tentative conclusions on the feasibility of practical applications.

Data

For this feasibility study, we do not use our biotechnology patent and publication dataset, but we instead selected six academic inventors from the Catholic University of Leuven – four from the medical faculty and two from the engineering faculty. All WO, EPO and USPTO patents were downloaded from *MicroPatent* (Thomson Reuters Micropatent) where one of the six professors appeared as inventors. After deduplication of the patent families and removal of patents without abstracts, 30

patents, ranging from 2 to 12 patents per academic inventor, were retained. Next, all publications of these professors appearing in the *WOS* database (Thomson Reuters ISI Web of Science) were downloaded. This resulted in 437 publications, ranging from 33 to 106 publications per professor (again only publications with an abstract were retained). Together, the dataset contains 467 documents.

Text mining options: delineation of selected parameters.

To assess the similarity between patent and publication documents, the distance between every (seed) patent and all publications of the same academic inventor is calculated using a variety of text-mining-based distance measures based on Latent Semantic Analysis (LSA). Stop-word removal using the SMART stop-word list was applied before indexing, as well as stemming using the Snowball analyser with the Porter stemming algorithm. Without these options, distance measures tend to yield unreliable results because too much non-relevant information is introduced. There is some debate about the reliability of Porter's stemmer for scientific and technological language. The rules this stemmer is composed of were conducted from natural languages examples; applied on the somewhat distinct language of science and technology, stemming errors might introduce too much unwanted homonyms. We decided to include stemming as our previous research experience showed significant better results when using stemming, but this issue definitely deserves more attention.

All terms occurring in only one document were removed. To further refine the index, some high frequency words that do not convey much information in the patent and publication context ("method", "present", "result", "studi" and "type") were also removed.

Most literature indicates TF-IDF weighting as a valuable step to obtain relevant distance measures by down-weighting less important terms. To verify the impact of weighting for smaller scale applications, and in combination with SVD dimensionality reduction, we included both TF-IDF weighting and no weighting (using the raw term frequencies) in our model.

The literature also suggests that LSA using SVD can improve significantly the performance of the distance measures compared to a cosine measure applied on the full vector space. Traditionally, rank-*k* approximations containing a few hundred dimensions are used. While this undoubtedly makes sense in large datasets containing thousands of documents – since the global vocabulary of these sets can contain ten thousands of terms – the relevance for small datasets is less clear, resulting in the inclusion of the level of dimensionality reduction by SVD in our research design.

Normally, a set of documents is indexed and weighted as a whole, and SVD is performed on the global index of all documents. In our set-up, we are only interested in relations within the set of patents and publications of the respective academic inventors. Accordingly, we have two options to perform weighting and SVD: index all documents of all academic inventors together and perform weighting and SVD on the global, unified vocabulary of all six academic inventors, or index the documents separately for each academic inventor and perform weighting and SVD on the case-specific vocabulary of the respective academic inventor. The individual or case-specific approach holds the promise that the weighting and SVD might be optimized for each professor individually; this may yield better results since we are only interested in relations within the document set of an academic inventor. But this case-specific approach implies that the individual document sets are small while one can expect that revealing the underlying latent structure in a document set by SVD requires large document collections. We included both the global unified vocabulary weighting and SVD and local case-specific weighting and SVD in our analysis. For the case-specific vocabulary approach, the highest rank-*k* approximation that can be used depends on the smallest document set of all academic inventors, which is 66 (a professor with 2 patents and 33 scientific publications). We opted to include rank-*k* approximations of 30, 20, 10, and 5 (as previous research on small document sets suggests the relevance of very low values of *k*, see Glenisson, Glänzel et al., 2005; and Glenisson, Glänzel & Persson, 2005). For the global unified vocabulary approach, the highest rank-*k* approximation possible is 467 (the total number of documents for all academic inventors). We opted to include rank-*k* approximations of 300 and 100, and also included 30, 20, 10, and 5 for comparison with the case-specific approach. For simplicity, we will denote the different rank-*k*

approximations by 'SVD' followed by the rank-*k* approximation (e.g. SVD 30 means we applied LSA using a rank-30 SVD approximation).

To summarize, we have incorporated the following options into our model: global unified document indexing (Index=G) and individual case document indexing (Index=C); no weighting (Weighting=NO) and TF-IDF weighting (Weighting=TI); and no SVD reduction and reduction to 5, 10, 20, 30, 100, and 300 concepts (the latter two only for the global unified document indexing). Final similarity scores are obtained by using a cosine measure in the vector space created by the indexing process according to the distinct options. Table 5-1 contains an overview of the options and obtained measures.

Table 5-1 : Overview of distance measures

Measure	Index	Weighting	SVD	Measure	Index	Weighting	SVD
1	U	NO	No SVD	15	C	N0	No SVD
2	U	NO	5	16	C	NO	5
3	U	NO	10	17	C	NO	10
4	U	NO	20	18	C	NO	20
5	U	NO	30	19	C	NO	30
6	U	NO	100				
7	U	NO	300				
8	U	TI	No SVD	20	C	TI	No SVD
9	U	TI	5	21	C	TI	5
10	U	TI	10	22	C	TI	10
11	U	TI	20	23	C	TI	20
12	U	TI	30	24	C	TI	30
13	U	TI	100				
14	U	TI	300				

There are fewer measures with local case document indexing because it is not possible to use SVD 100 and beyond for those measures because of the small datasets. Note in this respect that, while Table 5-1 lists 24 combinations, there are only 23 distinct measures. Indeed, when neither weighting nor SVD are applied, global unified document indexing and individual case document indexing yield the same distance scores for the set of relevant documents, hence measure 1 and measure 15 yield identical results.

All distances between all seed patents and all target publications were calculated using these different distance measure variants.

We deliberately decided not to apply more pre-processing tasks, like compound term and collocation detection, because we want to keep the processing simple and automated. These more advanced pre-processing tasks almost always imply more human involvement and manual attention. In this setting, we want to try out if a simple automatic approach will work.

5.3 Comparative analysis of distance measures

Differences in measure characteristics

An overview of the obtained descriptive statistics of all measures can be found in Appendix 5-1. It is clear that one group of measures (measures with no SVD or high rank-*k* SVD – SVD with high number of retained dimension) displays a highly skewed distribution, while other measures (measures with low rank-*k* SVD – SVD with low number of retained dimensions) are far less skewed.

Figure 5-1 : Distribution of distances for four representative measures

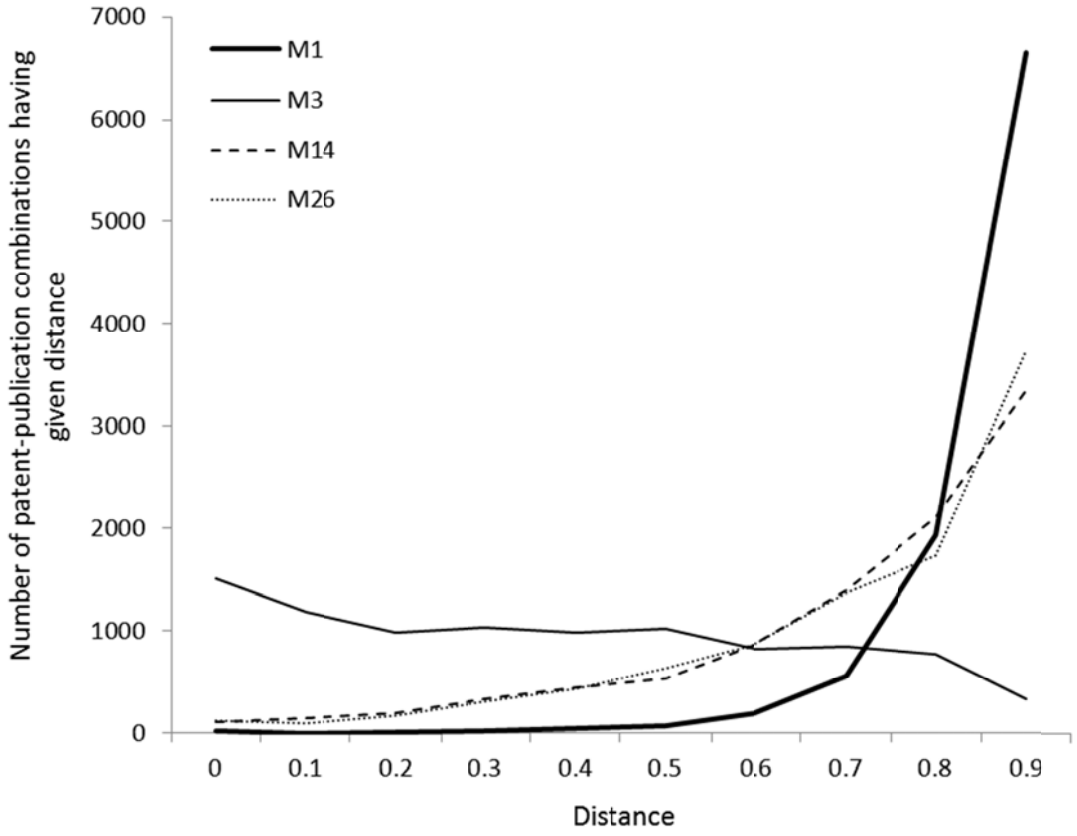


Figure 5-1 contains the distribution of four representative measures for the distance scores of all patent-publication combinations[24]. M1 is the measure with neither weighting nor SVD; M3 is a measure without weighting and high level of dimensionality reduction – i.e. low rank-*k* SVD (SVD 10) – performed on the global unified document set; M13 is a measure with weighting and medium level of SVD reduction (SVD 100) performed on the global unified document set; and M24 is a measure with weighting and medium level of SVD reduction (SVD 30) performed on the local case document set. The Y-axis indicates the number of patent-publication pairs having distances within the range indicated on the X-axis (distance buckets of 0.1).

The measure M3 with high levels of SVD reduction - hence only few retained dimensions - is very distinct from the other measures and has a counter-intuitive shape since one does not expect so many 'close' pairs – and certainly not more close pairs than distinct pairs (average distance is rather low, see Appendix 5-1), while for the other measures we observe distributions that are more in line with the expectations (i.e., less close pairs and more pairs that are more distant)[25]. It seems that high levels of dimensionality reduction maps too many unrelated terms to a small number of concepts, artificially creating close pairs. However, to arrive at such a conclusion, one needs to do more than inspecting descriptive statistics. In a next step, we compare the calculated similarity scores with similarity ratings obtained from independent ratings.

Assessing accuracy of measures

In order to compare and asses the accuracy of the different measures, patent-publication combinations have been rated independently. For the six professors in our study, 16 patents (all patents of four academic inventors and a selection of patents of the remaining two academic inventors - 3 out of 9 and 4 out of 12 patents respectively) were assessed independently in terms of relatedness.

[24] To have as a better idea of the distribution of obtained distance scores amongst similarity measures, all potential combinations between all patents and all publications are included, which yields far more combinations compared to the 2,345 combinations obtained if we only look at patent-publication combinations of each individual academic inventor.

[25] In line with the way the data for this figure was constructed: all combinations between all patents and all combinations of all academic inventors are included. As we expect that patents and publications of distinct academic inventors are rather unrelated, we expect far more distant patent-publication combinations compared to close combinations.

We opted for two independent ratings for each individual case (all patent-publication combinations of a given academic inventor) in order to be sure that this independent assessment was carried out in a consistent manner. In total, five different persons – all active and experienced in the field of science and technology studies – have been involved in this exercise for all six academic inventors.

Each validator was required to rate the relatedness between patent documents on the one hand, and publications on the other. Three categories have been used, ranging from 'highly related' to 'unrelated', with ‘somewhat related’ as the third category. In a next step, the scores of each patent-publication combination were compared and Kappa scores – indicating between-subject consistency – were calculated. In the case of two assessments differing greatly (highly related versus unrelated), both validators reviewed their assessments, potentially - but not necessarily - resulting in an adaptation of one or both scores. After this iteration, Kappa scores were obtained for every academic inventor, ranging from 0.62 to 0.90, signalling satisfactory and even excellent levels of consistency (average for the six academic inventors: 0.83).

For all measures, it now becomes feasible to assess the relation between the ‘expert’ assessment on the one hand, and the relatedness as obtained by the calculated measures on the other. That means, for the selected 16 patents, all patent-publication combinations were independently rated by two experts, and for every academic inventor, the expert score of all patent-publication combinations was used in an ANOVA analysis as independent variable, with the obtained distance scores according to the 24 measures as dependent variable.

Table 5-2 provides an overview of the average R^2 obtained for the measures under study. For every measure, this table contains the average of the R^2 values obtained for each academic inventor, i.e. R^2 value resulting from the ANOVA analysis with the expert scores as independent variable and obtained distance scores as dependent variable for all patent-publication combinations of the given academic inventor. The higher the observed R^2, the more calculated similarity scores coincide with the independent expert assessments.

Table 5-2 : Congruence levels obtained for different measures under study

Index	Weighting	SVD	Mean R^2	Std deviation R^2	N
Case	NO	5	0.247	0.257	16
		10	0.321	0.254	16
		20	0.362	0.274	16
		30	0.379	0.270	16
		No SVD	0.401	0.293	16
	TF-IDF	5	0.191	0.203	16
		10	0.356	0.265	16
		20	0.409	0.277	16
		30	***0.413 (4)***	0.295	16
		No SVD	***0.459 (3)***	0.288	16
Unified	NO	5	0.106	0.135	16
		10	0.195	0.273	16
		20	0.242	0.280	16
		30	0.285	0.324	16
		100	0.341	0.314	16
		300	0.386	0.286	16
		No SVD	0.401	0.293	16
	TF-IDF	5	0.133	0.185	16
		10	0.202	0.263	16
		20	0.251	0.296	16
		30	0.314	0.335	16
		100	0.340	0.324	16
		300	***0.482 (2)***	0.285	16
		No SVD	***0.489 (1)***	0.301	16

Mean R^2 values in bold and italic denote the four highest values obtained amongst all measures.

An inspection of Table 5-2 immediately reveals considerable differences between different measures. Measures coinciding most with independent assessment scores imply either high rank-*k* SVD – i.e. low levels of dimensionality reduction, like *k*=300 – or no SVD at all, in conjunction with a unified thesaurus and the application of TF-IDF weighting. Closely related – in terms of accuracy – are measures that combine weighting with a case-based thesaurus either without SVD or low levels of dimensionality reduction (*k*=30)[26]. Differences with less performing combinations are highly significant ($p<0.0001$).

Again according to Table 5-2, better performing measures share the characteristic that they are relatively modest in terms of information reduction. Applying no SVD by

[26] Note that, in this case-based indexing approach, SVD rank-*k* values of 30 can be considered as low dimensionality reduction as most dimensions are retained (given the size of case-based document sets).

definition implies refraining from reducing the initial word space, while applying SVD with a relatively large number of dimensions also respects the potential richness of the underlying information.

TF-IDF weighing also has a positive impact, albeit smaller than the application of SVD. The positive impact of weighting can be understood as distinct elements of documents being emphasized.

While the observations related to weighting may come as no surprise, the results on SVD are more counter-intuitive. As Table 5-2 reveals, SVD performs worst under all circumstances compared to a cosine measure on the full vector space, especially with a limited number of dimensions. The higher the number of dimension retained, the more the scores approximate the scores with no SVD applied, but there is no level of SVD reduction beating these scores. Given the premises of LSA, we expected better scores for at least some levels of SVD dimensionality reduction.

While the reduction in overall R^2 in Table 5-2 already illustrates the deterioration, scrutinizing specific patent-publication combinations really reveals the impact of parameter choices. Appendix 5-2 contains the title and abstract of one patent document and two publications, (co-) authored by an academic inventor under study. On reading these documents, it becomes apparent that one publication is 'highly related' while the other is ‘unrelated’. Table 5-3 provides a detailed insight with respect to the distances obtained under different indexing parameter choices. Note that low values indicate similarity – with a zero value indicating complete similarity – while values approaching 1 signal no relatedness at all. As Table 5-3 clarifies, applying an SVD solution with a limited number of dimensions (k=5) results in similarity measures that suggest that publication 2 is more related to the patent document than publication 1, while in fact the opposite holds true. This phenomenon manifests itself both when using a unified or a case-based thesaurus. This example illustrates how a strong reduction in underlying information may result in vector spaces that – when used to calculate distances between objects – yield distance measures of a misleading nature.

Table 5-3 : Example of impact of specific text mining choices on obtained distance scores

Seed Patent	Gluten biopolymers
Publication 1 (close to seed patent)	Designing new materials from wheat protein.
Publication 2 (far from seed patent)	In situ polymerization of thermoplastic composites based on cyclic oligomers.

Options taken to arrive at distance score			**Obtained distance values**		**Assessment**
Index	**Weighting**	**SVD**	**Publication 1 (highly related)**	**Publication 2 (unrelated)**	
Unified	NO	5	0.015	0.009	Misleading
Unified	TF-IDF	300	0.102	0.908	Accurate
Case	NO	5	0.051	0.036	Misleading
Case	TF-IDF	30	0.030	0.967	Accurate

At the same time, the two other measures shown in Table 5-3 (unified thesaurus, SVD 300 and case-based thesaurus, SVD 30) also illustrate the feasibility of applying text mining algorithms to detect similarity, even in the case of document sets stemming from different activity realms (patents and publications). Overall, these observations suggest that choices made, with respect to the setup of a vector space model and how to proceed when calculating similarity measures, affect considerably the outcomes obtained.

5.4 Conclusions, discussion, limitations, and directions for further research.

In this study, we applied and validated a set of content based similarity measures based on Latent Semantic Analysis (LSA) text mining techniques to construct distance measures that might allow us to grasp similarities between patent documents and scientific publications. We used small-scale patent and publication datasets of six academic inventors to examine the feasibility of matching patents with publications.

Several options for obtaining similarity measures within the framework of this model have been outlined and assessed in terms of accuracy. Our findings reveal that different options and methods coincide with considerable differences in terms of accuracy. While several combinations allow us to arrive at acceptable solutions, certain combinations display low levels of accuracy and even result in misleading similarity measures. For

relatively small datasets, options that respect the potential richness of the underlying data yield better results: either one opts for no dimensionality reduction (cosine on the full vector space) or one opts for dimensionality reduction retaining a relatively high number of dimensions. In addition, weighting has a beneficial impact under these conditions. For a set of small datasets, a global unified indexing and weighting (and SVD, if applied) approach does not yield worse results than an individual, case based, indexing and weighting approach. This is an interesting finding because a global unified indexing approach is far more convenient in practice. But LSA seems not to redeem its promise to deal with synonymy and polysemy problems in our setting; all measures involving SVD perform worse than a plain cosine measure on the full vectors space. We suspect this has to do with the low number of documents in the sample, especially for our case based indexing and SVD approach.

At the same time, this analysis has some limitations which might inspire future research. First, while our analysis might also contribute to the making of better-informed choices when confronted with larger and more heterogeneous document sets, further research might investigate which set of options yields better results when one works with larger datasets. Especially the effects of LSA deserve more attention (from which point onwards LSA improves results and how it deals with synonymy, polysemy and homonymy problems in practical datasets). Second, while several combinations yield relevant outcomes – and the specific example introduced in Table 5-3 clearly indicates the potential of text mining for the given purposes – average observed R^2 for the better set of options are not extremely high (approaching 0.50[27]). Improving accuracy levels might be feasible by further broadening the set of pre-processing options. For instance, when inspecting several patent-publication combinations, it became apparent that introducing more synonyms or collocations and phrase detection might further contribute to improving accuracy. Hence, research focusing on the precise impact of additional parameters not included in this design seems highly relevant. Finally, certain of our cases also seem to suggest that there is not much relatedness to be observed across patents and publications. Indeed, the question arises to what extent it is feasible

[27] Note that for some academic inventors R^2 of 0.80 has been obtained.

to define – for a given set of processing options – absolute values that would clearly detect the presence or absence of similarity (taking into account the inevitable trade-offs between recall and precision). While far from straightforward to conduct, the availability of a set of 'threshold' values would be especially beneficial for situations in which possibilities for extensive validation are limited. As the lack of extensive validation efforts will probably be the rule rather than the exception for most practical applications, the availability of validated threshold values might have a huge impact on the diffusion rate of text mining techniques in this and related fields. Accordingly, we hope that the analysis presented here will act as a source of inspiration for other researchers to engage in such efforts.

5.5 References

Bassecoulard, E. & **Zitt**, M. (2004). "Patents and publications: The lexical connection." In H. F. **Moed**, W. **Glänzel** & U. **Schmoch** (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*: 665–694. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Engelsman, E. C. & **van Raan**, A. F. J. (1994). "A patent based cartography of technology." *Research Policy*, 23 : 1–26.

Glänzel, W., **Meyer**, M., **Schlemmer**, B., **du Plessis**, M., **Thijs**, B., **Magerman**, T., **Debackere**, K. & **Veugelers**, R. (2003). *Domain Study Biotechnology: An analysis based on publications and patents.* Leuven, Belgium: Steunpunt O&O Statistieken.

Glenisson, P., **Glänzel**, W., **Janssens**, F. & **De Moor**, B. (2005). "Combining full-text and bibliometric information in mapping scientific disciplines." *Information Processing & Management*, 41 (6) : 1548–1572.

Glenisson, P., **Glänzel**, W. & **Persson**, O. (2005). "Combining full-text and bibliometric indicators: A pilot study." *Scientometrics*, 63 (1) : 163–180.

Hicks, D., **Martin**, B. R. & **Irvine**, J. (1986). "Bibliometric techniques for monitoring performance in technologically oriented research: The case of integrated-optics." *R&D Management*, 16 (3) : 211–223.

Hinze, S. & **Grupp**, H. (1996). "Mapping of R&D structures in transdisciplinary areas: New biotechnology in food sciences." *Scientometrics*, 37 (2) : 313–335.

Magerman, T., **Van Looy**, B. & **Song**, X. (2010). "Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications." *Scientometrics*, 82 (2) : 289-306.

Meyer, M. (2000). "Patent citations in a novel field of technology: What can they tell about interactions of emerging communities of science and technology?" *Scientometrics*, 48 (2) : 151–178.

Noyons, E. C. M., **van Raan**, A. F. J., **Grupp**, H. & **Schmoch**, U. (1994). "Exploring the science and technology interface–inventor author relations in laser medicine." *Research Policy*, 23 (4) : 443–457.

Rabeharisoa, V. (1992). "A special mediation between science and technology: When inventors publish scientific articles in fuel cells." In H. **Grupp** (Ed.), *Dynamics of science-based innovation*: 45–72. Berlin: Springer.

Schmoch, U. (2004). "The technological output of scientific institutions." In H. F. **Moed**, W. **Glänzel** & U. **Schmoch** (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*: 717–731. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Appendix 5-1 : Basic descriptive statistics for all measures.

Field descriptions:

Index: Union (U); Case (C)
Weighting: No weighting (NO); TF-IDF weighting (TI)
SVD: SVD rank-*k* reduction (0 = no SVD dimensionality reduction)
M: Measure identification number
Mean: Mean distance between patents and publications
Std Dev: Standard deviation of distance
Min/Max: Minimum / maximum distance
Median: Median distance
Low Q/Upp Q: Lower / upper quartile
Q Range: Quartile range
Kurt: Kurtosis
Skew: Skewness

Index	Weighting	SVD	M	Mean	Std Dev	Min	Max	Median	Low Q	Upp Q	Q Range	Kurt	Skew
U	NO	0	1	0.91	0.10	0.00	1.00	0.94	0.88	0.98	0.09	16.72	-3.16
U	NO	5	2	0.31	0.26	0.00	1.16	0.25	0.08	0.49	0.40	-0.56	0.71
U	NO	10	3	0.42	0.28	0.00	1.17	0.40	0.17	0.64	0.47	-1.09	0.22
U	NO	20	4	0.54	0.27	0.00	1.24	0.57	0.334	0.75	0.42	-0.92	-0.24
U	NO	30	5	0.62	0.26	0.00	1.19	0.66	0.44	0.83	0.38	-0.66	-0.51
U	NO	100	6	0.80	0.19	0.00	1.19	0.85	0.71	0.93	0.22	1.88	-1.38
U	NO	300	7	0.87	0.14	0.00	1.04	0.92	0.83	0.97	0.13	7.84	-2.37
U	TI	0	8	0.95	0.09	0.00	1.00	0.97	0.94	0.99	0.05	37.17	-5.04
U	TI	5	9	0.12	0.13	0.00	1.18	0.08	0.04	0.17	0.13	15.00	3.05
U	TI	10	10	0.31	0.27	0.00	1.23	0.23	0.07	0.49	0.41	-0.36	0.79
U	TI	20	11	0.45	0.29	0.00	1.24	0.44	0.17	0.70	0.52	-1.17	0.15
U	TI	30	12	0.53	0.29	0.00	1.19	0.57	0.27	0.79	0.52	-1.19	-0.23
U	TI	100	13	0.77	0.22	0.00	1.36	0.84	0.67	0.94	0.26	1.08	-1.23
U	TI	300	14	0.90	0.14	0.00	1.08	0.95	0.87	0.98	0.11	11.66	-3.02
C	NO	5	16	0.43	0.28	0.00	1.35	0.40	0.20	0.65	0.46	-0.91	0.32
C	NO	10	17	0.60	0.26	0.00	1.35	0.63	0.41	0.81	0.40	-0.79	-0.24
C	NO	20	18	0.73	0.22	0.00	1.19	0.77	0.60	0.89	0.29	0.32	-0.91
C	NO	30	19	0.78	0.20	0.00	1.17	0.83	0.69	0.93	0.23	1.72	-1.31
C	TI	0	20	0.96	0.08	0.00	1.00	0.98	0.96	0.99	0.03	55.53	-6.23
C	TI	5	21	0.33	0.29	0.00	1.50	0.23	0.08	0.54	0.46	-0.37	0.82
C	TI	10	22	0.54	0.30	0.00	1.21	0.57	0.29	0.80	0.51	-1.19	-0.15
C	TI	20	23	0.70	0.26	0.00	1.28	0.77	0.53	0.91	0.38	-0.35	-0.72
C	TI	30	24	0.78	0.22	0.00	1.35	0.84	0.67	0.95	0.28	1.11	-1.17

Appendix 5-2 : Title and abstract of one patent document and two publications (highly related and unrelated) authored by the same academic inventor.

Seed patent: Gluten biopolymers

This invention consists of a modified gluten biopolymer for use in industrial applications, such as composites and foams. In the present work, the fracture toughness of the gluten polymer was improved with the addition of a thiol-containing modifying agent. This work also resulted in the development of a gluten biopolymer-modified fiber bundle, demonstrating the potential to process fully biodegradable composite materials. Qualitative analysis suggests that a reasonably strong interface between the natural fibers and biopolymer matrix can form spontaneously under the proper conditions. Therefore this invention relates to a modified gluten biopolymer for use in industrial applications, such as composites, stabilized foams and molded articles of manufactures. The present invention relates to a new gluten based biopolymer with modified properties, such as an increase in impact strength, and prepared by using thiol-containing molecules. The multifunctional activity of the polythiol-containing molecules generates the potential for the development of a new material base for commodity plastics. The invention furthermore relates to a new composite material comprising gluten-coated fiber, its use and the method for preparing the composite material.

Publication 1 (highly related to the patent document): Designing new materials from wheat protein

We recently discovered that wheat gluten could be formed into a tough, plastic-like substance when thiol-terminated, star-branched molecules are incorporated directly into the protein structure. This discovery offers the exciting possibility of developing biodegradable high-performance engineering plastics and composites from renewable resources that are competitive with their synthetic counterparts. Wheat gluten powder is available at a cost of less than $0.5/lb, so if processing costs can be controlled, an inexpensive alternative to synthetic polymers may be possible. In the present work, we demonstrate the ability to toughen an otherwise brittle protein-based material by increasing the yield stress and strain-to-failure, without compromising stiffness. Water absorption results suggest that the cross-link density of the polymer is increased by the presence of the thiol-terminated, star-branched additive in the protein. Size-exclusion high performance liquid chromatography data of molded tri-thiol-modified gluten are consistent with that of a polymer that has been further cross-linked when compared directly with unmodified gluten, handled under identical conditions. Remarkably, the mechanical properties of our gluten formulations stored in ambient conditions were found to improve with time.

Publication 2 (unrelated to the patent document): In situ polymerization of thermoplastic composites based on cyclic oligomers

The high melt viscosity of thermoplastics is the main issue when producing continuously reinforced thermoplastic composites. For this reason, production methods for thermoplastic and thermoset composites differ substantially. Lowering the viscosity of thermoplastics to a value below 1 Pa.s enables the use of thermoset production methods such as resin transfer molding (RTM). In order to achieve these low viscosities, a low viscous mixture of prepolymers and catalyst can be infused into a mold where the polymerization reaction takes place. Only a limited number of polymerization reactions are compatible with a closed mold process. These polymerization reactions proceed rapidly compared to the curing reaction of thermosets used in RTM. Therefore, the processing window is narrow, and managing the processing parameters is crucial. This paper describes the production and properties of a glass fiber reinforced polyester produced from cyclic oligoesters.

6 Assessment of Latent Semantic Analysis (LSA) text mining algorithms for large scale mapping of patent and scientific publication documents.

The great tragedy of science:
the slaying of a beautiful hypothesis by an ugly fact.
Thomas Huxley

6.1 Introduction

In the previous chapter we examined the application of Latent Semantic Analysis (LSA) for smaller, pre-structured – i.e. both patent and publication selection is done at the level of an academic inventor – datasets (see also Magerman, Van Looy & Song, 2010). Results are promising but it became clear that generally accepted LSA options do not always yield the best results for our setup.

In this study we conduct a thorough assessment of the LSA text mining method and its options (pre-processing, weighting, …) to grasp similarities between patent documents and scientific publications on a larger scale, without the need for compiling datasets at the level of individual academic inventors (which is not convenient for larger datasets). We want to assess effectiveness (in terms of precision and recall) and derive best practices on weighting and dimensionality reduction for application on patent data, given the technical and juridical nature and hence different linguistic context of patent and scientific publication documents. Our primary goal is to set up a method to identify scientific publications that are protected by patents (so called 'patent-publication' or 'patent-paper' pairs, i.e. scientific publications from which the contents – methodology, findings, discovery/invention – is subject of a patent publication). We use LSA to derive similarity from a large set of patent and scientific publication documents (88,248 patent documents and 948,432 scientific publications) based on 40 similarity measurement variants; four weighting schemas – no term weighting; binary weighting; inverted

document frequency; and term frequency x inverted document frequency – are combined with ten levels of dimensionality reduction – no SVD reduction; 1,000; 500; 300; 200; 100; 50; 25; 10; 5 dimensions – and the cosine metric. In addition we also include three similarity measure variants into the comparison based on the number of common terms weighted by the total number of terms of the documents. A thorough validation is set up to compare the performance of those measure variants: the degree of similarity of 300 patent-publication combinations is rated by experts to compare with the outcomes of the text mining measures and about 30,000 patents from control sets are used to check the robustness of the expert validation results.

We first discuss the dataset used and the retained options to derive the measure variants. Next we will present aggregated results and first insights based on the distribution of similarity scores for a large set of patent-publication combinations, followed by results and insights based on an expert validation of 300 patent-publication pairs. Next we will elaborate on some LSA/SVD issues revealed by the results and finalise with conclusions.

6.2 Data and methodology

Match patents and publications based on content

We primarily want to identify so called patent-publication or patent-paper pairs, i.e. scientific publications for which the contents – methodology, findings, discovery/invention – is subject of a patent application. We do this by matching patent and publication documents based on content similarity using LSA text mining algorithms. For all patents, we derive similarity scores for all publications for a set of measurements variants based on LSA. Patent-publication combinations having a high content similarity are regarded to originate from the same inventive event. The choice of the best measure to grasp meaningful relations among patent and publication documents depends on a validation exercise.

We apply our method on patents and publications from the biotechnology field as it is known to be a science-intensive field with substantial science-technology interactions, and because we want to use the results and insights to check for the presence of an

anti-commons effect, which is especially relevant for biotechnology (see the introduction chapter). We compile a set of patent and publication documents related to biotechnology and calculate the content similarity between all patents and publications in the set to reveal patent-publication combinations originating from the same inventive event.

Selection of biotechnology patents

On the patent side, we limit ourselves to the OECD definition of biotechnology to identify biotechnology patents (OECD, 2005 and 2009), defining 30 International Patent Classification subclasses/groups related to biotechnology (see Appendix 6-1 for the list of IPC-subclasses/groups used for the selection). We use *PATSTAT* (EPO Worldwide Patent Statistical Database) to retrieve all EPO and USPTO granted patents with application and grant year between 1991 and 2008 according to the 30 defined IPC-subclasses/groups related to biotechnology. This leads to a set of 27,241 EPO and 91,775 USPTO granted patents (*PATSTAT* edition October 2009).

As text mining techniques are applied for the further identification of patent-publication pairs, only patents with titles and a minimum abstract length of 250 characters are withheld, resulting in a final patent dataset of 7,254 EPO and 80,994 USPTO biotechnology patents (hence 88,248 patents in total).

Selection of scientific publications

On the publication side, we select biotechnology publications (articles, letters, notes, reviews)[28] from the *WOS* database (Thomson Reuters ISI Web of Science) based on the Web of Science subject classification, for the same time period 1991-2008 (volume year). 243,361 publications are revealed from subject category 'Biotechnology and Applied Microbiology'.

However, to ensure that all potentially related scientific publications are present in the dataset, we extend this 'core' publication set with publications from nine related subject categories: 'Biochemical Research Methods'; 'Biochemistry & Molecular

[28] Articles are by far the biggest category (90% articles compared to 1.5% letters, 2% notes and 6.5% reviews).

Biology'; 'Biophysics'; 'Plant Sciences'; 'Cell Biology'; 'Developmental Biology'; 'Food Sciences & Technology'; 'Genetics & Heredity' and 'Microbiology Materials'[29]. This results in more than 1.75 million additional publications for the period 1991-2008 - a considerable computational challenge for the text mining method to identify patent-publication pairs. To lower the number of publications for ease of calculations without losing too much relevant documents, we only retain those publications from this extended set that are citing or are cited by at least one publication from our core set, sizing down the extended publication set to 683,674 publications.

Finally we also add all – not necessarily biotechnology - publications from three multidisciplinary journals ('Nature', 'Science' and 'Proceedings of the National Academy of Sciences of the United States of America') resulting in 97,970 additional publications.

Again we only retain publication documents with titles and a minimum abstract length of 250 characters, resulting in a final publication set of 948,432 biotechnology related publications[30].

Selection of control sets

To check the validity of our text mining method we also compile three control sets with patent documents that are not related to biotechnology: agriculture; automotive; and materials. For each of these control sets, we randomly select 2,500 EPO and 7,500 USPTO granted patent documents from the same time period based on IPC-codes (respectively IPC class A01 for agriculture; B60 and B62 for automotive, and IPC subclass G01N, G01R and HO1L for materials)[31]. As always we only retain documents with titles and a minimum abstract length of 250 characters, resulting in a control set of 29,952 patents related to agriculture, automotive and materials.

---

[29] We want to thank Wolfgang Glänzel for his kind help in the development of a search strategy for biotechnology publications.

[30] As all publication of three multidisciplinary journals are included, non-biotechnology publications will also be present as it is not straightforward to isolate biotechnology publications from those multidisciplinary journals.

[31] Patents of the control groups are selected in such a way that there is no overlap with biotechnology patents, i.e., patents classified in both biotechnology IPC classes and one of the control sets IPC classes are not selected for the control groups, only for the biotechnology group. This is of particular interest for the agriculture control group, as this group can be related to biotechnology and share some IPC codes (A01H 1/00 and A01H 4/00).

Combined dataset

In total, 1,174,021 patent and publication documents are originally selected based on the respective search keys, of which 1,066,632 documents are included in our final setup (all documents having an abstract of substantial length to allow text mining): 88,248 biotechnology patents; 9,952 agriculture patents; 10,000 automotive patents; 10,000 materials patents; 219,713 core biotechnology publications; 647,029 extended biotechnology publications and 81,690 publications from multidisciplinary journals.

6.3 Derivation of content similarity

Index parameters and comparison of measures based on Latent Semantic Analysis

We want to match patents and publications based on content similarity, and want to use LSA to derive content similarity from the patent and publication documents. In practice, applying this method involves multiple pre-processing steps to convert a document collection into a document-by-term matrix (tokenization, indexing, weighting, see previous chapters), and for every of those steps, multiple options are available, resulting in a myriad of choices to arrive at a document-by-term matrix as input for the LSA model. As stated before, the application of LSA in itself also requires a careful choice of the level of dimensionality reduction. Finally, multiple metrics are available to arrive at a similarity value. This heterogeneity in the process to derive content-based similarity measures makes the choice of the best similarity measure (and all corresponding pre-processing options required) very challenging for the purpose at hand, especially as best practices are not readily available. Moreover, our previous experience revealed that common practice does not always yield the best results (see previous chapter and Magerman, Van Looy & Song, 2010).

To shed a light on the feasibility and performance of LSA content-based measures for large scale patent-publication matching, we again compare multiple measures based on multiple weighting options and multiple levels of dimensionality reduction, combined with a limited number of pre-processing steps and the cosine metric. This allows us to check the effect of weighting and dimensionality reduction options on the performance of the matching method and select the best measure and procedure to arrive at reliable

matching results. In total we combine four weighting methods with ten levels of dimensionality reduction, resulting in a setup with 40 measures based on LSA. To complete the assessment of the performance of text mining based measures for patent-publication matching, we also add three measures based on a simple count of the number of common terms between documents.

Pre-processing choices

The first step in the process it to convert the document collection into a numerical dataset. LSA is based on the Vector Space Model: every document is represented by a vector in a highly dimensional space and every element in the vector represents the weight for a given term for the document at hand.

In practice this is done by an indexer splitting text documents into tokens or terms and compiling a list on how many times a given term appears in a given document. We use Apache-Lucene™, an open source text search engine library, for the indexing[32]. During the indexing process, a minimal number of stop words is removed[33], numbers are removed[34] and stemming is applied (Porter stemmer)[35].

Next we use MathWorks Matlab™, a commercial packet for mathematical and technical computing, for the construction of the vector space by converting the Lucene full text index into a document-by-term matrix[36] [37]. This results in a matrix with 1,066,632 rows (documents) and 729,761 columns (stemmed terms). After removal of terms/stems only appearing in one document, we end up with a document-by-term matrix with 1,066,632 documents and 301,697 terms/stems. This document-by-term matrix contains the raw term frequencies, i.e. the number of times a given term/stem appears in a given document.

[32] http://lucene.apache.org/java/docs/index.html
[33] Based on the Snowball English stop word list (http://snowball.tartarus.org/algorithms/englisch/stop.txt)
[34] Only numbers are removed, i.e. terms that only contain digits. Digits that are part of terms with alphanumerical characters (e.g. chemical formula) are untouched.
[35] http://tartarus.org/~martin/PorterStemmer/index.html
[36] http://www.mathworks.com/products/matlab/
[37] We want to thank Frizo Janssens who was so kind to share his proprietary Matlab code for the import of the full text index into Matlab and compilation of a document-by-term matrix.

We deliberately choose not to apply more pre-processing tasks, like compound term and collocation detection, because we want to keep the processing simple and automated. These more advanced pre-processing tasks almost always imply more human involvement and manual attention, while we want to opt for an automatic approach[38].

4 weighting methods

To improve retrieval and matching performance, raw term frequencies are weighted to take into account the relative importance of a term in a given document or in the complete corpus. Many weighting methods are available, and as in our previous setup we again choose TF-IDF weighting for our current setup as it is commonly used in text mining.

To get a better understanding of the impact of weighting, we again include a non-weighted variant (using the raw term frequencies) and two alternative weighting methods: binary weighting and IDF (inverted document frequency) weighting. In the binary weighting method, we only take the presence or absence of a given term in a given document into account and we ignore the number of occurrences, i.e., the binary weighted frequency of a given term *i* and document *j* is equal to 0 if TF_{ij}=0 and is equal to 1 if $TF_{ij} > 1$ (with TF – term frequency – the number of times a given term appears in a given document). In the IDF weighting method, we combine the binary weighting method with the inverted document frequency, i.e., we replace the raw term frequency for a given term *i* and document j by the IDF value of the given term *i*[39].

To summarize, we compare four weighting methods for the document-by-term matrix: (1) the raw term frequency (the number of occurrences of the given term in the given document); (2) the binary term frequency (0 if the given term is absent in the given document, 1 if the given term is present in the given document); (3) the inverted document frequency (0 if the given term is absent in the given document, the inverted document frequency value if the given term is present in given document); and (4) the TF-IDF value (multiplication of term frequency with inverted document frequency).

[38] A more elaborated overview on pre-processing options can be found in chapter 4.
[39] For more information on IDF and TF-IDF, see chapter 4.

10 levels of dimensionality reduction

Dimensionality reduction is an essential part in the LSA method. It truncates the vector space to reveal the underlying or 'latent' semantic structure in the document collection by mapping terms on latent concepts by combining terms in linear relationships. Truncation is done by applying Singular Value Decomposition (SVD) to get a rank-*k* approximation of the original matrix. Dimensionality reduction is supposed to remove the 'noise' due to polysemy and synonymy present in text documents, but the level of dimensionality reduction, or the best selection of the rank (*k*) of the truncated document-by-term matrix, is an open question. As mentioned before, empirical testing shows that the optimal choice for the number of dimensions ranges between 100 and 300 for large datasets (see chapter 4). For small datasets, low values of *k* (below 10) seem to work as well (Glenisson, Glänzel et al., 2005), although our previous experience suggests the use of large values of *k*, but also reveals that no dimensionality reduction at all might perform best[40].

In this study, we compare multiple levels of reduction (defined by *k*, the rank order of the truncated document-by-term matrix, i.e. the number of dimensions to retain). We include following nine levels of *k*: 1,000; 500; 300; 200; 100; 50; 25; 10; 5. And to assess the overall value of LSA and dimensionality reduction, we compare these nine levels of dimensionality reduction with a tenth variant, namely no dimensionality reduction at all (which is basically not an LSA-based measure anymore as it is just an application of the cosine metric on the full vector space).

40 LSA measures and 3 measures based on common terms

To summarize, we compare 40 measures based on LSA by combining 4 levels of term weighting with 9 levels of dimensionality reduction by SVD and no dimensionality reduction at all. For all these 40 measures, we use limited pre-processing options (stop word removal and stemming) and the cosine metric to arrive at a similarity value.

We also include three measures based on the count of the number of terms the patent and publication document have in common. For these three measures, not based on

[40] See previous chapter and Magerman, Van Looy & Song, 2010.

the cosine metric, we use the same pre-processing options as for the 40 measures based on LSA (stop word removal and stemming). To arrive at a similarity metric with values between 0 and 1 starting from the number of common terms, three variants of normalization are used: (1) divide the number of common terms by the minimum of the number of terms of the patent document on the one hand and the number of terms of the publication document on the other hand ('common terms MIN'); (2) divide the number of common terms by the maximum of the number of terms of the patent document and the publication document ('common terms MAX'); and (3) divide the number of common terms by the average of the number of terms of both documents ('common terms AVG')[41]. The second option is more restrictive compared to the first option and only attributes high similarities if both documents are almost identical (the intersection of both documents is equal to the union of both documents: $A + B = A \cap B$). The first option also attributes high similarity if one document is a subset of another document, even if the latter document contains far more information (the intersection of both documents is equal to one of the documents, but potential large remainder or complement of that one document is neglected: $A + B \neq A \cap B$ but $A = A \cap B$). Hence the first option will yield higher similarity values for the same document combinations than the second option, and the third options will be somewhere in-between.

6.4 Aggregated results

Similarity calculations

We calculate similarity scores between all 88,248 biotechnology patents and all 948,432 biotechnology publications according to the 43 defined similarity measures. For every patent, the closest 10,000 publications and corresponding similarity scores were retained for every of the 43 measure variants.[42]

We do the same for all 29,952 patents in the control set; for every control patent we calculate similarity scores with all of the 948,432 biotechnology publications according

[41] For the ease of reference, we will use 'common terms MIN', 'common terms MAX' and 'common terms AVG' to denote the measures based on the number of common terms and their respective normalization method throughout this document.

[42] Retaining all similarities of all 83 billion combinations 43 times is impossible because of current day storage limitations.

to the 43 defined similarity measure variants and retain again the closest 10,000 publications for every control patent and measure variant.

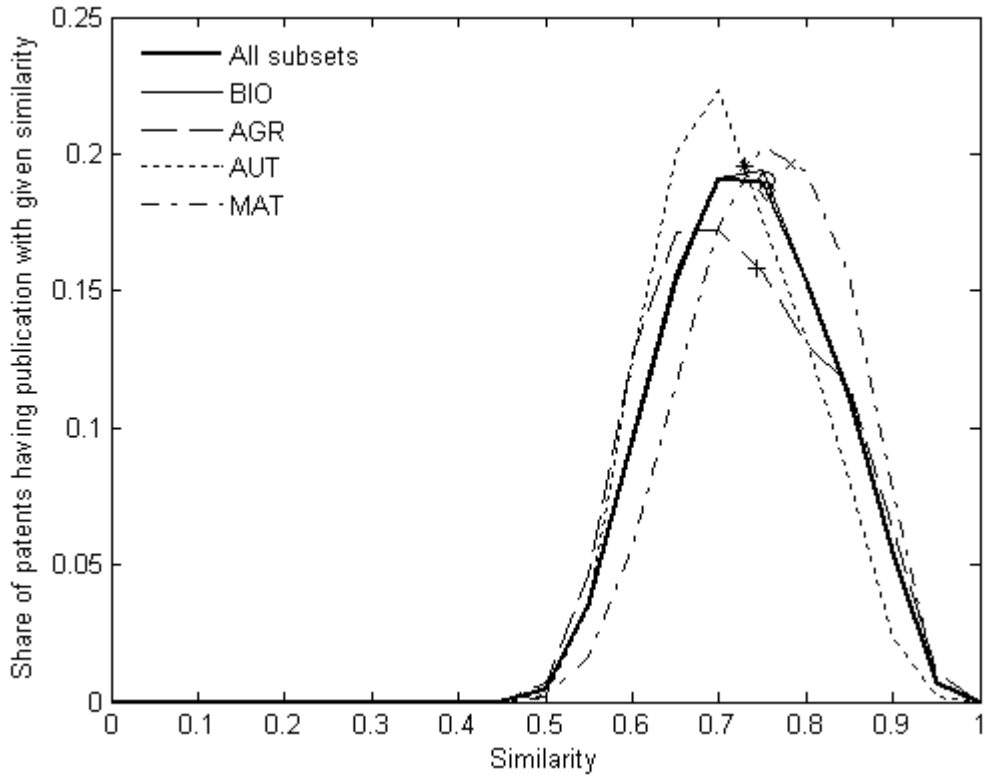
Comparison of distributions

To get a first look at the differences amongst measure variants, we compare the distribution of the obtained similarity scores amongst measures. For every measure, we take for every biotechnology patent the closest publication and the corresponding similarity score, hence 88,248 similarity scores for every measure variant. Based on those scores we derive relative distributions displaying the proportion of biotechnology patents having a closest biotechnology publication in a given similarity interval (one histogram for every similarity measure variant). We do the same for all patents related to agriculture, automotive and materials (again one histogram for every similarity measure variant and every control set). For any given measure variant, we can compare the distributions of the similarity scores of the biotechnology patents and the patents related to agriculture, automotive and materials as we used the relative share of patents having a closest publication within a given similarity interval and not the absolute number of patents.

Figure 6-1 shows the distribution of similarity scores for the patent groups (biotechnology and three control groups) for the similarity measure variant using TF-IDF weighting and SVD of rank 300, a commonly used measure. The Y-axis contains the proportion of patents having a closest publication with similarity given by the X-axis (with intervals of 0.05). Five distributions are combined: one for the biotechnology patents (solid thin line), one for every control group – agriculture (AGR), automotive (AUT) and materials (MAT) (non-solid lines) – and one for all patents together – biotechnology patents and all patents from all three control sets (thick solid line).

The distribution of similarity scores of the group of biotechnology patents (solid thin line) falls more or less together with the distribution of all patents (solid thick line) and almost has the same median value. Striking are the relative high similarity scores: the median similarity for all patents is 0.76, or 50% of all patents have a scientific publication with similarity above 0.76. These high average similarity scores are suspicious, although this might simply be a norming or calibration problem.

Figure 6-1 : Distribution of similarity scores of patents to closest publication according to TF-IDF SVD 300 (markers=median values)



More striking is the distribution of the similarity scores between patents related to materials and their closest biotechnology publication (dot-dash line). We expect the similarity score distributions of control set patents to be to the left of the similarity score distribution of biotechnology patents, as we expect that those control set patents are, on average, less related to biotechnology publications compared to biotechnology patents. However, here we observe that the distribution for materials patents is shifted to the right compared to the distribution of the group of biotechnology patents and the median value for these materials patents is 0.78. This means that, on average, patents related to materials are more closely related to biotechnology publications than patents related to biotechnology. This is very unlikely and suggests that similarity values based on TF-IDF weighting and SVD of rank 300 do not grasp the real relation between the patent and scientific publication documents.

We observe this phenomenon for all measure variants based on SVD, and the lower the number of retained dimensions, the worse (the more distributions of similarity scores

shift to the right and the less difference between the distribution of similarities for patents of the control groups – non-biotechnology patent to biotechnology publication - compared to the group of biotechnology patents). Weighting methods have some effect too: distributions based on binary weighting and IDF weighting are shifted more to the left compared to TF-IDF weighting and raw frequencies, regardless of the number of retained dimensions, and no weighting at all and binary weighting tend to suffer less from the phenomenon of patents of control groups being more similar to scientific biotechnology publications than biotechnology patents. SVD only seems to yield meaningful similarity values when a high number of dimensions are retained (500 or more) and not in combination with TF-IDF weighting (SVD with 1,000 dimensions and TF-IDF weighting still reveals unrealistic distributions).

Figure 6-2 : Distribution of similarity scores of patents to closest publication according to TF-IDF without SVD (markers=median values)

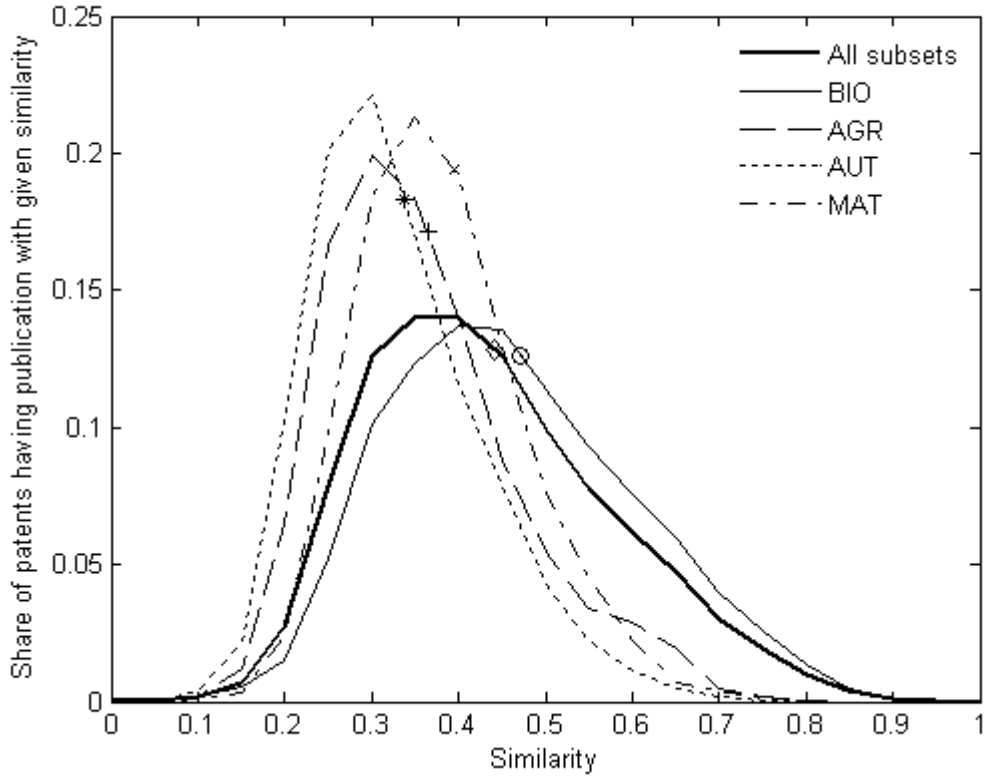
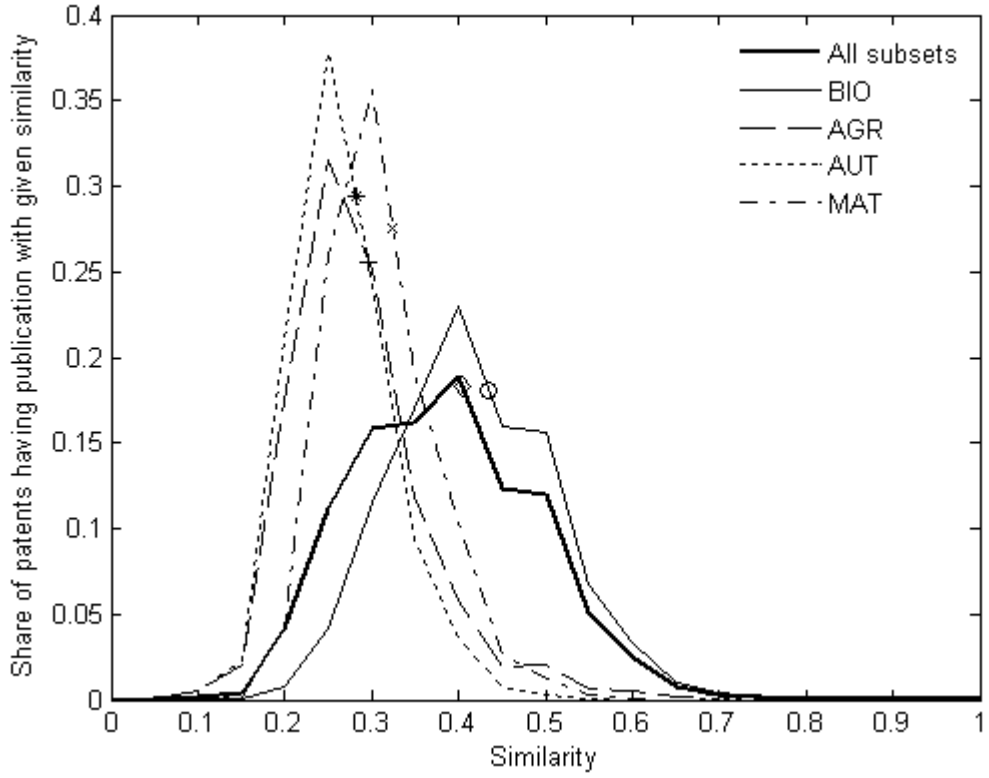


Figure 6-2 shows the distribution of similarity scores between patents and their closest biotechnology scientific publication according to the similarity measure variant using TF-IDF weighting without SVD dimensionality reduction. This distribution makes sense:

patents from control groups (agriculture, automotive, materials) are on average less similar to biotechnology patents. Even more, there are barely control set patents having high similarity with biotechnology patents. The other weighting methods, combined with no dimensionality reduction, yield similar distributions, although binary and IDF weighting results are slightly more peaked and shifted to the left.

Finally, Figure 6-3 shows the distribution of similarity scores of the measure variant based on the number of common terms normalized by the minimum of the term length of both documents ('common terms MIN').

Figure 6-3 : Distribution of similarity scores of patents to closest publication according to number of common terms normalized for minimum term length ('common terms MIN') (markers=median values)



Also here we observe an expected distribution with control set patents scoring significantly lower similarity scores compared to the biotechnology patents. All three measures based on the number of common terms yield expected results, although 'common terms MIN' yields the highest distinctive power between biotechnology

patents and control set patents. Even more, this measure seems to yield the best distinctive power of all measure variants under study

Preliminary conclusions

The comparison of the distribution of the similarity between patents and their closest scientific biotechnology publication and the pattern of biotech patent similarities compared to control patent similarities (agriculture, automotive, materials) amongst measure variants raises questions about the validity of LSA-based measures to match patent documents and scientific publications. Not only do LSA-based measures yield remarkably high similarity scores, they also do not score non-biotechnology patents as less similar to biotechnology publications compared to biotechnology patents, which suggests that these measure variants do not reflect real similarities present in the document collection. The less dimensions are retained, the more obtained similarity scores seem to deviate from the real relations between the documents. This effect is even reinforced when using TF-IDF weighting, a commonly used weighting method. Similarity measures based on the cosine metric without dimensionality reduction seem to perform better, in combination with any of the tested weighting schemas, as well as the three measures based on the count of common terms. The one normalized by the minimum of the number of terms of both documents ('common terms MIN') seems to yield the best results.

These remarkable results deserve a closer look to patent-publication combinations yielding high similarity values. Table 6-1 contains the similarity scores of a patent-publication combination scoring high on TF-IDF in combination with SVD (ranging from 0.928 to 0.995, depending on the number of retained dimensions). The title of the patent is: "Process and rotary milking parlor for the identification of a milking stall and an animal, in particular a cow, in a rotary milking parlor." And the title of the scientific publication is: "Growth-behavior of Lactobacillus-acidophilus and biochemical characteristics and acceptability of Acidophilus milk made from camel milk." Title and abstract of both documents make clear that both documents are only (very) slightly related; both are about milk, but the patent is about an apparatus for milking, while the publication is about a comparison of cow milk and camel milk for characteristics on

Lactobacillus acidophilus fermentation (see Appendix 6-2 for the full abstract of both documents). Obtained similarity scores contain considerable variation among weighting methods and dimensionality reduction options.

Table 6-1 : Similarity scores for patent US7104218 and publication A1994PC04400005 according to various measures

Weighting method	Dimensions retained (SVD)									
	ALL	1000	500	300	200	100	50	25	10	5
Raw	0.511	0.837	0.873	0.905	0.754	0.391	0.368	0.608	0.691	0.673
Binary	0.083	0.057	0.025	0.023	0.056	0.087	-0.030	0.492	0.763	0.750
IDF	0.095	0.168	0.162	0.260	0.375	0.403	0.504	0.532	0.698	0.738
TFIDF	0.364	0.928	0.973	0.986	0.991	0.991	0.995	0.980	0.959	0.960

Especially TF-IDF in combination with SVD yields high scores; other measures yield lower scores, better reflecting the limited relationship between both documents, although all weighting methods yield high values for high levels of dimensionality reduction (low values of *k*, right side of the table). In general, binary and IDF weighting yield lower scores compared to raw frequencies and TF-IDF weighting, although there are some exceptions. The measures based on the number of common terms yield low scores (0.10, 0.07 and 0.08 for 'common terms MIN', 'common terms MAX' and 'common terms AVG' respectively), in line with the real similarity between the two documents.

Mind also the non-linear relation between similarity scores and dimensionality reduction; lower number of retained dimensions do not necessarily yields the highest similarity scores (see e.g. the results for raw term vectors: starting from 0.837 for 1,000 dimensions it goes up to 0.905 for 300 dimensions, to go down to 0.368 for 50 dimensions to go up again for lower dimensions). This example proves again that the choice of the right level of dimensionality reduction is not straightforward and also that the weighting method has a considerable impact on the results.

6.5 First validation: comparison of the validity of the measures

Validation setup

To assess the validity of LSA-based measures and to get more insight in the contribution of weighting and dimensionality reduction levels in the performance of those measures,

we set up a validation at the level of individual patent-publication combinations. We select 250 patent-publication cases with variation in similarity scores amongst measure variants. For those 250 cases, we do an independent assessment of experts to rate the similarity on a five-level scale and we check the consistency between the expert assessment and the similarity scores obtained by each of the 43 measure variants for the 250 selected validation cases. This allows us to select the best performing measures.

Selection of 250 patent-publication combinations to validate

As almost all LSA-based measures tend to attribute high similarity scores to patent-publication combinations, we focus on the selection of patent-publication combinations with high obtained similarity scores to check whether these combinations are indeed similar. At the same time, we want to select patent-publication combinations for validation that have substantial variation in similarity scores amongst measures (it would not be very informative to select patent-publication combinations scoring high or low on all measures).

Starting point are the 88,248 biotechnology patents. For all those patents, the closest scientific biotechnology publication was selected according to a representative selection of 31 measures (all three measures based on common terms; the non-SVD cosine measure for the four weighting variants; and SVD cosine measure variants with their four weighting variants for *k*=10, 50, 200, 300, 500 and 1,000). For every measure in this selection, the 1,000 most similar patent-publication combinations are retained. After removal of duplicate patent-publication combinations (patent-publication combinations scoring within the top 1,000 for more than one measure), 16,717 patent-publication combinations are left. Out of this selection, 250 combinations were selected in groups of combinations that score high on one measure and low on as many other measures as possible[43] [44].

[43] 20 cases scoring high on 'common terms MIN'; 20 cases scoring high on 'common terms MAX'; 10 cases scoring high on 'common terms AVG'; 4x20 cases scoring high on the 4 weighting variants of the cosine measure without dimensionality reduction (one set of 20 cases for each weighting variant); 3x4x10 cases scoring high on the 4 weighting variants of the cosine measure with low (*k*=1000-500), medium (*k*=300-100) and high (*k*=50-5) dimensionality reduction respectively (one set of 10 cases for every weighting variant and dimensionality reduction level).

Expert assessment of 250 cases

A group of 9 people[45] rated all validation cases (patent-publication combinations) assessing the extent to which the content of the patent document and scientific publication cover the same invention/discovery using a five-level scale: not related at all (1), somewhat related (2), related (3), highly related (4) and identical (5). Every case was independently rated by two people: 176 cases got an identical score by the two raters; 21 cases got scores with a difference of one level; 8 cases got scores with a difference of 2 levels and 45 cases were judged complex. All complex cases and all cases with more than one level difference in scores were assessed by an additional rater resulting in 210 cases of total agreement; 39 cases of small disagreement (one level) and 1 remaining case of big disagreement (two levels).

The two independent scores were unified by taking the average of the two scores and rounded to arrive again at a 5 level score. To deal with the potential disagreement amongst raters, two final scores were retained: a conservative one by rounding the average of the scores down to the nearest integer, and an optimistic one by rounding the average up to the nearest integer. Table 6-2 contains the distribution of similarity levels amongst validated patent-publication combinations according to the conservative and optimistic validation.

Table 6-2 : Distribution of similarity levels amongst validated patent-publication combinations according to conservative and optimistic validation of experts

Score	Conservative	Optimistic
Identical	161	165
Highly related	8	15
Related	17	10
Somewhat related	10	27
Not related	54	33
Total	250	250

The fact that more than 50% of the cases are judged to be identical has to do with the selection method; our selection started from a set with the 1,000 most similar

[44] To compensate for the difference in distributions amongst measures, rank orders were used to evaluate high and low similarities instead of the absolute similarity values.

[45] All nine persons involved in the validation are familiar with patents and publications and IPR and 3 of them are also experts in biotechnology.

combinations for each and every measure, as we observe that SVD-based measures tend to attribute high similarity values to patent-publication combinations.

Check consistency between expert scores and 43 similarity measures

Given the expert assessment of the 250 validation cases, an ANOVA-type of analysis can be used to check the consistency between the expert scores (conservative and optimistic) and the calculated similarity values. Table 6-3 contains the results of the GLM regression based on 250 patent-publication validation cases for the conservative expert score. This table contains for every measure the R^2 value for the GLM regression with the conservative expert score as independent variable and the similarity values of the given measure as dependent variable (R^2 values higher than 0.50 are emphasized in bold and italic).

Table 6-3 : Congruence between (conservative) 5-level scale expert similarity assessment and calculated similarity measures (R^2 values of GLM regression based on conservative expert scores of 250 validation cases)

Measure		R^2	Measure		R^2
RAW	**No SVD**	***0.61***	TF-IDF	**No SVD**	***0.71***
	SVD 1000	0.34		**SVD 1000**	0.45
	SVD 500	0.31		**SVD 500**	0.34
	SVD 300	0.30		**SVD 300**	0.26
	SVD 200	0.31		**SVD 200**	0.21
	SVD 100	0.30		**SVD 100**	0.17
	SVD 25	0.22		**SVD 25**	0.14
	SVD 5	0.11		**SVD 5**	0.11
BIN	**No SVD**	***0.77***	IDF	**No SVD**	***0.80***
	SVD 1000	***0.65***		**SVD 1000**	***0.63***
	SVD 500	***0.63***		**SVD 500**	***0.57***
	SVD 300	***0.58***		**SVD 300**	***0.54***
	SVD 200	***0.51***		**SVD 200**	***0.51***
	SVD 100	0.45		**SVD 100**	0.49
	SVD 25	0.38		**SVD 25**	0.46
	SVD 5	0.20		**SVD 5**	0.21
Common terms (weighted by min number of terms)					***0.82***
Common terms (weighted by max number of terms)					***0.68***
Common terms (weighted by avg number of terms)					***0.75***

Mean R^2 values in bold denote values higher than 0.5.

Table 6-3 reveals that the application of SVD dimensionality reduction has a negative impact on the performance of similarity measures: for all weighting methods,

dimensionality reduction results in lower R^2 values, i.e. less congruence between the calculated similarity score according to the measure and the similarity level as assessed by the experts. And the larger the dimensionality reduction, the lower the obtained R^2 values. This is especially the case when raw frequencies or TF-IDF weighting is used – remarkable as the combination of TF-IDF weighting and SVD dimensionality reduction is commonly used. Binary and IDF-weighting (lower part of the table) outperforms raw frequencies and TF-IDF-weighting, whether or not SVD is used, and the combination of IDF-weighting without SVD, i.e. a cosine metric based on an IDF-weighted document-by-term matrix, yields the highest R^2 (0.80) of all cosine-based measures. Striking is also that simple metrics based on the number of common terms score very high, even more, the metric based on the number of common terms weighted by the minimum number of terms of both documents ('common terms MIN') yields the highest R^2 value (0.82).

When the optimistic expert scores are used instead of the conservative expert scores, results stay the same: despite small changes in R^2 (upward for some measures, downwards for others), conclusions about SVD dimensionality reduction, weighting method and best measures remain the same.

Also, when we convert the 5-level scale expert scores to 2-level scale expert scores (identical versus not-identical) to focus on the identification of patent-publication pairs, results stay the same.

First validation results

Our ANOVA results reveal that the similarity measure 'common terms MIN' best matches our expert validation. Of course it does not come to a complete surprise that measures based on the number of common terms perform that well: the more terms in common, the more you can expect both documents to be similar. But on the other hand these simple measures based on the number of common terms might miss relevant matches because they do not deal with language related issues like homonymy, polysemy and synonymy. It is remarkable that despite this lack of complexity these measures come closest to the expert assessment of similarity – clearly beating LSA measures that do claim to deal with typical language issues. Another remarkable observation is the consistency between 'common terms MIN' and the presence or

absence of a publication author in the list of patent inventors – a strong indication whether or not the patent and publication is identical, i.e. shares the same contents (methodology, findings, discovery). All patent-publication combinations with a similarity of 0.59 or above according to this measure do have a publication author listed as patent inventor, and all combinations with a similarity of 0.50 or below do not have a publication author listed as patent inventor (with one exception with a similarity value of 0.16). In between are 5 cases, 3 with and 2 without a publication author listed as patent inventor. This consistency is a strong indication of the validity of this measure[46].

If we take 0.55 as a threshold value (in between the zone with shared inventor/author and the zone without shared inventor/author) and translate the 5-scale expert score to a 2-scale score (identical or not identical – as we are primarily concerned about finding patent-publication pairs, hence primarily concerned about identical versus not-identical patent-publication combinations) we obtain a confusion matrix as displayed in Table 6-4 (using the conservative expert scores).

Table 6-4 : Confusion matrix for the measure based on the number of common terms weighted by minimum number of terms of both documents (based on conservative expert scores of 250 validation cases with threshold value of 0.55)

			Measure COMMON TERMS MIN	
			Identical	Not identical
			168	82
Expert opinion	Identical	161	160	1
	Not identical	89	8	81

This results in a precision of 0.95 (percentage of document combinations classified as related by the automated method that are correct according to the experts: 160/168 –

[46] Although the presence or absence of a shared inventor/author is a strong additional indication of content similarity, using this criterion on its own to identify patent-publication combinations is not straightforward – as stated in the introduction chapter – because of practical reasons (how to deal with spelling errors in names; presence or absence of initials and middle names; homonyms) and conceptual reasons (the same person can be involved in multiple discoveries/inventions, hence two documents of the same inventor/author can have a complete different contents). It is the combination of content relatedness and presence of shared inventor/author that yields a robust indicator.

to what extent is the automated method correct when it predicts a match) and a recall of 0.99 (percentage of document combinations that are related according to the experts that are classified as related by the automated method: 160/161 – to what extent does the automated method not miss relevant matches). When the optimistic expert scores are used, the number of identical patent-publication combinations according to the experts raise from 161 to 165, and this results in an even higher precision of 0.98 and recall of 0.99.

Although the validation seems to result in excellent precision and recall scores for an automated classification method, these scores are misleading because the distribution of the obtained similarity scores according to measure 'common terms MIN' is completely different within the validation sample and the total population. We have an underrepresentation of document combinations that are less related in our validation sample because the selection of validation cases was primarily based on combinations scoring high on at least one measure. As listed in Table 6-2, more than 65% of the cases in the validation sample were rated identical by the experts, while the relative number of patent-publication pairs in the full populations will be far, far lower. In reality, the number of document cases with average or low similarity scores will completely outnumber the cases with high scores. And although the relative number of misclassifications might by reasonable, this large group with average scores will result in a high number of mismatches in absolute terms, pulling down the relative number of correct classifications and negatively influencing precision and recall. To get reliable precision and recall scores, the relative number of mismatches has to be combined with the absolute number of document combinations to correct for the differences in distribution. We will come back on this issue later on when presenting validation results based on an extended validation set.

6.6 Additional validation: validation based on control sets

Validation setup

Apart from the expert validation, the control sets can be used for additional validation. As described earlier, three control sets were created with patents related to agriculture, automotive and materials, with 29,952 patents in total. These patents are presumed to

be unrelated to biotechnology publications, meaning that we do not expect to find biotechnology publications having a high content similarity with any of these control set patents.[47] If we apply our measure 'common terms MIN' on all combinations of the 29,952 control set patents and the 948,432 biotechnology publications, we expect not to find high similarity scores.

Additional validation results

In total we find 126 combinations of control set patents and biotechnology publications with a similarity value of 0.60 or above according to the measure 'common terms MIN' (about 0.04% of all control set patents), compared to 4,499 combinations of biotechnology patents and biotechnology publications (about 5.10% of all biotechnology patents). This significant difference in the ratio of patent-publication combinations with high content similarity between the group of biotechnology patents and the group of control patents is again an indication of the validity of our measure. Yet it might be interesting to dig into those 126 control set cases with high similarity. 51 of those cases have a similarity of 0.70 and above, and 12 cases even have similarities of 0.80 and above.

Appendix 6-3 contains an example of a combination of a control set patent and biotechnology publication (common terms min = 0.82; common terms max = 0.06). This example demonstrates the weakness of using the minimum number of terms of both documents as weight to normalize the number of common terms to arrive at a metric. The patent abstract is far shorter compared to the publication abstract, and as almost all terms of the patent abstract are present in the publication abstract, a high similarity is obtained when using the minimum number of terms of both documents as weighting factor. This approach seems to make sense in general; if the abstract of one document is a subset of the abstract of the other document, they can be regarded as identical. We checked this for multiple cases where there is a big difference in the similarity value based on measure 'common terms MIN' and measure 'common terms MAX' – an

[47] As stated before, patents of the control groups are selected in such a way that there is no overlap with biotechnology patents, i.e., patents classified in both biotechnology IPC classes and one of the control set IPC classes are not selected for the control groups, only for the biotechnology group. This is of particular interest for the agriculture control group, as this group can be related to biotechnology and share some IPC codes (A01H 1/00 and A01H 4/00).

indication of document combinations with unbalanced text length – and indeed, for the vast majority of those cases the longer document just contains more details or a longer introduction or results, but the actual relevant contents is the same. So there is some anecdotic evidence to back up the use of the minimum number of terms as weight (on top of the empirical results of the ANOVA-analysis revealing this measure as the best performing one). However, when one of the documents is too small, or when the difference in length is too big, using the minimum number of terms as weight leads to unreliable results (even for human experts it becomes difficult to assess similarity for these cases).

If we go back to our 126 control set patents with high similarity with a biotechnology publication (weighted by the minimum number of terms of both documents – measure 'common terms MIN'), it is striking that all of them do have low similarity values when the maximum number of terms of both documents is used as weight (measure 'common terms MAX') - i.e. there is a big difference in the length of both documents. Only 21 of those cases have a similarity 'common terms MAX' above 0.10, and only 2 above 0.20 (with a maximum of 0.25). In our validation set of 250 cases, 71 cases have a similarity 'common terms MAX' of 0.25 or below; and only 2 of those cases are rated as identical by the experts (one cases of 0.24 and one case of 0.20).

Additional criterion

The insights of the additional validation suggest that a correction is needed for document combinations with one small and one large document. For those cases, our best performing measure 'common terms MIN' might be misleading and an additional criterion based on document length is needed. Instead of adding an absolute criterion based on document size, we examine the impact of an additional relative measure, as we have already one measure available: measure 'common terms MAX'. So we combine the primary criterion based on measure 'common terms MIN' (e.g. above 0.55) with a secondary criterion based on measure 'common terms MAX' to correct for doubt cases. The results of the additional validation based on the control sets suggests that the threshold for this secondary criterion 'common terms MAX' is somewhere between 0.20 and 0.30 (almost all combinations from the control set score below 0.20 on this

criterion with a few exceptions between 0.20 and 0.30, and all combinations in our validation sample of 250 expert rated cases scoring below 0.20 on this criterion are rated not identical by the experts).

Applying this secondary criterion 'common terms MAX' with threshold around 0.20 does not influence the classification of the 250 expert rated cases in our validation sample because none of those cases with primary criterion 'common terms MIN' above 0.55 score below 0.20 on the secondary criterion (4 cases score between 0.20 and 0.30, and all four are rated identical by the experts).

However, the control set validation proves that setting the threshold value for the secondary criterion 'common terms MAX' does has a significant impact (e.g. setting the value to 0.20 would discard all matches found for the control set patents). If we look at the global biotechnology dataset (88,248 biotechnology patents and 948,432 biotechnology publications), and take the 1,000 closest combinations for every patent according to measure 'common terms MIN', there are 112,847 patent-publication combinations above 0.55 for the primary criterion 'common terms MIN', but the vast majority of those combinations score low on the secondary criterion 'common terms MAX'. Table 6-5 contains the distribution of the 'common terms MAX' scores for the combinations above 0.55 for 'common terms MIN'.

Table 6-5 : Distribution of second criterion scores 'common terms MAX' for all patent-publication combinations with primary criterion 'common terms MIN' above 0.55

Second criterion 'COMMON TERMS MAX' range	Number of patent-publication combinations	Number of patent-publication combinations (cumulative)
0.35 ≤ x	631	631
0.30 ≤ x < 0.35	262	893
0.25 ≤ x < 0.30	747	1,640
0.20 ≤ x < 0.25	2,856	4,496
0.15 ≤ x < 0.20	14,053	18,549
0.10 ≤ x < 0.15	56,093	74,642
0 ≤ x < 0.10	38,205	112,847

The figures in table Table 6-5 make clear that matching results are extremely sensitive to threshold setting; even within the range of 0.20-0.30 the impact on the number of

matches is significant (from 4,496 combinations labelled as identical by the automatic method for a threshold value of 0.55 for 'common terms MIN' and 0.20 for 'common terms MAX' to 893 combinations labelled as identical for the same threshold value for 'common terms MIN' but a threshold value of 0.30 for 'common terms MAX').

6.7 Final validation: selection of 50 additional cases for expert validation

Validation setup

As the first validation set of 250 cases for validation cases does not allow for careful selection of the threshold value for the secondary criterion - as we do not have enough cases with low scores on 'common terms MAX' in our validation sample - 50 additional cases were selected. For the selection of these additional cases, we do not only look for cases with low scores on the secondary criterion 'common terms MAX', but also for potential false negatives and false positives for the primary criterion 'common terms MIN'. The idea is the create a robustness check for a classification method based on 'common terms MIN' and 'common terms MAX' by deliberately selecting additional validation cases that are though, i.e. difficult to classify because they are in the grey zone between identical and not-identical combinations or cases that are expected to be misclassified based on the information we have. To obtain a balanced and representative selection, we split up the selection of additional validation cases by similarity range for the primary criterion 'common terms MIN':

0.81-1.00 : this is normally the save zone were we only expect to find patent-publication combinations that are identical (the first validation only revealed one case not rated as 'identical' by the experts). For this zone, we are interested in potential false positives introduced by the primary criterion, so we select 5 cases without shared inventor/author because these are unlikely to be identical (all those cases happen to have a low score on the secondary criterion).

0.71-0.80 : this is still rather a save zone (the first validation only revealed one or two cases not rated as 'identical' by the experts - depending whether conservative or optimistic expert scores are used). For this zone, we are interested in potential false positives introduced by the primary criterion and in potential false negatives introduced

by the secondary criterion. We take 5 cases without shared inventor/author (potential false positives) and with high scores on the secondary criterion (how to set second criterion threshold to discard false positives); 5 cases with shared inventor/author and low scores on the secondary criterion (to what extend will the secondary criterion introduce false negatives); and 10 cases with shared inventor/author and a secondary criterion value around 0.3 (to help finding a solid threshold for the secondary criterion), so 20 cases in total.

0.61-0.70 : this is a grey zone with multiple mismatches according to the first validation. We follow the same logic for the selection of cases as for the previous range: 8 cases without shared inventor/author and with high scores on the secondary criterion and 7 cases with shared inventor/author and low scores on the secondary criterion, so 15 cases in total.

Below 0.61 : here we are interested in false positives in the frontier zone ('common terms MIN' in the range of 0.55-0.60) and in false negatives for lower values on the primary criterion 'common terms MIN'. We select 5 cases scoring high on the primary criterion (within this range) and without shared inventor/author or low value on the secondary criterion, and 5 cases scoring low on the primary criterion (within this range) and with shared inventor/author.

Final validation results

Those 50 cases were again rated by two experts as in the first validation. Table 6-6 contains the result of the validation for every range and subset based on conservative expert scores.

Table 6-6 : Expert validation results (conservative) for 50 additional cases by primary criterion range ('common terms MIN') and validation subset

Range primary criterion	Validation subset	Total cases	IDENTICAL ACCORDING TO EXPERTS		NOT IDENTICAL ACCORDING TO EXPERTS	
			Cases	Range secondary criterion	Cases	Range secondary criterion
0.81-1.00	Potential false positives	5	0		5	0.11-0.18
0.71-0.80	Potential false positives	5	1	0.28	4	0.25-0.33
0.71-0.80	Potential false negatives	5	0		5	0.10-0.16
0.71-0.80	Secondary criterion around 0.3	10	9	0.31-0.41	1	0.36
0.61-0.70	Potential false positives	8	1	0.32	7	0.29-0.45
0.61-0.70	Potential false negatives	7	2	0.24-0.26	5	0.20-0.26
< 0.61	Potential false positives	5	1	0.35	4	0.30-0.51
< 0.61	Potential false negatives	5	2	0.36	3	0.35-0.40

For the first range ('common terms MIN' in the range of 0.81-1.00) results are good for the false positives: all cases that score high on 'common terms MIN' but that were suspect of being false positives (because they had no shared inventor/author) are rated as not identical by the expert validation. It is clear that the secondary criterion based on 'common terms MAX' easily discards all those false positives with a clear threshold value of 0.18).

For the second range (0.71-0.80), results are still reasonable, although a proper selection of the threshold value for the secondary criterion is not clear. A threshold value below 0.36 will introduce false positives, but a threshold value above 0.28 will introduce false negatives, so no clear cut-off point exists and a trade-off has to be made (e.g. a threshold value of 0.30 results in 3 false positives and 1 false negative).

For the third range (0.61-0.70), the overlap gets bigger and the choice for a threshold value for the secondary criterion gets complicated. A threshold value below 0.45 will introduce false positives, but a threshold value above 0.24 will introduce false negatives, so again no clear cut-off point exists and a trade-off has to be made (e.g. a threshold value of 0.30 results in 2 false negatives and 5 false positive).

Finally, the last range (< 0.61) requires even higher values for the secondary criterion to discard false positives (0.51) but the overlap is not much bigger compared to the previous range (at least 0.36 to prevent false negatives). We observe identical combinations (according to the expert validation) up to a similarity of 0.46 for 'common terms MIN', but distinction from false positives is difficult, even with 'common terms MAX' as secondary criterion.

Based on these results, it makes sense to add a secondary criterion based on the number of common terms weighted by the maximum number of terms of documents to eliminate potential false positives with minimal introduction of false negatives. But the results in Table 6-6 also reveal that for lower values of 'common terms MIN' a clear distinction between identical and non-identical combinations is not possible.

It is clear that setting thresholds on the primary criterion ('common terms MIN') and secondary criterion ('common terms MAX') is a trade-off between false positives and false negatives, or precision and recall.

Table 6-7 contains precision and recall for different thresholds on the primary and secondary criteria (optimal precision, optimal recall, and balanced precision/recall) based on all 300 cases rated by experts (both for the conservative and optimistic expert scores).

Table 6-7 : Precision and recall for different thresholds on primary and secondary criterion (optimal precision, optimal recall, balanced precision) (based on conservative and optimistic expert scores for 300 validated cases)

Primary criterion	Secondary criterion	CONSERVATIVE EXPERT OPINION		OPTIMISTIC EXPERT OPINION	
		Precision	Recall	Precision	Recall
0.50	0.10	0.81	0.99	0.88	0.98
0.50	0.32	0.91	0.92	0.94	0.88
0.50	0.61	0.98	0.55	1.00	0.51
0.55	0.10	0.82	0.98	0.88	0.97
0.55	0.30	0.90	0.93	0.93	0.89
0.55	0.61	0.98	0.55	1.00	0.51
0.60	0.10	0.83	0.95	0.98	0.94
0.60	0.29	0.91	0.92	0.94	0.88
0.60	0.61	0.98	0.55	1.00	0.51

Optimal precision scores can be obtained with a recall around 0.55/0.51, optimal recall scores can be obtained with a precision around 0.81/0.88 and balanced precision/recall scores around 0.90 are possible for both precision and recall at the same time (e.g. 'common terms MIN' above 0.55 and 'common terms MAX' above 0.30).

As stated before, we have to keep in mind that the precision and recall figures listed in Table 6-7 are not representative for the total population because obtained similarity values for 'common terms MIN' and 'common terms MAX' are not equally distributed in the validation sample and the total population (very high number of identical document combinations in the validation sample). As there are only a very limited amount of document combinations scoring high on the proposed measures in the total population, it is more appropriate to derive precision and recall measures based on validation cases scoring around the threshold values. Take for instance a threshold value of 0.55 for 'common terms MIN' and 0.30 for 'common terms MAX'. According to the conservative expert validation, this would result in a precision of 0.90 (184 combinations classified as pair by the automated method in the validation set of which 165 are real pairs according to the experts) and a recall of 0.93 (165 real pairs retrieved by the automated method in the validation set compared to 177 real pairs identified by the experts). Applied on the full population this would mean that we would label 893 patent publication combinations as identical (see Table 6-5), and by doing so, about 89 of those cases would be wrongly labelled identical (10%), and at the same time we would miss about 61 cases (7%). However, if we look at the cases with 'common terms MAX' between 0.30 and 0.25 in our validation set, we find 9 cases of which 5 cases are assessed as identical by the experts, or 55%. These matches will be missed by the automated method when the threshold for the secondary criterion is set to 0.30. If this 55% match rate is representative for the whole population in the range of 0.25 and 0.30 for 'common terms MAX', we would miss 411 patent-publication pairs in this range for 'common terms MAX'[48]. In the range of 0.20-0.25 for 'common terms MAX', we observe 34% matches in the validation sample. Again according to Table 6-5 we would miss an

[48] According to table Table 6-5, we have 747 cases in the total population with 'common terms MIN' above 0.55 and 'common terms MAX' in the range of 0.25 and 0.30, of which 55% or 411 cases are expected to be identical according to the validation.

additional 971 patent-publication pairs. If we continue this reasoning, we and up with a match rate of 20% for the range 0.15-0.20 resulting in another 2,811 missed patent publication pairs. This makes a total of 4,193 expected missed patent-publication pairs, far more than the 61 cases we initially expected. According to these estimations, the real recall rate is 16%.

The same problem occurs for precision rates. In the range 0.30-0.35 for 'common terms MAX', the precision rate in the validation set 53%. Again according to table Table 6-5 this would mean already 123 false positives. For 'common terms MIN' equal to 0.35 and above, precision rate is 94% hence 38 additional false positives, or 161 false positives in total, again more than the 89 false positives initially expected according to the precision rates in table Table 6-7. According to the estimations based on the full dataset, the real precision rate is 82%.

The button line is that precision and recall rates derived from the validation sample are not representative for the whole population because we have far, far more patent-publication combinations scoring low on the proposed distance measures while we initially calculated precision and recall rates from sample data with an overrepresentation of patent-publication combinations scoring high on the respective distance measures. Especially recall rates are suffering from this issue. However, the magnitude of the difference between the precision and recall rates averaged over the validation sample and the real rates based on the distribution in the global population heavily depends on the representativeness of those cases scoring low in the validation set. As the previous examples describes, these derived numbers are based on only 28 cases scoring less than 0.30 for 'common terms MAX' (given a score of 0.55 or above on 'common terms MIN'). More validation cases with lower scores are needed to get a more reliable estimate of the real precision and recall. But to be at the save side, the threshold on 'common terms MIN' has to be increased to get acceptable precision rates (e.g. to 0.60).

Precision can be improved by introducing a third criterion: the presence of a shared inventor/author. Although this extra criterion helps to make results more robust, large scale application on big datasets might not be straightforward.

6.8 Where does it go wrong for TF-IDF and SVD

Weighting issues

The most remarkable finding of this study is the bad performance of SVD-based measures, even with commonly used pre-processing options and levels of dimensionality reduction (e.g. TF-IDF weighting in combination with SVD with 300-1,000 dimensions).

When it comes to the influence of weighting, Table 6-3 reveals that weighting methods taking into account term frequencies (raw frequencies and TF-IDF weighting) perform worse compared to weighting methods ignoring term frequencies for all levels of dimensionality reduction. In line with these findings we also observe better performance for the measures based on the number of common terms, measures which also ignore term frequencies.

Looking at individual cases gives some insight in the implications of the choice of a weighting method. In general, including term frequencies is expected to generate better results as the number of times a given term appears in one document is an indication of the importance of that term in that particular document. However, for our patent-publication document combinations (mostly of a rather moderate length and with highly technical content), this additional notion of importance derived from term frequencies seems to be of less relevance in the assessment of similarity of the documents. Indeed, when looking at multiple document combinations, the human judgement on similarity is far more driven by the kind of terms in the documents rather than the number of times a particular term appears in a document. This observation explains why weighting methods taking into account term frequencies do not perform better, but not why they perform worse. Again looking at individual cases reveals some additional insights.

First of all, stemming errors and tokenization and parsing issues sometimes cause artificial inflation of term frequencies, magnifying the impact of the underlying stemming and tokenization errors.

Appendix 6-4 contains an example of a patent-publication combination where the amplification of a stemming error results in misleading similarity scores for weighting methods taking into account term frequency. The patent document is about an incubator with external gas feed. The publication document is about gibberellin metabolism in suspension-cultured cells of raphanus-sativus. Both documents have nothing in common, yet score high on some measures (and score significantly higher for measures including term frequencies). Both documents have only two (stemmed) terms in common, 'feed' and 'ga'. But the stemmed term 'ga' occurs 9 times in the patent document and 29 times in the publication document, resulting in high weights when the term frequency is included. But the stemmed term 'ga' in the patent document is a stemming error derived from 'gas', while the stemmed term 'ga' in the publication document is an abbreviation of 'gibberellin' and has nothing to do with the stemmed term 'ga' in the patent document. For weighting methods not taking term frequency into account, this stemming error counts as just one (be it wrongly) matching term, but for weighting methods using term frequency, this stemming error is magnified and leads to erroneous results.

Appendix 6-5 contains an example of a patent-publication combination where tokenization and parsing issues result in misleading similarity values for weighting methods taking into account term frequencies. Again both documents are not related and have only two terms in common: 'alpha' and 'beta'. Both of these terms occur a lot in both documents as part of chemical formulas, and these high term frequencies result in higher similarity values for weighting methods based on term frequencies. But the larger chemical formulas these terms are part of, are not related. It would probably be better to parse and index those formulas as one piece, but this is not straightforward.

Secondly, we observe that words with a particular meaning and hence very relevant in the assessment of similarity tend to have smaller term frequencies compared to natural language words with a more general meaning. For weighting methods including term frequencies, the weight of these more general natural language words becomes too influential in the derivation of similarity by the cosine metric. This issue might be specific to the technical nature of the documents – i.e. for our set of patent and

publication documents, low frequency technical words are far more important for the assessment of similarity compared to higher frequency natural language words. Weighting terms by their respective IDF values does only partially correct this problem; TF-IDF performs better than no weighting at all, and IDF performs (slightly) better than binary weighting, but TF-IDF still performs worse compared to binary or IDF weighting.

Given these insights, it might be worthwhile to investigate to leave stemming out of the pre-processing steps and to devote additional efforts for a more advanced tokenization and parsing, especially to better deal with chemical formula. Another approach is to improve feature selection to eliminate or further down-weight terms which are too general in meaning to be significant in the derivation of similarity.

SVD issues

It is not clear why LSA – or SVD – fails, or why SVD tends to assign unrealistic high similarity scores to document combinations - mind in that respect the high similarity scores for the patents in the materials control set. While some anecdotic evidence exists to explain differences in weighting performance, disentangling the bad performance of SVD in general is of a different level of complexity. Looking at individual cases is not very informative as it is virtually impossible to trace back term vectors after SVD to the original terms and contents. The document-by-concept matrix compiled by the SVD solution contains the scores of all documents on newly formed latent concepts, and every latent concept consists of a linear combination of all original terms, i.e. a linear combination with 301,697 components.

There are however some general reasons why LSA or SVD might fail for our dataset. The first reason is that the dataset might not be large enough to derive the latent structure. This is however very unlikely for our dataset as it contains almost one million documents. A related issue might be that the individual documents are not long enough to grasp the contents of the documents. This issue might be relevant for our dataset as we work with titles and abstracts, and especially patent abstracts tend to be rather

small (about 39 unique terms on average for patents and 65 unique terms on average for publications[49]). We will come back to this issue in the next sections.

Another reason for the unfulfilled expectations might be that the chosen levels of dimensionality reduction are not appropriate for our dataset, or that our derived SVD solutions for our selection of *k*-values accidentally do not grasp the latent structure of the data. The former deserves more attention, although literature suggests 300 to 1,000 concepts is enough to capture the topics in a document set (see chapter 4), which is the range we included in our setup. We will also come back to this issue in the next sections. The latter is very unlikely: SVD solutions are based on the singular values of the full document-by-term matrix. Changing the number of retained dimensions/concepts does not alter the values of the singular values, it only alters the number of singular values and singular vectors taken into account to approximate the original document-by-term matrix. As singular values are ordered by magnitude and values drop significantly, small changes in the number of retained dimensions cannot have big effects once beyond the first tens or hundreds of singular values. Moreover, we have four fundamentally different SVD derivations because of the four weighting methods, and all those variants yield SVD based measures that underperform compared to cosine measures on the full vector space.

A complete different kind of issues resides in the technical nature of the documents. Maybe the specific context of patent and publication documents does not allow the method to achieve its full potential. LSA is intended to derive meaning from text based on large samples of 'narrative' documents. The distinctive language use within our dataset might not be appropriate (might especially be of a concern for patent documents where phrasing might reflect tactical and strategic consideration more than technical disclosure, e.g. to maximize legal claims to get broad application protection or to disguise the real contents to mislead competitors). A related issue might be that we are combining patent documents with scientific publications, two document spheres that might be too different to derive a latent structure that fits both. These potential causes of failure might look farfetched, but it is clear when reading patent abstracts that

[49] After stop word removal, stemming and removal of words appearing in only one document.

such documents have little in common with typical applications as e.g. the ones described in the Handbook of Latent Semantic Analysis (Landauer, McNamara et al., 2007). However, finding evidence for these raised issues is not straightforward. One could set up validation exercises as the ones deployed in this chapter to evaluate LSA performance on a distinct subset with only patents and a distinct subset with only scientific publications. Another avenue might be to compare the LSA performance when more descriptive abstract are used, e.g. using the *Derwent Abstracts* as available in the *Derwent World Patent Index* (Thomson Reuters Derwent World Patent Index), which are abstracts rewritten by scientifically-trained editors detailing claims and disclosures of the invention and highlighting main use and advantages. Pursuing such additional research efforts might reveal interesting information on the applicability of LSA on patent and publication data, but goes beyond the limitations of this dissertation.

One final reason why LSA might fall short is the limitation to Euclidean geometry as imposed by the assumption of LSA that documents are represented as vectors in a vector space. In an Euclidean space, similarity should be symmetric and not violate triangle inequality - $d(x,z) \leq d(x,y) + d(y,z)$ - placing strong constraints on the location of point in a space given a set of distances (Griffiths, Steyvers & Tenenbaum, 2007). This issue however is a general one and not directly related to the limitations of our patent and publication dataset; it is related to problems when dealing with high-dimensional spaces ('curse of dimensionality'). In high dimensional spaces all data appear to be sparse and dissimilar, preventing efficient identification of communalities. Other text mining techniques not relying on spatial representations might be more appropriate, like generative topic models as Probabilistic Latent Semantic Modelling (Hofmann, 1999) and Latent Dirichlet Allocation (Blei, Ng & Jordan, 2003), but the exploration of those methods are again beyond the limits of this dissertation.

In the next sections, we will elaborate more on the impact of document size and the impact of the number of retained dimensions/concept on the performance of SVD.

Impact of document size on SVD performance

The poor performance of SVD might be related to the document size, as especially patent abstracts tend to be short. To get more insight in this issue, we include

document size when we check for the congruence between obtained calculated similarity scores and the expert validation scores. For all patent-publication combinations in the validation sample, we use the minimum document size, i.e. the minimum of the number of terms of the patent document and the publication document, as indicator of the document length. If we include this in the regression analysis, i.e. if we take again the compiled similarity measure variants as dependent variables and we take the expert score and the document size as independent variables, results reveal that document size has no impact on the similarity scores for our measure 'common terms MIN', but that the impact is significant (at the 5% level) for all cosine based measures without SVD. For binary and IDF weighting in combination with SVD, document size also has a significant impact on the similarity score, but not for SVD in combination with TF-IDF weighting or raw frequencies. Overall, the impact is small, except for binary and IDF weighting in combination with SVD, where R^2 values can improve with 8 to 11 percentage points compared to the model with only the expert score as independent variable. Whether we use the total number of terms or the number of distinct terms does not make a lot difference, although results are somewhat softened when the number of distinct terms is used to derive the document size indicator.

Likewise, we also had a look at the difference in document size within a patent-publication document, as we know there are many combinations with a small document combined with a large document. Now we used the ratio between the number of terms of the smallest document and the number of terms of the largest document as indicator of document size difference. If we take the expert scores and the document size difference as independent variables, we see comparable results as for the document size effect, but with stronger impact. Again the impact is not significant for our measure 'common terms MIN' but is significant for all cosine based measures without SVD, for binary weighting and IDF weighting with SVD, and for TF-IDF weighting and raw frequencies with SVD 1,000. The impact of the document size difference is higher than the impact of the document size, with R^2 values increasing with 15 to 20 percentage points for binary and IDF weighting with SVD.

If we also include both document size and document size difference in the model, and the interaction between document size and document size difference, we observe that the significance of the document size disappears and the impact of document size difference remains.

These results tend to suggest that SVD based measures in combination with binary and IDF weighting are influenced by the document size difference, which might be an explanation for the poor performance. However, this is not a complete explanation as this impact is rather moderate for SVD in combination with TF-IDF weighting and no weighting at all, while those measures perform worst.

In a last analysis trying to disentangle the relation between document size and performance, we looked at direct influence of document size and measure performance. All patent-publication pairs in the validation sample where uniformly divided into three groups: group one with small documents (measured as before by the minimum number of terms of the patent and publication document); group two with medium size documents; and group three with large documents. Now we perform a regression analysis with the similarity measures as dependent variable and the expert scores as independent variable for each of the three groups. For measure 'common terms MIN', performance goes down for larger documents (R^2 of 87% for small documents to 61% for large documents). For cosine based measures without SVD, performance of binary and IDF weighting also goes down by about 15 percentage points; for raw frequencies performance goes up considerable (R^2 from 38% to 72%), and for TF-IDF weighting performance remains more or less constant (R^2 around 64%). For SVD with low levels of dimensionality reduction (k=1,000) we see the performance slightly going down for larger documents for binary and IDF weighting, but heavily going up for TF-IDF weighting and no weighting. If we do the same kind of analysis with document size difference instead of document size, we observe a considerable increase in performance of SVD with low levels of dimensionality reduction (k=1,000) for TF-IDF weighting and raw frequencies for patent-publication combinations that are balanced in document size, while performance remains more or less constant for binary and IDF weighting.

To summarize, we see that document size and document size difference has a different impact on the measures depending on the weighting schema used. The three measures based on the number of common terms, and the cosine measures without SVD and without term frequencies (binary and IDF weighting) tend to perform worse for larger documents and/or documents of equal size. There seems to be a normalization problem for these measures. More important is that SVD-based measures and especially TF-IDF measures in combination with low levels of dimensionality reduction perform far better for larger and more balanced patent-publication combinations, although not yet beating our preferred measures 'common terms MIN'. In this respect it would be interesting to combine both document size and document size difference in the same analysis. The problem is that the number of observations in the validation set becomes low for some combinations of document size and document size difference, and that the variance amongst expert scores becomes very low for some of these combinations.

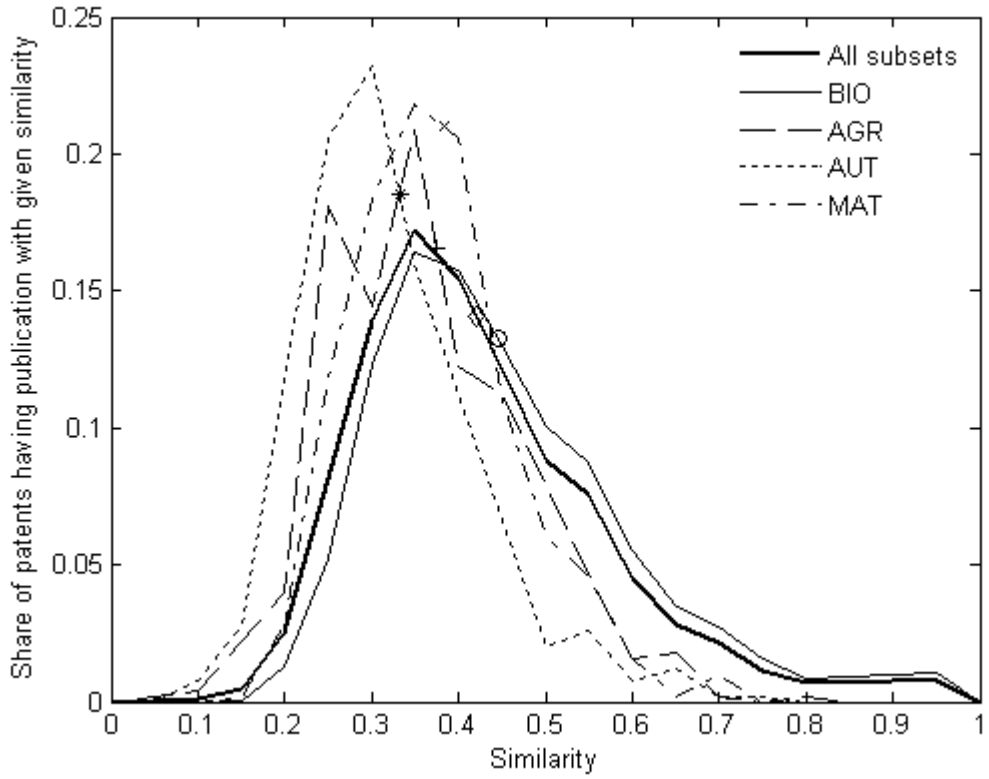
We have to be careful in deriving hard evidence from these analyses, as the validation sample is rather small - especially when split up by size and size difference - but it seems that document size and document size difference have an impact on SVD based measures and might at least partially explain the bad performance of e.g. TF-IDF weighting in combination with SVD because of our rather short patent abstract documents. A larger validation sample with a more balanced design when it comes to document size and document size difference is required to disentangle this further.

Impact of the number of retained dimensions/concepts and stability of the SVD solution

As stated before, the choice of *k*, the number of retained dimensions or concepts, is not straightforward. In this study, the maximum value of *k* taken into consideration was 1,000 because of computational limitations. Although literature suggest to take 100 to 300 concepts (see chapter 4), the variety of topics present in our patent and publication set might require more concept to be taken into account to grasp the latent structure of the dataset. Computationally limitations prevent us from deriving SVD solution with more than 1,000 retained dimensions/concepts for our large dataset, but using a smaller sample allows us to go beyond 1,000 retained concepts in the SVD calculation.

We started from the original raw document-by-term matrix, the document-by-term matrix after IDF weighting, and the document-by-term matrix after TF-IDF weighting. Every time we selected 5% random patents from each group (biotechnology, agriculture, automotive and materials) and 5% random publications from our original dataset. However, we made sure that patent-publication combinations that are present in our expert validation set are also present in all samples. This resulted in three different subsamples of 53,332 patent and publication documents, based on three different weighting methods (mind that not only the weighting method is different, but that selected patents and publications are also different, except for the patent-publication combinations present in our validation sample, which are present in all three subsamples). For all three subsamples, we performed SVD with k=5,000 and calculated distances between all patents and publications within the samples[50].

Figure 6-4 : Distribution of similarity scores of patents to closest publication according to TF-IDF SVD 5000 (based on 5% sample) (markers=median values)



[50] Going beyond 5,000 retained dimensions/concept when deriving an SVD solution from our 5% sample takes an extremely amount of computing time and was not feasible for more than one variant.

This results in the same kind of information as described in section 6.4 ('Aggregated results'), except for a smaller sample. This means that we can again plot distributions and compare similarity scores of patents and their closest publications for the biotechnology patents and control group patents. Figure 6-4 shows the distribution of similarity scores for the similarity measure using TF-IDF weighting and SVD of rank 5,000. The distributions are more shifted to the left compared to the distributions for TF-IDF with lower rank SVD, and there is also a clear distinction between biotechnology patents and control set patents. In short, these distributions are more in line with the expectations, and with the results of TF-IDF weighting without SVD or the similarity measures based on the number of common terms. For the raw document-by-term matrix, and the one after IDF weighting, we find the same kind of results when applying SVD with rank 5,000. In short, we find realistic distributions when SVD with a high number of retained dimensions/concepts is used. Remind that the selection of patents and publication is different for the three samples, so we observe the same kind of improvement for 3 independent subsets. As all validated patent-publication combinations are present in all three subsets, we can again check the congruence between the obtained similarity scores according to those three measures and the expert scores, as we did in section 6.5 ('First validation: comparison of the validity of the measures').

Table 6-8 is an extension of Table 6-3 for the three weighing variants for which we derived a 5% sample and calculated a rank-5,000 SVD solution, and contains again the results of the ANOVA-type of analysis to check consistency between the expert scores and calculated similarity scores. We see that higher number of retained dimensions/concepts have a significant positive effect; for IDF weighting and TF-IDF weighting, obtained results even approach the variants without SVD.

Table 6-8 : Congruence between (conservative) 5-level scale expert similarity assessment and calculated similarity measures, including high rank-*k* SVD based on 5% sample (R^2 values of GLM regression based on conservative expert scores of 250 validation cases)

Measure		R^2	Measure		R^2
RAW	**No SVD**	***0.61***	**TF-IDF**	**No SVD**	***0.71***
	SVD 5000 (5% sample)	***0.56***		**SVD 5000** (5% sample)	***0.68***
	SVD 1000	0.34		**SVD 1000**	0.45
	SVD 500	0.31		**SVD 500**	0.34
	SVD 300	0.30		**SVD 300**	0.26
	SVD 200	0.31		**SVD 200**	0.21
	SVD 100	0.30		**SVD 100**	0.17
	SVD 25	0.22		**SVD 25**	0.14
	SVD 5	0.11		**SVD 5**	0.11
BIN	*NA*		**IDF**	**No SVD**	***0.80***
				SVD 5000 (5% sample)	***0.79***
				SVD 1000	***0.63***
				SVD 500	***0.57***
				SVD 300	***0.54***
				SVD 200	***0.51***
				SVD 100	0.49
				SVD 25	0.46
				SVD 5	0.21
Common terms (weighted by min number of terms)					***0.82***
Common terms (weighted by max number of terms)					***0.68***
Common terms (weighted by avg number of terms)					***0.75***

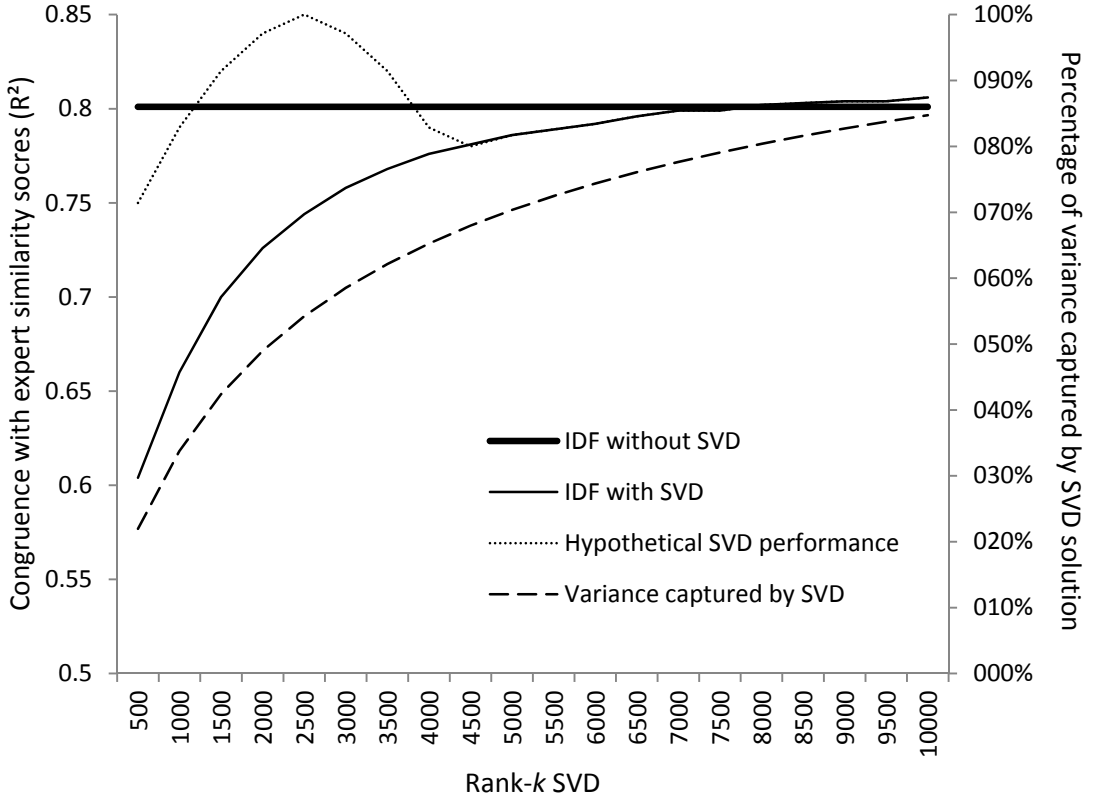
Mean R^2 values in bold denote values higher than 0.5.

To be sure obtained R^2 values are not the result of an accidental good fit of the particular SVD solutions, we created a second 5% subsample based on IDF weighting, derived another rank-5000 SVD solution, calculated similarities of validated patent-publication combinations again and checked congruence with the expert scores, and obtained an R^2 value of 0.793, remarkably close to the R^2 value of 0.786 of the first 5% subset based on IDF weighting. This suggest that SVD solutions are rather stable for a given rank-*k* solutions for subsets within a large document collection.

Table 6-8 also suggest a positive relationship between the number of retained dimensions and the congruence with expert scores. Figure 6-5 contains more ANOVA results and lists the obtained R^2 value for the measure based on IDF weighting in

combination with high level rank-*k* SVD in the range 500 to 10,000 in steps of 500 for the 5% subset, together with the captured variance[51].

Figure 6-5 : Congruence between (conservative) 5-level scale expert similarity assessment and calculated similarity measures based on IDF weighting for high rank-*k* SVD based on 5% sample (R^2 values of GLM regression based on conservative expert scores of 250 validation cases)



The solid horizontal thick line represents the R^2 value obtained when all dimensions are taken into account - cosine on the full vector space defined by the 5% sample after IDF weighting (left axis, 0.801). The thin solid line represents the R^2 value obtained by a given rank-*k* SVD solution based on IDF weighting (left axis, R^2 of 0.60 for 500 dimensions to 0.806 for 10,000 dimensions). The more dimension/concepts are retained, the more the R^2 values of the SVD-based measure approach the one for the full vector space. The dashed line represents the percentage of variance captured by a given rank-*k,* i.e. the variance present in the approximate document-by-term (document-by-concept) matrix after SVD compared to the total variance in the original

[51] We were only able to go to *k*=10,000 for one variant because of the extreme amount of computing time required.

document-by-term matrix of the 5% sample (right axis, 22% for 500 dimension to 85% for 10,000 dimensions).

This table again confirms the positive relationship between the number of retained dimensions/concepts and the congruence with expert similarity scores, approaching the R^2 of 0.801 of a plain cosines on the full vector space based on IDF weighting of the 5% sample.

Striking is the resemblance between the obtained R^2 value from the original dataset and the 5% sample for the same absolute level of dimensionality reduction. Both for the original large dataset as for the 5% sample, the obtained R^2 value when applying the cosine measure to the full vector space is equal to 0.80. The same for the rank-500 and rank-1,000 SVD solutions: R^2 of 0.60 for rank-500 SVD from the 5% sample compared to 0.57 from the full dataset, and R^2 of 0.66 for rank-1,000 from the 5% sample compared to 0.63 from the full dataset. Again an indication that SVD solutions are rather stable for the same rank-*k* level regardless of the data sample used to derive the SVD solution within a large dataset. Mind that the rank-1,000 SVD solution from the full dataset maps 301,697 (stemmed) terms to 1,000 concepts, a reduction to 0.33% of the original number of dimensions. As the 5% sample subset contained 48,561 (stemmed) terms, a rank-1,000 SVD solution from that sample represents a reduction to 2% of the original dimension (capturing 33% of the original variance). It seems that the absolute number of retained dimensions is more important than the relative number of retained dimensions, i.e. relative to the total number of dimensions in the original vector space.

More important, Figure 6-5 does shed some light on the most pressing question: does a further increase in the number of retained dimension/concepts allow the SVD-based measure to perform better compared to the cosine measure applied on the full vector space, or are obtained similarity scores – and hence performance – merely converging to the ones obtained by the cosine measure on the full vector space. By definition obtained scores of the SVD-based measure will be identical to cosines scores obtained from the full vector space for levels of rank-*k* solutions approaching the original number of dimensions. The point of LSA/SVD is that there are intermediate levels of dimensionality reduction where the SVD-based measure will perform better compared

to the cosine measure obtained from the full vector space. The dotted line in Figure 6-5 represents hypothetical R^2 values in function of the number of retained dimensions/concepts after SVD as claimed by the LSA method: for levels of dimensionality reduction that are too low, performance will be inferior compared to the cosine measure applied on the full vector space (horizontal thick solid line) because too much relevant information is not taken into account. But for a given range of *k* – in this hypothetical example between *k*=1,500 and *k*=3,500 – performance will be superior because noise – biasing full cosine calculations – is removed from the data, to slide down again beneath the level of the full cosine, to eventually approaching again the performance of the full cosine for values of *k* approaching the original number of dimensions. In our real sample, we do not observe this behaviour, i.e. we do not observe ranges of *k* where the SVD-based measure, based on IDF weighting, performs clearly better than the cosine measure calculated on the full vector space after IDF weighting. We only see the SVD-based measure approaching the full cosine measure. However, there might be ranges beyond 10,000 dimensions/concepts where the SVD-based measure performs better, but unfortunately we cannot check that because of computational limitations to derive those SVD solutions. One can observe at the very right of the figure, for *k*-values beyond 8,000, that the R^2 values of the SVD-based measure are slightly higher compared to the R^2 value for the cosine measure on the full vector space (0.806 versus 0.801), however we strongly believe this is due to rounding errors. Given the curve of the performance of the SVD-based measure in function of the number of retained dimensions, the scenario of (very) high *k*-values resulting in the SVD-based measure to perform better than the cosine measure on the full vector space seems very unlikely; our solution with rank 10,000 already captures 85% of the original variance, and the R^2 curve of the SVD-based measure is almost flat in this region of *k*-values. As SVD solutions are dependent of singular values in descending order, one can expect that the dashed curve representing the captured variance in function of the number of retained dimensions will continue to increase at very slow rates beyond k=10,000, and that the obtained R^2 values will follow this pattern, making further significant increases in R^2 values for the SVD-based measures unlikely. Anyhow, if LSA/SVD indeed requires such very high levels of *k* to perform, and if it would be

feasible to derive such SVD solutions for very high levels of *k* – e.g. by using a sample of documents to derive the SVD solution and projecting or folding in all other documents into the newly created truncated vector space – the method would still be virtually impossible to apply for big datasets because of a lack of storage to retain those big full matrices.

To conclude, we doubt whether further increasing the number of retained dimensions/concepts will result in similarity measures that perform better than a cosine measure derived from the full vector space, let alone the practical feasibility of such a solution. It seems that the dimensionality reduction imposed by SVD is not only cutting off noise, but also relevant information, resulting in the observed pattern of the SVD-based measure approaching but never beating the cosine measures based on the full vector space. This brings us back to the question why this SVD approach would work for some datasets but not for ours.

6.9 Conclusions, discussion, limitations, and directions for further research

In this study we thoroughly assessed Latent Semantic Analysis (LSA) as a text mining technique to match patent and publication documents based on their contents. The goal is to find patent and publication documents that are related by the topics they address, the methods they use, the results they obtain and the inventions or discoveries they address. This would bypass limitation of current approaches like IPC-codes, non-patent references, and patent inventor and patentee name matching, and allow to compile large scale datasets for a broad range of applications in innovation studies.

As off-the-shelve text mining solutions are not readily available and experience with patent data is limited, we have set up a large comparison exercise based on the LSA method combining four weighting methods and ten levels of dimensionality reduction, and added three measures based on the number of common terms. Similarity value distributions obtained after application on a large dataset revealed unexpected patterns for LSA-based measures with unrealistic high average similarities and non-biotechnology control set patents being – on average – not less similar to biotechnology publications than biotechnology patents. These results suggest that LSA-based

measures tend to overestimate similarity and not grasp the real topic similarity of patent and publication documents.

Expert validation of 250 cases confirmed the poor performance of LSA based measures. SVD dimensionality reduction results in less congruence with the expert assessment of similarity compared to cosine measures applied on the full vector space, and the less dimensions retained, the less congruence. The term weighting method used also effects the performance; binary and IDF weighting yielded better results compared to TF-IDF weighting and no weighting at all, a remarkable observation as TF-IDF in combination with SVD retaining 300-500 dimensions is a commonly used method. We observed that a cosine metric applied on the fill vector space after binary or IDF weighting yields the best results. However, measures based on the number of common terms between documents perform slightly better. As in the previous study (see previous chapter), the claim that LSA can outperform such simple measures based on common terms or co-occurrence because of a better understanding of the meaning of language of this former method is not backed up by our data.

The weighting method has a significant impact on the performance of the method and it seems that methods taking into account term frequencies perform worse, partly because of stemming and parsing issues, partly because common natural language words tend to get too much weight in the similarity derivation. Better stemming and parsing will probably improve performance.

We propose a combination of measures that allow a more robust identification of similar patent and publication documents: 'common terms MIN', the measure based on the number of common terms weighted for the minimum of the number of terms of the patent and the publication document, as a primary criterion to identify similar documents, combined with 'common terms MAX', the measure based on the number of common terms weighted for the maximum number of terms of the patent and the publication document, as a secondary criterion to eliminate doubt cases due to combinations of short and long documents. Especially when precision is important, those measures deliver good results. When recall gets important, things get more complicated because there are no threshold values that allow a clear cut distinction

between the two groups. The typical trade-off between precision and recall remains a though one, especially as final results are very sensitive to threshold values: small changes in the threshold values for both the primary as secondary criterion result in big differences in the number of matches in the total population. This is particularly problematic for the secondary criterion 'common terms MAX', needed to clear out doubt cases: the vast majority of potential matches based on the primary criterion 'common terms MIN' score very low on 'common terms MAX', so small changes in the range of 'common terms MAX' to discard doubt cases (0.20-0.35) have a huge impact. It seems that our method suffers from too many documents with short abstracts that are very difficult to judge, even for human experts. A potential remedy is to extend document sizes by including patent claims or full documents contents, and not only title and abstract, into the analysis, or use extended abstracts as the ones supplied by the *Derwent World Patent Index*.

When it comes to the identification of patent-publication pairs, i.e. scientific publications from which the contents is covered by patent protection, quality of the results can greatly benefit from an additional third criterion based on the presence or absence of a shared inventor/author. Although inventor-author name matching is not straightforward for larger datasets because of homonymy problems, spelling errors and variation, and use of middle names and initials, the combination of a content based measure like our 'common terms MIN' and 'common terms MAX' and the presence of a shared inventor/author might be the way to go, because the biggest challenge in inventor-author name matching – the homonymy issue – is largely controlled for when combined with a content bases measure.

A remarking observation is the poor performance of SVD-based measures. It is not clear why the specific context of our data does not allow the LSA-method to achieve its full potential. It is unlikely that our dataset size is not large enough, nor that we did not retain enough dimensions/concepts. There are indications that the document size and document size differences are negatively influencing the SVD-based measures. But it might also be due to the particular language use in our patent and publication dataset. Again, using the full text of patent and publication documents, or extended abstracts as

supplied by the *Derwent World Patent Index*, might resolve this, although we lack hard evidence that larger or better abstracts would resolve the issues.

What is clear is that, for our dataset, the dimensionality reduction imposed by LSA/SVD is cutting off valuable information instead of noise. We observe a gradually increasing performance for increasing number of retained dimensions, but we do not observe a range of dimensions for which the performance is better than that of a cosine measure applied on the full vector space; the performance of LSA/SVD is just approaching the performance of a cosine measure on the full vector space for higher numbers of retained dimensions, in contradiction to the claims of LSA that dimensionality reduction would improve results (understanding the 'latent' structure).

Another remarkable observation is that patents of our materials control set are – on average – more related to biotechnology publications than are biotechnology patents. This information can act as a source of inspiration to reveal the shortcomings of SVD on our data. However, looking into many individual cases did not reveal significant information to explain the higher obtained similarity scores nor the poor performance of SVD.

A final reason why LSA might fall short is the limitation to Euclidean geometry as imposed by the assumption of LSA that documents are represented as vectors in a vector space. In an Euclidean space, similarity should be symmetric and not violate triangle inequality - $d(x,z) \leq d(x,y) + d(y,z)$ - placing strong constraints on the location of point in a space given a set of distances (Griffiths, Steyvers & Tenenbaum, 2007). Other text mining techniques not relying on spatial representations, like generative topic models as Probabilistic Latent Semantic Modelling (Hofmann, 1999) and Latent Dirichlet Allocation (Blei, Ng & Jordan, 2003), might be more appropriate to deal with this aspect of the curse of dimensionality.

To conclude, the debate about the value of more complex text mining methods for application on patent and scientific publication data – complex in the sense that they try to deal with the characteristics of text and language – compared to simpler methods based on common terms or co-occurrence does not end here. For our purpose, the

identification of patent-publication pairs, a simple measure based on the number of common terms performs best. While claims of LSA are not backed up by our observations, and simpler seems to be better – in line with Occam's razor principle – other text mining techniques are available and it is worthwhile to investigate the application of those techniques on our data, like the generative topic models mentioned before.

6.10 References

Blei, D. M., **Ng**, A. Y. & **Jordan**, M. I. (2003). "Latent Dirichlet allocation." *Journal of Machine Learning Research*, 3 : 993-1022.

Glenisson, P., **Glänzel**, W., **Janssens**, F. & **De Moor**, B. (2005). "Combining full text and bibliometric information in mapping scientific disciplines." *Information Processing & Management*, 41 : 1548-1572.

Griffiths, T. L., **Steyvers**, M. & **Tenenbaum**, J. B. (2007). "Topics in Semantic Representation." *Psychological Review*, 114 (2) : 211-244.

Hofmann, T. (1999). "Probabilistic latent semantic indexing." *Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval*, 50-57.

Magerman, T., **Van Looy**, B. & **Song**, X. (2010). "Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications." *Scientometrics*, 82 (2) : 289-306.

OECD (2005). A framework for biotechnology statistics. Paris: OECD publishing.

OECD (2009). OECD Biotechnology Statistics. Paris: OECD publishing.

Appendix 6-1 : OECD biotechnology IPC codes (OECD, 2005 and 2009).

IPC codes	Title
A01H 1/00	Processes for modifying genotypes
A01H 4/00	Plant reproduction by tissue culture techniques
A61K 38/00	Medicinal preparations containing peptides
A61K 39/00	Medicinal preparations containing antigens or antibodies
A61K 48/00	Medicinal preparations containing genetic material which is inserted into cells of the living body to treat genetic diseases; Gene therapy
C02F 3/34	Biological treatment of water, waste water, or sewage: characterised by the micro-organisms used
C07G 11/00	Compounds of unknown constitution: antibiotics
C07G 13/00	Compounds of unknown constitution: vitamins
C07G 15/00	Compounds of unknown constitution: hormones
C07K 4/00	Peptides having up to 20 amino acids in an undefined or only partially defined sequence; Derivatives thereof
C07K 14/00	Peptides having more than 20 amino acids; Gastrins; Somatostatins; Melanotropins; Derivatives thereof
C07K 16/00	Immunoglobulins, *e.g.* monoclonal or polyclonal antibodies
C07K 17/00	Carrier-bound or immobilised peptides; Preparation thereof
C07K 19/00	Hybrid peptides
C12M	Apparatus for enzymology or microbiology
C12N	Micro-organisms or enzymes; compositions thereof
C12P	Fermentation or enzyme-using processes to synthesise a desired chemical compound or composition or to separate optical isomers from a racemic mixture
C12Q	Measuring or testing processes involving enzymes or micro-organisms; compositions or test papers therefor; processes of preparing such compositions; condition-responsive control in microbiological or enzymological processes
C12S	Processes using enzymes or micro-organisms to liberate, separate or purify a pre-existing compound or composition processes using enzymes or micro-organisms to treat textiles or to clean solid surfaces of materials
G01N 27/327	Investigating or analysing materials by the use of electric, electro-chemical, or magnetic means: biochemical
G01N 33/53*	Investigating or analysing materials by specific methods not covered by the preceding groups: immunoassay;
G01N 33/54*	Investigating or analysing materials by specific methods not covered by the preceding groups: double or second antibody: with steric inhibition or signal modification: with an insoluble carrier for immobilising immunochemicals: the carrier being organic: synthetic resin: as water suspendable particles: with antigen or antibody attached to the carrier via a bridging agent: Carbohydrates: with antigen or antibody entrapped within the carrier
G01N 33/55*	Investigating or analysing materials by specific methods not covered by the preceding groups: the carrier being inorganic: Glass or silica: Metal or metal coated: the carrier being a biological cell or cell fragment: Red blood cell: Fixed or stabilised red blood cell: using kinetic measurement: using diffusion or migration of antigen or antibody: through a gel
G01N 33/57*	Investigating or analysing materials by specific methods not covered by the preceding groups: for venereal disease: for enzymes or isoenzymes: for cancer: for hepatitis: involving monoclonal antibodies: involving limulus lysate
G01N 33/68	Investigating or analysing materials by specific methods not covered by the preceding groups: involving proteins, peptides or amino acids
G01N 33/74	Investigating or analysing materials by specific methods not covered by the preceding groups: involving hormones
G01N 33/76	Investigating or analysing materials by specific methods not covered by the preceding groups: human chorionic gonadotropin
G01N 33/78	Investigating or analysing materials by specific methods not covered by the preceding groups: thyroid gland hormones
G01N 33/88	Investigating or analysing materials by specific methods not covered by the preceding groups: involving prostaglandins
G01N 33/92	Investigating or analysing materials by specific methods not covered by the preceding groups: involving lipids, *e.g.* cholesterol
* Those IPC codes also include subgroups up to one digit (0 or 1 digit). For example, in addition to the code G01N 33/53, the codes G01N 33/531, GO1N 33/532, etc. are included.	

Appendix 6-2 : Example of a patent-publication combination with high but misleading similarity according to the measure based on TF-IDF and SVD.

Following patent-publication combination is an example of a combination with high similarity scores according to the measures based on TF-IDF and SVD. Similarity scores for these measures range from 0.928 to 0.995 depending of the number of dimensions retained (see last line in the table). Title and abstract of both documents make clear that both documents are only (very) slightly related; both are about milk, but the patent is about an apparatus for milking, while the publication is about a comparison of cow milk and camel milk for characteristics on Lactobacillus acidophilus fermentation.

The measures based on the number of common terms yield low scores (0.10, 0.07 and 0.08 depending whether the minimum number of terms, the maximum number of terms or the average number of terms of both documents is used as weighting factor).

Biotechnology patent title and abstract:

Process and rotary milking parlor for the identification of a milking stall and an animal, in particular a cow, in a rotary milking parlor.

For the determination of the occupancy of a milking stall by an animal, in particular a cow, in a rotary milking parlor with a plurality of milking stalls which are disposed on a rotatable milking platform, a process is proposed in which the identification of the animal only takes place after it has entered the milking stall in which it is supposed to be milked.

Biotechnology publication title and abstract:

Growth-behavior of Lactobacillus-Acidophilus and biochemical characteristics and acceptability of Acidophillus milk made from camel milk.

Acidophilus milk was made from camel milk and compared to that made from cow milk. Although the camel milk supported the growth of L. acidophilus, the quality of acidophilus milk from cow milk was superior. Bovine acidophilus had firm curd while that made from camel milk had flocks with no curd formation. The initial proteolysis of raw camel milk provided ready substrates for L. acidophilus for more protein breakdown in the acidophilus milk made from it.

Similarity values

Weighting method	Dimensions retained (SVD)									
	ALL	**1000**	**500**	**300**	**200**	**100**	**50**	**25**	**10**	**5**
Raw	0.511	0.837	0.873	0.905	0.754	0.391	0.368	0.608	0.691	0.673
Binary	0.083	0.057	0.025	0.023	0.056	0.087	-0.030	0.492	0.763	0.750
IDF	0.095	0.168	0.162	0.260	0.375	0.403	0.504	0.532	0.698	0.738
TFIDF	0.364	0.928	0.973	0.986	0.991	0.991	0.995	0.980	0.959	0.960

Appendix 6-3 : Example of a patent-publication combination of a control set patent and biotechnology publication with high but misleading similarity according to the measure based on the number of common terms weighted by the minimum of the number of terms of both documents ('common terms MIN').

9 common terms out of 11 terms of the patent document and 141 terms of the publication document (after stemming, stop word removal and removal of terms only appearing once in the document set): common terms min = 0.82; common terms max = 0.06.

Control patent title and abstract:

Inbred maize line PHBG4

An inbred maize line, designated PHBG4, the plants and seeds of inbred maize line PHBG4, methods for producing a maize plant produced by crossing the inbred line PHBG4 with itself or with another maize plant, and hybrid maize seeds and plants produced by crossing the inbred line PHBG4 with another maize line or plant.

Biotechnology publication title and abstract:

Major QTLs for disease resistance and other traits identified in recombinant inbred lines from tropical maize hybrids

Major QTLs (quantitative trait loci) with large genetic effects often provide the basis for rapid genetic gains with quantitative traits like disease anti pest tolerance. This study sought to identify major QTLs in maize through the creation and use of recombinant inbred lines (RILs) based uniquely on hybrids of elite tropical and temperate inbreds. Nine single crosses involving ten inbreds served as the source of 1072 RILs created through six cycles of single seed descent in the absence of selection in Hawaii. About 30 sublines of each of the ten parental inbreds were bred to estimate means and variances of quantitative traits under study. These parameters were then used to predict RIL segregations of major QTLs based on normal probability distributions, designated here the RIL-NP method. Segregations were also tested for fit to expected ratios by the use of maximum likelihood estimators. The nine sets of RILs were grown selectively under disease epiphytotics at experimental stations in the United States, Korea, Mexico, Nigeria, and the Philippines. Major QTLs apparently acting monogenically (segregating 1:1 in RILs) were identified to control general resistance to the following diseases: Southern rust: Common rust, Northern le:lf blight, Southern leaf blight, Bacterial leaf blight, Stewart's bacterial wilt, Maize mosaic virus and Maize streak virus. Digenic segregations with additive gene action appeared to characterize QTLs governing resistance to Striga witchweed and to European corn borer. Major QTLs were also observed for polymorphisms in ear height, plant height, maturity, tassel branch number and central tassel-spike length. Examples are cited of molecular mapping based on these RILs. The potential use of major QTLs in marker-assisted selection is discussed in relation to the transfer to temper;Ice germplasm of tolerances to disease, insect and stress from the largely untapped tropical germplasm.

Appendix 6-4 : Example of a patent-publication combination with stemming error with high impact on weighting methods including term frequencies.

Both documents have nothing in common, yet score significantly higher for measures based on weighting methods including term frequencies). Both documents have only two (stemmed) terms in common (after stemming and stop word removal), 'feed' and 'ga'. But the stemmed term 'ga' occurs 9 times in the patent document and 29 times in the publication document, resulting in high weights when the term frequency is included. But the stemmed term 'ga' in the patent document is a stemming error derived from 'gas', while the stemmed term 'ga' in the publication document is an abbreviation of 'gibberellin' and has nothing to do with the stemmed term 'ga' in the patent document. For weighting methods not taking term frequency into account, this stemming error counts as just matching term, but for weighting methods using term frequency, this stemming error is magnified and leads to erroneous results.

Biotechnology patent title and abstract:

Incubator with external gas feed.

An incubator with an external gas feed is disclosed, wherein a gas is supplied to an interior space of the incubator to maintain an interior atmosphere with a constant gas-to-air ratio. The gas is supplied to the interior space through a gas nozzle forming a gas jet. The gas jet draws in the interior atmosphere through an injector effect, thereby thoroughly mixing the gas with the interior atmosphere.

Biotechnology publication title and abstract:

Gibberellin metabolism in suspension-cultured cells of raphanus-sativus.

Gibberellin A(1) (GA(1)), GA(4), GA(9), GA(19) and GA(20), which are known to be native to Japanese radish (Raphanus sativus), were applied as [H-3]GAs and [H-2]GAs to cell suspension cultures of R. sativus. As the metabolites in [H-2]GA-feeds, [H-2]GA(8) from [H-2]GA(1), [H-2]GA(1) and [H-2]GA(2) from [H-2]GA(4), [H-2]GA(1), [H-2]GA(4) and [H-2]GA(20) from [H-2]GA(9), [H-2]GA(20) from [H-2]GA(19), and [H-2]GA(1) and [H-2]GA(20)-15-ene from [H-2]GA(20) were identified by GC-SIM. The distribution of [H-3]GA metabolites after HPLC corresponded closely with that of the [H-2]GA metabolites, except in the case of the [H-2]GA(20)-feeds. Based on the metabolic patterns of applied GAs, it is supposed that 13-hydroxylation from GA(4) is much more dominant than 3 beta-hydroxylation from GA(20) in pathways leading to GA(1) in suspension cultured cells of R. sativus.

Similarity values

Weighting method	Dimensions retained (SVD)									
	ALL	**1000**	**500**	**300**	**200**	**100**	**50**	**25**	**10**	**5**
Raw	0.688	0.881	0.960	0.958	0.638	0.294	0.229	0.362	0.361	0.443
Binary	0.072	0.088	0.017	0.052	0.056	0.066	0.066	0.144	0.307	0.525
IDF	0.056	0.128	0.128	0.171	0.222	0.218	0.220	0.270	0.471	0.689
TFIDF	0.594	0.941	0.972	0.984	0.988	0.936	0.847	0.886	0.928	0.961

Appendix 6-5 : Example of a patent-publication combination with tokenization and parsing issues with high impact on weighting methods including term frequencies.

Both documents are not related and have only two terms in common: 'alpha' and 'beta' (after stemming and stop word removal). Both of these terms occur a lot in both documents as part of chemical formulas, and these high term frequencies result in higher similarity values for weighting methods based on term frequencies. But the larger chemical formulas these terms are part of, are not related. If would probably be better to parse and index those formulas as one piece, but this is not straightforward.

Biotechnology patent title and abstract:

alpha -mannosidase inhibitors

¢4S-(4 alpha ,4a beta ,5 beta ,6 alpha ,7 alpha ,7a alpha)!-Octahydro-1H-1-pyrindine-4,5,6,7-tetrols and ¢4R-(4 alpha ,4a alpha ,5 alpha ,6 beta ,7 beta ,7a beta)!-octahydro-1H-1-pyrindine-4,5,6,7-tetrols are useful as inhibitors of alpha-mannosidase and are useful immunostimulants, chemoprotective and radioprotective agents and antimetastatic agents.

Biotechnology publication title and abstract:

Taxanes from Taxus mairei

Four new taxane diterpenes, 9 alpha-hydroxy-14 beta-(2-methylbutyryl)oxy-2 alpha, 5 alpha, 10 beta-triacetoxytaxa-4(20), 11-diene, 2 alpha, 5 alpha, 9 alpha, 10 beta, 14 beta-pentaacetoxytaxa-4(20),11-diene, 5 alpha-(cinnamoyl)oxy-7 beta-hydroxy-9 alpha, 10 beta-13 alpha-triacetoxytaxa-4(20), 11-diene and 5 alpha-hydroxy-9 alpha, 10 beta, 13 alpha-triacetoxytaxa-4(20), 11-diene, along with 12 known taxa-4(20),11-dienes, have been isolated from twigs of Taxus mairei and their structures determined by spectral methods. Copyright (C) 1996 Elsevier Science Ltd.

Similarity values

Weighting method	Dimensions retained (SVD)									
	ALL	**1000**	**500**	**300**	**200**	**100**	**50**	**25**	**10**	**5**
Raw	0.705	0.928	0.930	0.946	0.952	0.970	0.985	0.993	0.999	0.997
Binary	0.089	0.248	0.259	0.295	0.328	0.349	0.198	0.159	0.420	0.785
IDF	0.011	0.248	0.298	0.353	0.389	0.545	0.507	0.397	0.412	0.517
TFIDF	0.199	0.935	0.970	0.982	0.988	0.994	0.996	0.998	0.999	0.997

EMPIRICAL PART II :
POTENTIAL PITFALLS – IN SEARCH OF ANTI-COMMONS EVIDENCE

7 In search of anti-commons evidence: patent-publication pairs in biotechnology. An analysis of citation flows.

There are three reasons why lawyers are being used more and more in scientific experiments.
First, every year there are more of them around.
Second, lab assistants don't get attached to them.
And, third, there are some things that rats just won't do.
Anonymous

7.1 Introduction

In this chapter we want to close the circle by looking at one potential pitfall of academic patenting and increasing commercialization of science in general: the potential introduction of an anti-commons effect. We started our journey by looking at the relation between the science-intensity of patents and technological performance of countries as we observe increasing science-intensity of patents throughout the last years. One aspect of the 'scientification' of patents is the increasing trend of academic patenting, which is to be encouraged to the extent it has a positive impact on technological performance (see chapter 3). However, concerns arise about the privatization of science and the creation of an anti-commons effect (see introduction chapter). In this chapter we want to contribute to the research on an anti-commons effect in biotechnology by comparing citation patterns of patents and scientific publications for a large dataset containing all EPO and USPTO biotechnology patents from the *PATSTAT* database (EPO Worldwide Patent Statistical Database) and scientific publications published in journals covered by the *WOS* database (Thomson Reuters ISI Web of Science) from 1991 to 2008. First we investigate whether biotechnology publications for which a counterpart exists in the patent system (so called 'patent-publication pairs' or 'patent-paper pairs', scientific publications from which the contents

- methodology, findings, discovery - is part of a patent application) are cited differently (more/less) within scientific journals, compared to similar biotechnology publications without a patent counterpart. If an anti-commons effect is present, we expect to observe less forward citations for the publications that are part of a patent-publication pair (related to a patent) as scholars would refrain from building upon those publications because of uncertainty imposed by IPR claims. Next, we engage in a similar analysis, focusing this time on 'technological' citations: to what extent are patents that are part of a patent-publication pair (related to a scientific publication) cited differently by other patents compared to biotechnology patents without a counterpart in the scientific literature.

The former will allow us to shed some light on the fear that exploitation of scientific findings is hampering scientific development by pruning promising developments due to the introduction of (potentially blocking) patents. The latter will allow us to look at the technological impact of scientific developments that become translated into a patent.

An important methodological aspect for this kind of studies relates to the identification of those patent-publication pairs, scientific publications for which a patent equivalent is present. To obtain a broad set of patent-publication pairs, we stepped down from a manually guided process of mapping patent and scientific publications and developed a new approach of automated, large scale, mapping of patents and scientific publications based on content similarity by relying on text mining algorithms, as developed in the previous methodological chapters. This approach allows large scale processing of patents and scientific publications to detect patent-publication pairs.

Within the next pages, we first outline the selection of the data used for this analysis, followed by a description of the methodology adopted to assess the similarity between patents and scientific publications. This section is followed by reporting the findings, for scientific citations and patent citations respectively. We conclude with outlining the limitations of our work and suggest avenues for further research in this area.

7.2 Data and methodology

Field selection

As throughout the whole dissertation we focus again on patents and scientific publications in the field of biotechnology because it is a field well known for the presence of science-technology linkages and because the large scale exploitation of biomedical research makes it more susceptible to an anti-commons effect (Heller & Eisenberg, 1998).

Patents and publications were selected based on technological and scientific classification schemes respectively. Patent-publication pairs were identified by matching the content of titles and abstracts of patents and scientific publications using text mining algorithms.

Selection of biotechnology patents and publications

For this study we use the same dataset as compiled in the previous chapter to develop our method to map patents and publications (OECD definition of biotechnology[52] for patents from the *PATSTAT* database - EPO and USPTO - and 10 relevant subject categories plus three major multidisciplinary journals for publications in the *WOS* database). This wraps up to 88,248 granted EPO and USPTO biotechnology patents and 948,432 biotechnology publications from the period 1991-2008.

Text mining oriented identification of patent-publication pairs

The identification of patent-publication pairs (scientific publications from which the contents - methodology, findings, discovery - is part of a patent application) is based on the content similarity of titles and abstracts of patents and publications, as developed in the previous chapter. For all patents, the similarity with all publications is derived based on content similarity metrics. Patent-publication combinations with similarity scores beyond thresholds are retained as patent-publication pairs under the condition that at least one of the patent inventors is listed as publication author.

Based on the insights from our large comparative study as outlined in the previous chapter, two metrics are combined for the classification of patent-publication

[52] OECD, 2005 and 2009. See also previous chapter.

combinations. The number of common terms, divided by the minimum of the number of terms of the patent document on the one hand and of the publication document on the other hand, is used for a first selection of patent-publication combinations with significant content similarity (CommonTermsMin ≥ 0.60). A second criterion, based on the number of common terms divided by the maximum of the number of terms of the patent document and publication document, is used to filter out ambiguous cases (CommonTermsMax ≥ 0.30). These two content-based criteria are combined with an additional criterion based on authorship: at least one of the patent inventors has to be listed as a publication author. Together those three criteria allow an accurate identification of patent-publication pairs; threshold values on the measures were set in a conservative way to eliminate false positives at the cost of a lower coverage (high precision but lower recall).

Identified patent-publication pairs

The starting point for the identification of patent-publication pairs is the combined dataset of 88,248 biotech patents and 948,432 biotech publications from 1991 to 2008[53]. Application of the first matching criterion, a content similarity of at least 0.60 based on the number of common terms weighted for the minimum of the number of terms of both documents, yields 27,250 related patent-publication combinations out of the more than 80 billion combinations under examination. Application of the second matching criterion, a content similarity of at least 0.30 based on the number of common terms weighted for the maximum of the number of terms of both documents, results in 645 patent-publication pairs. Application of the last criterion, at least one patent inventor being listed as a publication author, results in a final set of 584 patent-publication pairs. 17 patents are matched with multiple publications (up to three publications), which seems to be cases of (partly) disclosure of the same results in multiple scientific articles. At the same time, 115 publications are matched to multiple patents (up to seven patents), which revealed to be members of the same patent family. Hence we have 566 distinct biotechnology patents having a paired

[53] Only patents and publications with titles and abstracts of sufficient length are retained to allow for content-based matching.

biotechnology publication, and 400 distinct biotechnology publications having a paired biotechnology patent.

Remember that we deliberately opted for a very conservative selection to identify patent-publication pairs. Especially the second criterion filters out a lot of ambiguous cases, so we can be confident that the described patent-publication matching method reveals real patent-publication combinations.

7.3 Findings on citation patterns of scientific publications (publication-to-publication citations)

Within this section we report and discuss the empirical results obtained when analysing scientific citations - i.e. citations from other scientific publications - to scientific publications that are part of a patent-publication pair. This analysis implies a comparison with scientific citations to scientific publications which do not belong to a patent-publication pair (that do not have a patent counterpart).

Descriptive statistics

Table 7-1 shows the evolution of the number of biotechnology publications and the number of forward citations from other scientific publications. The left part of the table contains data for all biotechnology publications; the right part of the table contains data for those biotechnology publications that are paired with a patent, i.e. publications that are part of a patent applications (patent-publication pairs).

The number of biotechnology publications in our dataset is steadily growing from 31,381 in 1991 to 76,004 publications in 2008. After a first period characterized by double-digit growth figures (from 1991 to 1995 - 10 to 12 per cent annual growth in publication outcome), we observe a period of moderate growth (4.3 to 8.0 per cent between 1996 and 1999) followed by a period of volatility during the most recent years (-2.5% to 2% with some upward outliers in 2003, 2005 and 2008).

The average number of forward publication citations (publication-to-publication citations counted by a 10-year citation window: year of publication plus following nine

years)[54] for the biotechnology publications follows a more or less stable pattern; within a first time period observed citation rates vary between 45 and 50 (till 2000) followed by a decrease from 2000 onwards, reflecting the shorter time window of observation (citation counts based on a 10-year citation window cannot be complete for the last 9 years). The average number of forward citations for all publications between 1991 and 2000 is 46.9 (median number of forward citations is 20).

Table 7-1 : Number of biotechnology publications and forward citations per year

	ALL BIOTECHNOLOGY PUBLICATIONS		PAIRED BIOTECHNOLOGY PUBLICATIONS	
Publication year	**Number of publications**	**Average number of forward citations**	**Number of publications**	**Average number of forward citations**
1991	31,381	50.53	16	229.31
1992	35,185	49.29	22	77.82
1993	38,677	49.46	25	150.44
1994	42,764	47.11	40	133.05
1995	48,092	45.97	43	140.23
1996	50,788	44.43	35	224.94
1997	53,175	45.91	40	233.88
1998	57,361	45.99	41	204.83
1999	59,866	45.76	40	78.23
2000	61,072	47.12	26	104.88
2001	62,299	43.29	29	286.48
2002	63,409	38.57	13	69.00
2003	66,564	33.51	14	62.79
2004	65,705	28.47	8	28.38
2005	72,378	22.24	4	43.25
2006	70,529	15.99	2	63.50
2007	69,756	10.58	1	2.00
2008	76,004	4.97	1	16.00
Total/Average	1,025,005	34.64	400	156.51

For the period 1991-2000 (the period relevant for comparison because of the 10-year citation window) we have 328 publications that are part of a patent-publication pair, starting from 16 in 1991 and rapidly growing to 40 in 1994, to smooth out around 40 between 1994 and 1999, and ending with a decrease to 26 in 2000. For those publications, the average number of forward citations is far more volatile throughout

[54] For all forward publication citation counts in this chapter we use citation counts based on a 10-year citation window except when explicitly mentioned otherwise.

the years, ranging from a minimum of 77.8 forward citations on average in 1992 to a maximum of 233.9 citations on average in 1997, with no clear trend. The average number of forward citations for all publications that are part of a patent-publication pair for the total period of 1991-2000 is 161.8 (median number of forward citations is equal to 65).

On average we clearly observe substantially higher forward citation counts for publications that are part of a patent-publication pair and other publications (mean of 161.8 versus 46.9, median of 65 versus 20 for the time period 1991-2000). But not only the average numbers are higher, the complete distribution of forward citation counts is shifted to the right in favour of publications that are part of a patent-publication pair.

Figure 7-1 : Distribution of the number of forward publication citations for all biotechnology publications and biotechnology publications part of a patent-publication pair (1991-2000)

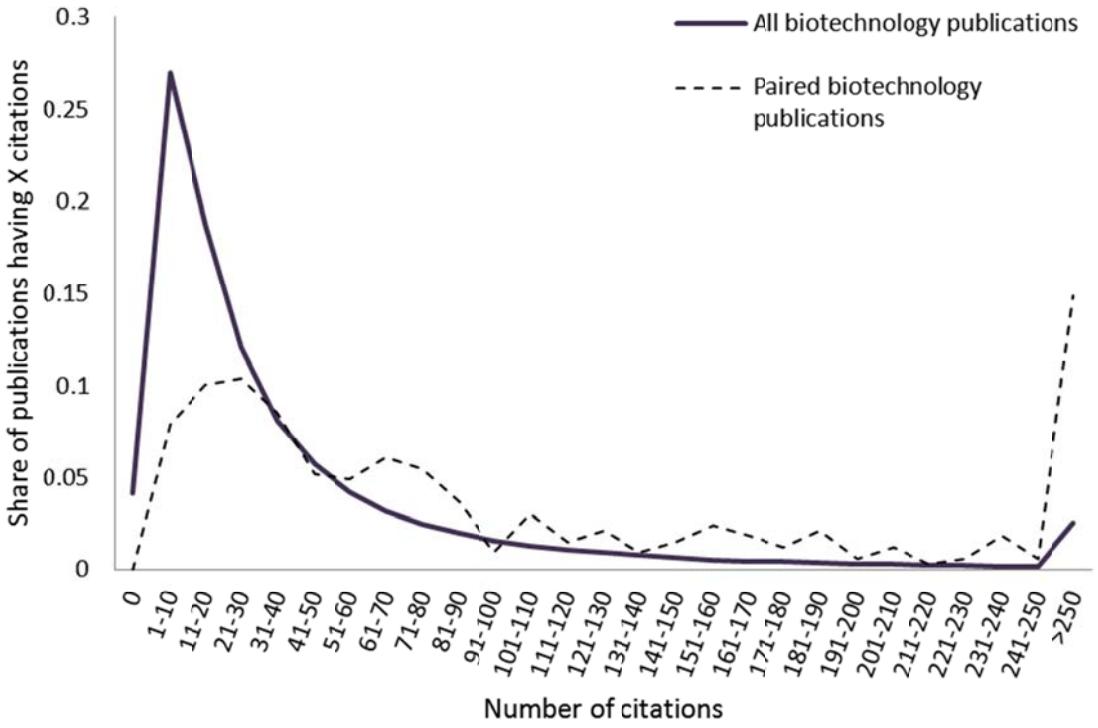


Figure 7-1 shows the distribution of the number of forward publication citations for all biotechnology publications and biotechnology publication part of a patent-publication pair for the period 1991-2000. 25% of paired biotechnology publications have 27 or less citations compared to 7 or less citations for the first quartile for all biotechnology

publications; 50% of paired biotechnology publications have 65 citations or less (20 citations for all biotechnology publications) and 75% of paired citations have 160 or less citations (48 citations for all biotechnology publications). At the right side of the distribution we observe substantial outliers, especially for publications that are related to a patent.

One potential explanation for the higher number of forward citations for paired publications might be the difference in the number of authors. Publications having more authors tend to have more forward citations - as is confirmed by our data (an average of 38 forward citations for single authored papers up to 46 citations for publications with 5 authors and 86 citations for publications with 10 authors)[55]. We indeed observe a higher number of authors for paired publications (26% more authors on average), but this seems not to be a satisfactory explanation for the differences in citation behaviour; for publications with the same number of authors, the average number of forward citations is again substantially higher for paired publications, with a notable exception for single authored publications (an average of 19 citations for single authored paired publications up to 135 citations for paired publications with 5 authors and 345 citations for paired publications with 10 authors)[56].

Another, more important, consideration when observing the difference in forward citation counts is the presence of a selection bias for paired publications towards higher quality publications. For the large overall biotechnology publication sample, all kind of quality levels will be present in the dataset. But for publications that are part of a patent-publication pair, one can expect to find more publications of higher quality than average, i.e. publications valuable enough to justify costs and effort to apply for a patent. We need to correct for quality differences to obtain a fair comparison between paired (publications with a patent counterpart) and non-paired (publications without a patent counterpart) publications. We to take into account the journal in which

[55] 78% of the biotechnology publications in our sample have 5 or less authors, 20% have 6 to 10 authors.
[56] When comparing the number of forward citations for groups of publications with a given number of authors with a bin size of 5, paired publications always have a substantial higher number of forward citations. For publications having 1, 2, … 10 authors (the vast majority of publications), paired publications always have higher citation counts for all levels of the number of authors, except for single authored publications.

publications are published as an indication of the quality level of publications (i.e. we assume underlying journal impact factors are a good indication of the average quality of publications appearing in that journal).

Table 7-2 : Top publishing and top cited journals for all biotechnology publications and for biotechnology publications with a paired patent (1991-2000)

	ALL BIOTECHNOLOGY PUBLICATIONS				BIOTECHNOLOGY PUBLICATIONS WITH PAIRED PATENT			
		Journal	Average citations	Share of all biotech publications	Journal	Average citations	Share of all paired biotech publications
Top publishing journals	1	JOURNAL OF BIOLOGICAL CHEMISTRY	67.14	4.96%	PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA	108.99	22.26%
	2	BIOCHEMISTRY	45.03	1.67%	SCIENCE	550.68	8.54%
	3	JOURNAL OF BACTERIOLOGY	34.46	1.66%	CELL	366.36	6.71%
	4	APPLIED AND ENVIRONMENTAL MICROBIOLOGY	33.01	1.59%	JOURNAL OF BIOLOGICAL CHEMISTRY	98.67	6.40%
	5	BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS	32.74	1.31%	NUCLEIC ACIDS RESEARCH	57.9	3.05%
Top cited journals	1	NATURE REVIEWS MOLECULAR CELL BIOLOGY	400.56	0.00%	NATURE	803.13	2.44%
	2	ANNUAL REVIEW OF BIOCHEMISTRY	374.20	0.06%	MOLECULAR CELL	617.33	0.91%
	3	ANNUAL REVIEW OF CELL BIOLOGY	305.20	0.01%	SCIENCE	550.68	8.54%
	4	CELL	296.03	0.78%	CELL	366.36	6.71%
	5	ANNUAL REVIEW OF CELL AND DEVELOPMENTAL BIOLOGY	280.95	0.02%	GENES & DEVELOPMENT	307.00	1.52%

Table 7-2 contains the most important journals for biotechnology publications for the period 1991-2000. The top of the table contains the most important journals in terms of the number of biotechnology publications – expressed in share of all biotech publications – while the bottom of the table contains the most important journals measured by the average number of forward citations for the biotechnology

publications[57]. The left side of the table contains the most important journals for all biotechnology publications in our sample and the right side contains the most important journals for biotechnology publications that are part of a patent-publication pair (publications with a patent counterpart). For every journal the average number of forward citations (for the biotechnology publications in our sample) and the share of biotechnology publications within our sample are listed[58].

Paired sample T-tests

To test whether forward citations to paired publications - publications that are part of a patent-publication pair, i.e. have a scientific counterpart - differ from publications that are not related to patents, controlling for quality differences, we perform a series of paired sample T-tests. Every time we split up our set of biotechnology publications into paired publications (with patent counterpart) and non-paired publications (without patent counterpart) and control for quality differences by grouping forward citation counts by journal and publication year. For every journal and publication year containing at least one paired publication, we compare the average number of forward citations of paired biotechnology publications with the average number of forward citations of non-paired publications for that same year and journal. This allows for a comparison taking into account quality differences as expressed by the differences in the journal quality. Multiple T-tests are performed based on different comparison of paired and non-paired publications to test the robustness of the findings. Table 7-3 contains the results of the paired sample T-tests.

[57] The three multidisciplinary journals that were added to our selection of biotechnology patents also represent a large share of all in our dataset (PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA: 5.3%; NATURE, 3.1% and SCIENCE: 2.8%) but this is misleading as all publication of those journals were included in our dataset – and not only the biotechnology publications - as there is no straightforward way to identify biotechnology publications within these journals.

[58] For the right side of the table, the share of paired biotechnology publications is listed.

Table 7-3 : Results of paired sample T-tests for paired and non-paired publications (1991-2000)

Test		N	Mean 1	Mean 2	Difference	t value	Pr > |t|
Paired vs **non-paired**	Forward citations	190	130.47	74.24	56.23	4.33	0.0001
	Without self citations	190	116.01	65.02	50.99	4.07	0.0001
Paired vs **non-paired** (at least 2 paired publications)	Forward citations	59	224.97	131.63	93.34	3.12	0.0028
	Without self citations	59	202.7	117.88	84.82	2.97	0.0043
Paired and grey zone vs **all others**	Forward citations	764	60.57	42.69	17.88	5.72	0.0001
	Without self citations	764	53.09	36.48	16.61	5.59	0.0001
Paired and grey zone vs **all others** (at least 2 paired or grey zone publications)	Forward citations	281	96.41	59.64	36.77	5.57	0.0001
	Without self citations	281	85.85	51.76	34.09	5.43	0.0001
Academic, government/non-profit, hospital patentee vs **affiliation**	Forward citations	24	122.94	104.83	18.11	0.88	0.3899
	Without self citations	24	106.81	91.89	14.92	0.76	0.4521

In total 5 tests are performed. In the first test, we take all 328 distinct biotechnology publications from the period 1991-2000 (to allow a full 10-year citation window) that are paired to a biotechnology patent and compare the average number of forward citations (130.47) with the average number of forward citations of all 106,027 non-paired biotechnology publications from the same period 1991-2000 (average of 74.24)[59]. We group the average number of forward citations by journal and publication year, and compare the 190 resulting journal/year group averages using a paired sample T-test and find that the difference is significant. We do the same correcting the number of forward citations by the number of self citations to get the net number of citations without self citations and again find a significant difference (116.01 versus 65.02)

[59] Only biotechnology publication suited for text mining - i.e. with a title and abstract of substantial length - are included in the comparison. This additional filter is applied on all samples of all performed T-tests.

To allow more variation for the paired publications, we performed a second test based on the data of the first test but only retaining journal/year combinations having at least two paired publications, reducing the number of journal/year combinations to 59 with 197 paired publications versus 60,848 non-paired publications. Again we find a significant difference in the number of forward citations, both for all citations and for net citations (corrected for self citations).

The problem with the samples used in the first two T-tests is that the number of paired publications is very low compared to the total number of biotechnology publications because of the very conservative selection of patent-publication pairs. As a robustness check, we relax the patent-publication pair selection criteria and include additional patent-publication pairs. In practice, we add a grey zone of patent-publication combinations scoring 0.50 or above on the first content similarity based criterion (number of common terms weighted for the minimum number of terms of both documents) and scoring 0.25 or above on the second content similarity based criterion (number of common terms weighted for the maximum number of terms of both documents) - compared to respectively 0.60 and 0.30 for the original selection of patent-publication pairs. This results in 3,432 additional combinations with 1,979 distinct patents and 1,939 distinct publications. The third T-test uses this sample to compare forward citations between the group of paired publications, including the additional paired publications retrieved by the relaxed thresholds, with all other biotechnology publications. Again citation averages by journal/year combinations are compared (764 journal/year combinations; 1,681 paired publications versus 197,556 non-paired publications). The difference in the average number of citations is again significant, both for all citations and for net citations (corrected for self citations).

Again, to allow more variation for the paired publications, we perform a forth test based on the data of the third test but only retaining journal/year combinations having at least two (extended) paired publications, reducing the number of journal/year combinations to 281 with 1,198 paired publications versus 135,409 non-paired publications. Again we find a significant difference in the number of forward citations, both for all citations and net citations (corrected for self citations).

The T-last test is a bit different and looks for differences in the nature of paired publications. We observe that the majority of our paired biotechnology publications are linked to a patent with at least one patentee from the academic or government/non-profit or hospital sector (253 out of 328 paired publications between 1991 and 2000). 71 paired biotechnology publications are linked to a patent with at least one company patentee, of which 10 are copatented with an organization of the academic, government/non-profit or hospital sector. This does not necessarily mean that companies are less involved in both publishing and patenting the same invention/discovery in se, it is just a reflection of the fact that companies are less involved in publishing in journals covered by the *WOS* database.

However, it is a well-known phenomenon that hospital, government/non-profit and especially academic institutions do not always act as a patentee even when the invention is related to activities within these institutions. So we know that the number of academic, government/non-profit and hospital patents is underestimated when only looking at the patentee sector. This can be resolved by identifying patent inventors residing in academic, government/non-profit or hospital institutions not acting as patentee. We used an indirect way to identify those inventors for those publications paired with a patent. We look at the affiliations of the related publication, and if at least one of the affiliations is an academic, government/non-profit or hospital organization, we label the related patent as having an inventor from the academic, government/non-profit or hospital sector, based on the fact that, because of our selection process, every patent-publication pair has a shared inventor/author. This construct is not a perfect indicator: a publication might be linked to more than one affiliation, but as our *WOS* dataset does not allow to identify which author belongs to which affiliation, we have no clue whether the shared inventor/author is really the one belonging to the academic, government/non-profit or hospital affiliation.

Using our identification method, we found 61 paired publications with an academic, government/non-profit or hospital affiliation where this affiliation is not acting as patentee on the related patent. In total this makes that 314 out of 328 paired

publications between 1991 and 2000 are linked to an organization from the academic, government/non-profit or hospital sector.

The last T-test tests whether there are differences in the average number of forward citations between paired publications linked to a patent with a patentee from the academic, government/non-profit or hospital sector and paired publications without such patentee but with an affiliation to an academic, government/non-profit or hospital organization. Because of the limited number of observations, citations averages are not compared on the level of journal and publication year combinations, but at the level of journals. There are 24 journals having paired publications from both groups: 180 paired publications linked to a patent with a patentee from the academic, government/non-profit or hospital sector, and 56 paired publications linked to a patent without such patentee but with an affiliation to one of those sectors. Now we find that there is no significant difference in average number of citations between those two groups, nor for the total number of citations, nor for the net number of citations corrected for self citations ($Pr>|t| = 0.39$ and 0.45 respectively). The idea to conduct this test was triggered by findings of Czarnitski, Glänzel & Hussinger (2009) in the debate of the quantitative and qualitative effects of academic patenting. They take into account the heterogeneity of patenting activities and find that academic patenting with non-profit organizations increases publication quantity and quality and are complementary to academic activities, while academic patenting with corporations have a negative impact. Our last T-test makes the same distinction: it compares average forward citation counts between paired publications linked to a patent with a non-profit patentee and paired publications linked to a patent with no non-profit patentee (hence with a company as patentee) but with an affiliation to an academic, government/non-profit or hospital institution (hence indication of an academic or non-profit patent despite the absence of an academic or non-profit patentee). As stated before, we do not find a significant difference in the number of forward citations, albeit that our sample is very small.

Results of the T-tests reveal that paired publications have significantly more forward citations on average, and that there is no indication that paired publications linked with

a non-profit patentee have a significantly different citation counts compared to paired publications with a non-profit inventor (but commercial patentee).

However, this increase in the forward citations for paired publications is not present for all journals. For 49 journals we find an increase, for 31 journals we find a decrease. Table 7-4 contains the journals with the highest increase and decrease – in absolute terms - in the number of forward citations for paired publications.

Table 7-4 : Journals with highest increase and decrease of average forward citations for paired publications

Journals with highest increase - in relative terms - in the number for forward citations for paired publications		
Journal	**Absolute increase**	**Relative increase**
Nature	+ 604	× 4.0
Molecular Cell	+ 457	× 3.8
Journal of Immunological Methods	+ 56	× 3.3
Current Microbiology	+ 19	× 3.0
Current opinion in chemical biology	+ 125	× 2.9
Journals with highest decrease – in relative terms – in the number of forward citations for paired publications		
Journal	**Absolute decrease**	**Relative decrease**
Bioinformatics	- 62	÷ 8.8
Canadian Journal of Botany	- 11	÷ 6.7
Journal of Applied Microbiology	- 12	÷ 3.0
Biochimica et biophysica acta-molecular …	- 34	÷ 2.8
Biotechnology Techniques	- 4	÷ 2.8

Multivariate analysis

We also performed a multivariate analysis to verify the significance of the observed difference when controlling for other factors. Given the nature of the data (citation data) we opted for a negative binomial regression with the number of forward citations as dependent variable and a dummy variable indicating whether a publication is part of a patent-publication pair or not (i.e has a patent counterpart or not) as independent variable.

To adjust for the expected difference in average quality between paired and non-paired publications (due to the potential selection bias of publications that are part of a patent-publication pair), we only include publications from journals that have at least one publication that could be paired with a patent, i.e., we only use publications that are comparable in average impact factor because they originate from the same set of

journals. This leaves 400 biotechnology publications that are part of a patent-publication pair, and 451,803 biotechnology publications that are not part of a patent-publication pair for the whole period 1991-2008[60].

For this analysis, we use net citation counts, i.e. citations counts corrected for self citations, as dependent variable. We further control for journal of publications (105 distinct journals), publication document type (article, letter, note, review), number of backward publication-to-publication citations, and finally, the number of authors. We also include a time variable (1 for the first year, 1991, up to 18 for the last year, 2008) and a squared time variable to accommodate evolutions over time.

Table 7-5 : Results of negative binomial regression for 1991-2008 (dependent variable: number of net forward publication citations – i.e. without self citations – of publications)

			95% Wald Confidence Interval		Hypothesis Test		
Parameter	**B**	**Std. Error**	**Lower**	**Upper**	**Wald Chi-Square**	**df**	**Sig.**
(Intercept)	2.966	.1258	2.719	3.213	555.643	1	.000
Pair (Y/N)	.450	.0506	.350	.549	78.945	1	.000
Document type:							
Article	-.574	.0113	-.596	-.552	2589.688	1	.000
Letter	-.774	.0590	-.890	-.659	172.469	1	.000
Note	-.567	.0175	-.601	-.533	1051.989	1	.000
Review	0	.	.	.	.	.	.
Number of backward publication citations	.013	.0001	.013	.014	10416.453	1	.000
Number of authors	.033	.0005	.032	.034	4613.407	1	.000
Time	.125	.0015	.122	.128	7191.199	1	.000
Time²	-.012	.0001	-.013	-.012	29450.994	1	.000
Journal dummies (n=104)			Included				

Table 7-5 reports the results of the regression analysis of forward publication citations of publications. Publications being part of patent-publication pairs have significantly more forward publication citations (Pair Y/N). One also notices a positive relationship between forward citations and the number of authors as well as the number of backward citations. Citation rates differ between document types: reviews receive more citations compared to articles, letters and notes. The number of forward citations differ

[60] Again only publications with titles and abstracts of substantial length were included in the analysis.

significantly between journals (journal dummies have been included, but not reported, n=104). Finally, the observed citation rates reflect an inverse U pattern over time.

When removing outliers, i.e. all publications with a forward citation count larger than the mean plus three times the standard deviation, similar results are obtained then the ones reported in Table 7-5.

Comparison of citation counts before and after patent grant

Inspired by the observation of Murray & Stern (2007) - a relative decline in citation patterns after patents have been granted – we also verify whether the citation rates differ before and after a patent has been published or granted. We follow the reasoning of Murray and Stern, stating that if a patent grant comes to a complete surprise to follow-on researchers, i.e. if researchers that continue working on previous discoveries are not aware of pending patent applications on those previous discoveries, a drop in citation rate can be an indication of the presence of an anti-commons effect. The reasoning behind this construct is that if researchers are not aware whether a given piece of knowledge is subject to patent filing, they will use (cite) this knowledge (publication) in a normal way. As soon as a patent covering that piece of knowledge is granted, those follow-on researchers might stop using (citing) this knowledge because of the perceived 'price' (patent rights) of building on the prior discovery. Hence in case of the presence of an anti-commons effect, forward citations of publications that are part of a patent-publication pair are expected to drop as soon as the corresponding patent is granted.

To test this we split up forward citation counts for all paired publications into the number of citations received before and after the grant year of their corresponding patent[61] [62] [63]. These numbers are aggregated at the level of journals and publication years, resulting in two average citations counts; one for the pre-grant period, and one for the post-grant period. Next, for every observed journal and publication year having

[61] For those publications linked to multiple patents (multiple members of patent families), the earliest patent grant data was used to split citations into a pre-grant and post-grant period.
[62] For this analysis, the total number of citations was used, including self citations.
[63] Only publications of period 1991-2000 are included to have a full 10-year citation window for all publications and to make use of the fact that USPTO applications were not made public before 2001, making the changes of a 'surprise' grant to follow-up researchers more likely.

paired publications, we construct a control group that consists of all non-paired biotechnology publications published in that given journal and year. For these publications, forward citations are split up in exactly the same manner as to reflect the pre- and post-grant period. This is done as follows: if for a given journal and publication year only one paired publication is present, we split citation counts for all non-paired publications published in the same year and journal based on the lag between the publication year of the journal and grant year or the corresponding patent. This is again aggregated at the level of the journal and publication year, resulting in an average citation count pre- and post-grant for non-paired patents for the given journal and publication year. If a given journal has multiple publications with a paired patent in a given publication year, we split up forward publication citation counts for the non-paired publications multiple times, once for every lag between the publication year and the grant year of the corresponding patents. All these numbers are aggregated at the level of the journal and publication year, resulting in an average citation count pre- and post-grant for non-paired publications (pre- and post-grant of the corresponding patents linked to paired publications of the same journal and publication year). Finally, for all combinations of journals and publication years in which paired publications have been observed, we calculate the ratio between citation received by paired publications and non-paired publications two times: for the pre-grant period as well as the post-grant period. If an anti-commons effect would manifest itself, the citation ratio between paired publications and non-paired publications would drop significantly after granting of the corresponding patents.

As Table 7-6 indicates, the ratio of average citations received by paired publications and non-paired publications equals to 1.71 for the pre-grant period and 1.74 for the post-grant period. Controlling for journal and publication year, these figures mean that paired publications, i.e. publications having a patent counterpart, receive on average 1.71 more citations in the pre-grant period compared to publications without patent counterparts for the same period, and that paired publications receive on average 1.74 more citations in the post-grant period compared to publications without patent counterpart for the same period. While these descriptive statistics do not indicate a decline, a formal t-test reveals that both ratios are not significantly different ($p=0.86$).

As such, our data do not show any sign of anti-common effects that become visible after patent rights have been granted.

Table 7-6 : Results of independent sample T-test – Ratio average citations paired/non-paired publication pre-grant versus post-grant (1991-2000)

Variable	Class	N	Lower cl mean	Mean	Upper cl mean
Ratio average citations pairs/non-pairs	Pre-grant	288	1.42	1.71	2.00
Ratio average citations pairs/non- pairs	Post-grant	288	1.48	1.74	2.00
Diff	(1-2)		-0.43	-0.03	0.36

T-TESTS

Variable	Method	Variances	DF	t value	Pr > |t|
Ratio average citations pairs/non-pairs	Pooled	Equal	574	-0.17	0.8666
Ratio average citations pairs/non-pairs	Satterthwaite	Unequal	565	-0.17	0.8666

EQUALITY OF VARIANCES

Variable	Method	Num DF	Den DF	F value	Pr > F
Ratio average citations pairs/non-pairs	Folded F	287	287	1.29	0.0299

Murray & Stern (2007) made use of a natural experiment as USPTO patent applications prior to 2001 were not published until grant, making the granting of a patent an external shock. However, we have both EPO and USPTO in our dataset, and EPO patents are published 18 months after filing. This means that the granting of an EPO patent does not come to a complete surprise as researchers can be aware of pending patent applications far before the granting of the EPO patent. Hence, if an anti-commons effect is present, citation rates can already go down after the first publication of the EPO application, depending on the behaviour of the researchers when they are aware of pending patent applications. As splitting our analysis for USPTO patents (based on grant year) and EPO patents (based on publication year) is not straightforward because of patent family issues, we performed an alternative independent sample T-test in which we do not split citation counts into citation counts received before and after patent grant, but before and after the first publication of the patent document. For EPO patents, this publication date is 18 months after patent filing; for USPTO patents, this

publication date is when the patent is granted for patents up to 2001 and 18 months after patent filing for patents after 2001. After recalculation of citation lags based on publication date instead of grant date, we can again compare citation rates before and after publication (instead of grant). For this test we find similar results as the one based on grant date: controlling for journal and publication year, we find that paired publications, i.e. publications having a patent counterpart, receive on average 1.49 more citations in the pre-publication period compared to publications without patent counterparts for the same period, and that paired publications receive on average 1.54 more citations in the post-publication period compared to publications without patent counterpart for the same period. Again a formal t-test reveals that both ratios are not significantly different (p=0.71).

7.4 Findings on citation patterns of patents (patent-to-patent citations)

Within this section we report and discuss the empirical results obtained when analysing patent citations - i.e. citations from other patent document - to patent documents that are part of a patent-publication pair. This analysis implies a comparison with patent citations to patent documents which do not belong to a patent-publication pair.

Descriptive results

Table 7-7 provides a summary overview of the number of (granted) biotechnology patents under study as well as the observed average number of forward patent citations, organized by application year. Again the left side of the table contains data for all biotechnology patents, while the right side contains data for the paired biotechnology patents, i.e. patents with a counterpart in the scientific literature (patent-publication pairs).

The number of biotechnology patent grants in our dataset is starting at 3,069 patents in 1991 and exponentially growing to 9,881 patents in 1995[64]. In 1996 the number of patents falls down to 5,635 to level around (and later above) 7,000 patents in the period

[64] All patent counts mentioned in this chapter are granted patents by application year for patents having a substantial title and abstract length to be of use in text mining, unless stated otherwise.

1997 to 2001. After 2001 the number of patents gradually diminishes.[65] The average number of forward patent citations (patent-to-patent citations) for the biotechnology patents follow a negative trend, starting around 16 from 1991 to 1993 and steadily going down from 1994 onwards (a decrease of roughly 1.5 for every year)[66] [67]. The average number of forward citations for all biotechnology patents is 8.9 (median number of forward citations is 4). For the period 1991-2000 (to allow for a sufficient time lag for citations) the average number of forward citations is 11.4 (median is equal to 5).

Table 7-7 : Number of biotechnology patents and forward citations per year

	ALL BIOTECHNOLOGY PATENTS		PAIRED BIOTECHNOLOGY PATENTS	
Application year	Number of patents	Average number of forward patent citations	Number of patents	Average number of forward patent citations
1991	3,069	16.21	9	14.56
1992	3,727	16.14	11	24.09
1993	4,392	16.01	25	12.68
1994	6,170	14.39	37	11.16
1995	9,881	14.60	71	13.51
1996	5,635	12.13	33	6.45
1997	7,097	10.12	56	11.68
1998	6,974	8.30	70	8.84
1999	7,742	7.35	58	5.47
2000	7,798	5.46	65	3.52
2001	7,509	4.43	49	2.86
2002	6,315	3.06	30	3.17
2003	4,554	2.50	19	1.26
2004	3,590	2.16	23	11.52
2005	2,342	1.67	7	0.57
2006	1,170	1.61	3	0.33
2007	275	1.03	0	N/A
2008	8	0.75	0	N/A
Total/Average	88,248	8.94	566	8.21

[65] For more recent years, trends in granted patent numbers per application year are not reliable because of declining grants due to the grant lag in patent systems.

[66] In contrast with the publication citation counts (publication-to-publication citations), patent counts in this study are not counted by a fixed citation window but continuously for all succeeding years up to 2009, the last year for which we have information available. This explains the early fall in average number of citations.

[67] The patent citation counts are corrected for patent families, both at the cited as at the citing side. At the cited side, all citations to the patent and one of its DOCDB patent family members are added together. At the citing cite, citations of multiple members of the same DOCDB patent family are counted as one.

The number of biotechnology patents linked to a publication (566) follows a trend similar to the overall evolution of biotechnology patents: first an exponential growth phase starting from 9 in 1991 to 71 in 1995, followed by a drop to 33 in 1996 and a phase with numbers fluctuating around 63 between 1995 and 2000. Again the numbers gradually diminish after 2001, with a notable exception of 2004 (23 patents for 2004 versus 19 patents for 2003 and 7 patents for 2005). The average number of forward patent citations (patent-to-patent citations) for the biotechnology patents linked to a scientific publication follow a less stable pattern and fluctuate around 13 for the period 1991-1997 (with a significant positive raise to 24 in 1992 and negative drop to 6.45 in 1996) and steadily goes down from 1997 onwards (with a steep increase to 11.52 in 2004; compared to 1.26 in 2003 and 0.57 in 2005). The average number of forward citations for biotechnology patents linked to a publication is 8.2 (median number of forward citations is 3). For the period 1991-2000, the average number of forward citations is 9.5 (median is equal to 4). These averages are about 8% lower compared to non-paired patents.

As can be expected, patent-publication pairs are largely related to academic patenting; 52% of biotechnology patents that are linked to a publication have at least one academic patentee, compared to 18% for all non-paired biotechnology patents. Patents with at least one government or non-profit patentee are also overrepresented in the set of patents closely related to publications (23% for paired patents versus 10% for non-paired patents).

Multivariate analysis

In order to assess whether observed differences are statistically significant, we performed a negative binomial regression with the number of forward patent-to-patent citations as dependent variable and a dummy variable indicating whether a patent is or is not part of a patent-publication pair as independent variable. We use all 88,248 biotechnology patents having a title and abstract of substantial length (to be suited for text mining): 566 patents that are part of a patent-publication pair and 87,682 patents that are not part of a patent-publication pair.

We further control for the patent system (EPO or USPTO), the number of IPC-codes (technological specialization), the presence of an academic patentee, the number of backward scientific non-patent citations, the number of backward patent citations, the number of forward publication citations (citations from *WOS* publications to the particular patent), the number of inventors and the number of patentees. We also included dummy variables for all 11 biotechnology IPC-subclasses (4 digits) present in our selection of biotechnology patents (see Appendix 6-1 of previous chapter for all IPC-codes as used in the OECD biotechnology definition). Again we include a time variable (1 for the first year, 1991, up to 18 for the last year, 2008) and a squared time variable to include the evolution over time.

Table 7-8 : Results of negative binomial regression for 1991-2008 (dependent variable: number of forward patent citations of patents – corrected for DOCDB patent family members, both at cited and citing side)

			95% Wald Confidence Interval		Hypothesis Test		
Parameter	**B**	**Std. Error**	**Lower**	**Upper**	**Wald Chi-Square**	**df**	**Sig.**
(Intercept)	2.300	.0197	2.262	2.339	13585.101	1	.000
Pair (Y/N)	.058	.0460	-.032	.148	1.599	1	.206
Patent system							
EPO	-.193	.0140	-.221	-.166	190.450	1	.000
USPTO	0	.	.	.	.	.	.
Subfield dummies (IPC-subclasses) (n=11)				Included			
Number of IPC codes	.043	.0008	.042	.045	3265.759	1	.000
Has university patentee (Y/N)	.036	.0097	.017	.055	13.462	1	.000
Number of backward scientific non-patent citations	.003	.0002	.003	.003	248.658	1	.000
Number of backward patent citations	.016	.0003	.016	.017	3370.269	1	.000
Number of forward publication citations from WOS publications	.141	.0052	.131	.152	744.627	1	.000
Number of inventors	.018	.0019	.014	.022	92.591	1	.000
Number of patentees	-.010	.0074	-.025	.004	1.945	1	.163
Time	-.063	.0042	-.071	-.055	228.478	1	.000
Time²	-.007	.0003	-.008	-.007	766.940	1	.000

Table 7-8 reports the results of the regression analysis of forward patent citations of patents. Patents being part of a patent-publication pair have more forward publication

citations (variable Pair Y/N), but the difference is not significant. USPTO patents have more citations than EPO patents. All other controlling variables have a significant and positive impact, except for the number of patentees, which has a negative but not significant impact, and time, which displays a decreasing, curvilinear relationship with patent citations.

After removing outliers, i.e. all patents with a forward citation count larger than the mean plus three times the standard deviation, similar results are obtained as the ones reported in Table 7-8. Finally, when we limit the time period to all patents applied for between 1991 and 2000 – in order to allow all patents to have at least 10 years of forward patent citations – patent that are part of a patent-publication pair have less forward patent citations, but also this difference is not significant (both when including and excluding outliers). Overall, we observe no significant difference in terms of (forward) patent citations when comparing patents that are associated with a scientific publication with their solitary counterparts.

7.5 Conclusions, discussion and directions for further research

In this study, we have applied a text mining methodology to examine the possible presence of anti-commons effects in biotechnology research. Inspired by previous work undertaken by Murray, Stern and others, we analysed citation flows stemming from patent-publication pairs present within the field of biotechnology (scientific publications from which the contents - methodology, findings, discovery - is part of a patent application). The delineation of the biotechnology domain was based on the use and the refined application of existing classification schemes. An elaborate text mining scheme was developed and implemented in order to identify and validate the patent-publication pairs. A total of 584 patent-publication pairs were ultimately included in the citation analysis. The necessary validation and control strategies were introduced and executed. After taking into account these controls and studying the citation patterns of the documents included in the patent-publication pairs, we were not able to detect a significant anti-commons effect on the basis of the 584 pairs identified. On the contrary, scientific publications belonging to a patent-publication pair receive significantly more

scientific citations than their counterparts for which no patent document has been identified. We also do not find a significant difference for citations rates before and after patent publication or grant. Also we do not find differences in the citation rates of publications linked to a patent with an academic or non-profit patentee compared to publications linked to a patent with an academic or non-profit inventor but no academic or non-profit patentee (hence commercial patentee). As such, our findings do not reveal the presence of anti-commons effects once scientific findings become translated into intellectual property rights (in this case, patents). In terms of technological citations, we observed no difference between patents belonging to a patent-publication pair and patent documents that are not associated directly with a scientific publication. As such, no additional impact – on future technological developments – is observed when patent documents are situated in the vicinity of science.

These findings add to the current stock of insights on the interaction between patenting and publication behaviour. Through the design and application of text mining techniques on a broad set of data, we intended to take the current insights a step further. Extensive validation efforts were undertaken in order to confirm the results obtained. These results definitely are an invitation to further examine the joint effects of patenting and publishing activities by scientists.

However, our current approach also has limitations. A first point of attention is the limited number of patent-publication pairs identified by the method; 584 patent-publication pairs for all biotechnology patents and (*WOS*-covered) publications from 1991 to 2008 is low compared to 169 paired publications found by Murray & Stern (2007) in *Nature Biotechnology* in the period 1997-1999. Although relaxing our criteria to identify patent-publication pairs and using these additional patent-publication pairs in our analyses do not undermine our findings, it is worth to find out why we are missing patent-publication pairs compared to the manual method of Murray & Stern (we only find 9 Nature Biotechnology publications paired to a patent for the same period). An inverse search approach, in which we first match patents and publications based on inventor/author name matching followed by a text mining approach to assess content similarity and eliminate false matches because of homonyms might reduce

recall, without jeopardizing high precision levels compared to traditional name matching techniques, and is definitely worth trying.

A second point of attention is the difficulty to control for the heterogeneity in the large set of biotechnology publications compared to the small set of publications that are part of a patent-publication pair. This makes it difficult to distinguish underuse because of a potential anti-commons effects from general qualitative differences when observing differences in citation counts (especially because there is a bias for paired publications as we can expect to find more publications of higher quality in that group – i.e. publications valuable enough to justify costs and efforts to apply for a patent). Murray & Stern (2007) use a natural experiment based on the non-disclosure of USPTO patent filings prior to 2001 to grasp direct anti-commons effects. Besides this approach, we control for general quality differences by matching paired and non-paired publications on journal and volume year, assuming journal impact factors are a good indication of the average quality and citation rate of published publications. However, robustness of results would benefit from additional matching criteria or control variables to further rule out general quality differences. E.g. sector and country of affiliations of publications, and citation network information might be included. Another interesting and feasible experiment is to compile a dataset of all publications of all authors having a publication paired with a patent, and look for differences in citation rates between the paired and non-paired publications of the same author.

Another point of attention that arises is the one of generalization towards other fields of 'techno-scientific' economic activity. Can we substantiate the current findings in technology domains such as materials or in other fields? And can we corroborate and consolidate the robustness of the text mining methodology that was deployed, as well as a further, independent, confirmation of the optimal identification algorithm. A last point pertains to the continuous cross-validation of the results obtained with our method with the results obtained by sets of patent-publication pairs that have been constructed manually by experts, like the Murray & Stern dataset mentioned before.

Besides previous points, disentangling patent-publication pairs by their nature deserves more attention. In line with Czarnitzki, Glänzel & Hussinger (2009) we did already look

at the differences between patent-publication pairs with an academic or non-profit patentee compared to those with an academic or non-profit inventor, but additional research is needed to get more insight in the dynamics and heterogeneity of patentees and publishers.

Finally, the absence of an anti-commons effect does not imply that we have reached the end of the patent-publication debate. On the contrary, we still need a far better understanding of the many, often multidimensional, spillovers that involvement of scientists in both patent and publication activities can bring and generate. These spillovers do not only occur at the material level, but also at the immaterial, cognitive level. Understanding them and linking them to the performance of scientists in setting and advancing their research agendas, remains a question of primary importance. A better insight into these substantive relationships, both at the personal level and at the institutional level, can indeed only improve our understanding of the effective and fruitful management of scientific activity.

7.6 References

Czarnitzki, D., **Glänzel**, W. & **Hussinger**, K. (2009). "Heterogeneity of patenting activity and its implications for scientific research." *Research Policy*, 38 (1) : 26-34.

Heller, M. A. & **Eisenberg**, R. S. (1998). "Can Patents Deter Innovation? The Anticommons in Biomedical Research." *Science*, 280 : 698-701.

Murray, F. & **Stern**, S. (2007). "Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis." *Journal of Economic Behavior and Organization*, 63 : 648-687.

OECD (2005). A Framework for Biotechnology Statistics, 29-32. Paris: OECD Publishing.

OECD (2009). OECD Biotechnology Statistics. Paris: OECD publishing.

8 Methodological part: application of text mining techniques to identify science-technology interactions.

Essentially, all models are wrong, but some are useful.
George E. P. Box

8.1 Summary and conclusions

In our methodological part we thoroughly assessed LSA as a text mining technique to match patent and publication documents based on their contents. The goal is to find patent and publication documents that are related by the topics they address, the methods they use, the results they obtain and the inventions or discoveries they address. This would bypass limitations of current approaches based on IPC-codes, non-patent references, and patent inventor and patentee matching with publication author and affiliation, and allow to compile large scale datasets for a broad range of applications in innovation studies. On the one hand we used small-scale patent and publication datasets of six academic inventors to examine the feasibility of matching patents with publications using a Vector Space Model and a Latent Semantic Analysis text mining approach. On the other hand, we did a large scale matching exercise for biotechnology patents and publications.

As off-the-shelve text mining solutions are not readily available and experience with patent data is limited, we have set up a large measure variant comparison exercise based on the LSA method. Several options for obtaining similarity measures within the framework of this model – based on multiple weighting methods and multiple levels of dimensionality reduction – have been outlined and assessed in terms of precision and recall.

Our findings reveal that different options and methods available coincide with considerable differences in terms of accuracy. While several combinations allow us to

arrive at acceptable solutions, certain combinations display low levels of accuracy and even result in misleading similarity measures, both for the exercise with the small samples as for the exercise with the large set of biotechnology patents and publications. These results suggest that LSA-based measures tend to overestimate similarity and not grasp the real topic similarity of patent and publication documents. LSA seems not to redeem its promise to deal with synonymy and polysemy problems in our setting; SVD dimensionality reduction results in less congruence with the expert assessment of similarity compared with cosine measures applied on the full vector space, and the less dimensions retained, the less congruence.

The term weighting method used also effects the performance; for small datasets we only compared raw frequencies and TF-IDF weighting and observe that TF-IDF weighting works best. However, for the large dataset we also included binary weighting and IDF weighting and observed that those weighting methods yield better results compared to TF-IDF weighting and raw term frequencies, a remarkable observation as TF-IDF in combination with SVD retaining 300 to 500 dimensions is a commonly used method. Weighting methods including term frequencies suffer from tokenization and parsing errors, explaining the better performance of binary and IDF weighting. We believe this is a dataset specific problem caused by tokenization and parsing errors, reinforced because of term frequencies, of chemical formula which are rather common in our technical dataset.

We observe that a cosine metric applied on the full vector space after binary or IDF weighting yields the best results. However, simple measures based on the number of terms documents have in common perform slightly better, in line with Occam's razor principle. The claim that LSA can outperform such simple measures based on common terms or co-occurrence because of a better understanding of the meaning of language of this former method is not backed up by our data.

We propose a combination of measures that allow a more robust identification of similar patent and publication documents: 'common terms MIN', the measure based on the number of common terms weighted for the minimum of the number of terms of the patent and the publication document, as a primary criterion to identify similar

documents, combined with 'common terms MAX', the measure based on the number of common terms weighted for the maximum number of terms of the patent and the publication document, as a secondary criterion to eliminate doubt cases due to combinations of short and long documents. Especially when precision is important, those measures deliver good results. When completeness or recall gets important, things get more complicated because there are no threshold values that allow a clear cut distinction between groups.

When it comes to the identification of patent-publication pairs, i.e. scientific publications from which the contents is covered by patent protection, quality of the results can benefit from an additional third criterion based on the presence or absence of a shared inventor/author. Although inventor-author name matching is not straightforward for larger datasets because of homonymy problems, spelling errors and name variation, and the use of middle names and initials, the combination of a content based measure like our 'common terms MIN' and 'common terms MAX' and the presence of a shared inventor/author might be the way to go, because the biggest challenge in inventor-author name matching – the homonymy issue – is largely controlled for when combined with a content bases measure.

8.2 Limitations and directions for further research

It is clear that the outlined automated method still have limitations and does not work well in all circumstances. The typical trade-off between precision and recall remains a though one, especially as final results are very sensitive to threshold values: small changes in the threshold values for both the primary as secondary criterion result in big differences in the number of matches in the total population. This is particularly problematic for the secondary criterion 'common terms MAX', needed to clear out doubt cases: the vast majority of potential matches based on the primary criterion 'common terms MIN' score very low on 'common terms MAX', so small changes in the range of 'common terms MAX' to discard doubt cases (0.20-0.35) have a huge impact. It seems that our method suffers from too many documents with short abstracts that are very difficult to judge. On the other hand, this seems not to be due to the shortcomings

of an automated method, but due to the characteristics of the dataset. There are simply too many documents with short abstracts that seem to have some relatedness with documents with large abstracts, and even for human experts these combinations are very difficult to assess. We might have to conclude that for these document combinations accurate assessment of similarity is simply impossible. A potential remedy is to extend document sizes by including patent claims or full document contents - and not only title and abstract - into the analysis, or use extended abstracts as supplied e.g. by the *Derwent World Patent Index*.

Regardless of the problems with document combinations that differ largely in size, the number of patent-publication pairs revealed seems low compared to the total population of patents and publications involved. Although precision is high, recall is a problem. The measurement of the thru recall rate in the overall population is difficult. Our validation set does not contain that many missed matches, and hence recall rates seems high, but the global number of patent-publication pairs seems so low compared to the total population of patents and publications that there is no doubt a substantial amount of patent-publication pairs is missed. Ways to improve recall are not clear because we only do have a very limited amount of false negatives in our validation sample. Cross-validation with other datasets with patent-publication pairs would be very valuable to get more insights why patent-publication pairs are missed.

Improving precision and recall levels might be feasible by further broadening the set of pre-processing options. For instance, when inspecting several patent-publication pairs, it became apparent that introducing more synonyms or collocations and phrase detection might further contribute to improving results. More advanced feature selection techniques would also improve the performance of weighting methods that take term frequencies into account, like TF-IDF weighting. Hence, research focusing on the precise impact of additional parameters not included in this design seems relevant. However, practical use might be limited because these additional processing options require considerable manual intervention and this might not be feasible for every single text mining exercise. Unless one would be able to derive synonym lists and collocations that are specific for patents or publications, or that are relevant for a particular science

or technology field. One route definitely worth pursuing is the improvement of tokenization and parsing to eliminate errors - especially from chemical formulas which are rather common in our technical dataset - which are getting reinforced because of term frequencies. We believe this will improve the performance of TF-IDF weighting. Related to this issue are stemming errors, also reinforced by term frequencies. One might consider not to use stemming and try to solve synonymy problems using more advanced feature selection techniques.

A remarking observation is the poor performance of SVD-based measures. It is not clear why the specific context of our data does not allow the LSA-method to achieve its full potential. It is unlikely that our dataset size is not large enough, nor that we did not retain enough dimensions/concepts. There are indications that the document size and document size differences are negatively influencing the SVD-based measures. But it might also be due to the particular language use in our patent and publication dataset. Again, using the full text of patent and publication documents, or extended abstracts as supplied by the *Derwent World Patent Index*, might resolve this, although we lack hard evidence that larger or better abstracts would resolve the issues.

What is clear is that, for our dataset, the dimensionality reduction imposed by LSA/SVD is cutting off valuable information instead of noise. We observe a gradually increasing performance for increasing number of retained dimensions, but we do not observe a range of dimensions for which the performance is better than that of a cosine measure applied on the full vector space; the performance of LSA/SVD is just approaching the performance of a cosine measure on the full vector space for higher numbers of retained dimensions, in contradiction to the claims of LSA that dimensionality reduction would improve results (understanding the 'latent' structure).

Another remarkable observation is that patents of our materials control set are – on average – more related to biotechnology publications than are biotechnology patents when SVD-based measures are used. This information can act as a source of inspiration to reveal the shortcomings of SVD on our data. However, looking into many individual cases did not reveal significant information to explain the higher obtained similarity scores nor the poor performance of SVD.

Disentangling the bad performance of SVD in general is very difficult. Looking at individual cases is not very informative as it is virtually impossible to trace back term vectors after SVD to the original terms and contents. The document-by-concept matrix compiled by the SVD solution contains the scores of all documents on newly formed latent concepts, and every latent concept consists of a linear combination of all original terms, i.e. a linear combination with 301,697 components. Although multiple reasons of the limited performance of the LSA application on our dataset can be put forward, hard evidence is limited. We are almost certain that the chosen level of dimensionality reduction is not too low and the problem resides in the fact that LSA is not grasping the latent structure and cutting valuable information. There are some indications that the document size and size differences are causing problems, and we are almost certain that some of the observed weighting issues are due to tokenization, parsing and stemming errors. However, testing each and every suspicion is very time consuming. We propose first to deal with tokenization and parsing issues, and next conduct additional tests on the impact of document size and size differences and try our larger documents (full patent and publication documents or e.g. *Derwent Abstracts*).

One final reason why LSA might fall short is the limitation to Euclidean geometry as imposed by the assumption of LSA that documents are represented as vectors in a vector space. In an Euclidean space, similarity should be symmetric and not violate triangle inequality - $d(x,z) \leq d(x,y) + d(y,z)$ - placing strong constraints on the location of points in a space given a set of distances (Griffiths, Steyvers & Tenenbaum, 2007). This issue however is a general one and not directly related to the limitations of our patent and publication dataset; it is related to problems when dealing with high-dimensional spaces ('curse of dimensionality'). In high dimensional spaces all data appear to be sparse and dissimilar, preventing efficient identification of communalities. Other text mining techniques not relying on spatial representations might be more appropriate, like generative topic models as Probabilistic Latent Semantic Modelling (Hofmann, 1999) and Latent Dirichlet Allocation (Blei, Ng & Jordan, 2003).

To conclude, the debate about the value of more complex text mining methods for application on patent and scientific publication data – complex in the sense that they try

to deal with the characteristics of text and language – compared to simpler methods based on common terms or co-occurrence does not end here. While claims of LSA are not backed up by our observations, and simpler seems to be better – see also Occam's Razor principle - denoting where exactly it goes wrong with LSA remains tough, but this does not refrain us to continue with a method to detect patent-publication similarities based on content relatedness for dealing with our empirical questions.

9 Empirical part: impact of science-intensity of patents and the potential threat of an anti-commons effect.

Once a new technology rolls over you,
if you're not part of the steamroller,
you're part of the road.
Stewart Brand

9.1 Summary and conclusions on the impact of science-intensity of patents

First we had a look at the impact of science-intensive patents on technological development at the country level. We selected all granted USPTO biotechnology patents and *WOS* publications (Thomson Reuters ISI Web of Science) for the period 1992-1999 for 20 countries having patent activity in biotechnology throughout the whole time period. We used the number of examiner-given front page non-patent references of the USPTO patents as an indicator of science-intensity, and the number of granted patents, divided by the population, as indicator for the technological productivity of a country. We correct for the scientific productivity and used the number of publications, again divided by the population, as indicator for the scientific productivity of a country.

Our findings revealed that within the field of biotechnology, the science-intensity or science proximity – as measured by the amount of non-patent references – of patents is positively associated with technological productivity. The relationship between science and technology within the field of biotechnology indeed reveals itself here as reciprocal and bi-directional rather than unidirectional or linear. These findings corroborate the construct validity of indicators based on non-patent references found within patents. In addition, the positive relationship between science-intensity, or stated otherwise, the closeness between science and technology, and technological productivity, corroborates the relevancy of policy frameworks that foster interaction between

knowledge/science generating institutions (universities, research centres) and technology producers (companies).

9.2 Limitations and directions for further research on the impact of science-intensity

These findings also suggest interesting avenues for further research. While we focused on one specific field (biotechnology), refining the insights obtained in terms of their field specific nature requires extensions towards other fields. Likewise, introducing extended time frames would allow assessing whether differential effects are to be observed related to technological life cycle dynamics. Extending the analysis to include other patent system and different counting methods (see Guellec & Van Pottelsberghe, 2001) seems more than worthwhile to pursue in order to assess the robustness or the peculiarities of the findings obtained. Finally, additional insights in the value of non-patent references as indicator of scientific intensity and additional indicators of science-intensity are very valuable to include in this exercise. In this respect we think about the presence of an academic inventor or an inventor that also publishes scientific publications, or patents cited by scientific publications. And of course, content base text mining methods, as developed in this dissertations, could also be instrumental to get a better grasp of science-intensity, both directly (identification of patent-publication pairs) and indirectly (helping to solve homonymy problems when identifying academic inventors and inventors involved in scientific publishing).

9.3 Summary and conclusions on the potential threat of an anti-commons effect

As observed in our first study, the increasing science-intensity of patenting seems to have a positive effect on technological performance. One aspect of this 'scientification' of patenting is the proliferation of academic patenting. The backside of this phenomenon is the 'privatization' of scientific commons, altering the model of open science and potentially hampering scientific progress because of the blocking power of patent holders. This fear is nicely expressed by the metaphor of the 'Tragedy of the anti-

commons' by Heller, referring to the underuse of scarce resources because of too much ownership.

In this study, we have applied a text mining methodology to examine the possible presence of anti-commons effects in biotechnology research. Inspired by previous work undertaken by Murray, Stern and others, we analysed citation flows stemming from patent-publication pairs present within the field of biotechnology. The delineation of the biotechnology domain was based on the use and the refined application of existing classification schemes. An elaborate text mining scheme was developed and implemented in order to identify and validate the patent-publication pairs. A total of 584 pairs were ultimately included in the citation analysis. The necessary validation and control strategies were introduced and executed. After taking into account these controls and studying the citation patterns of the documents included in the patent-publication pairs, we were not able to detect a significant anti-commons effect on the basis of the 584 pairs identified. On the contrary, scientific publications belonging to a patent-publication pair receive significantly more *scientific* citations than their counterparts for which no patent document has been identified. We also do not find a significant difference for citations rates before and after patent publication or grant. Also we do not find differences in the citation rates of publications linked to a patent with an academic or non-profit patentee compared to publications linked to a patent with an academic or non-profit inventor but no academic or non-profit patentee. As such, our findings do not reveal the presence of anti-commons effects once scientific findings become translated into intellectual property rights (in this case, patents). In terms of technological citations, we observed no difference between patents belonging to a patent-publication pair and patent documents that are not associated directly with a scientific publication. As such, no additional impact – on future technological developments – is observed when patent documents are situated in the vicinity of science.

These findings add to the current stock of insights on the interaction between patenting and publication behaviour. Through the design and application of text mining techniques on a broad set of data, we intended to take the current insights a step

further. Extensive validation efforts were undertaken in order to confirm the results obtained.

Our findings backup policy frameworks that encourage science-technology interactions and the concept of the entrepreneurial university and academic patenting. The threat of an anti-commons effect did not reveal itself in our data. To the extent that academic patenting has a positive influence on both academia and industry – complementarities and spillovers, market of ideas, additional funding for (basic) research – this observation might be reassuring. However one has to bear in mind that a potential anti-commons effect is not the only threat of academic patenting. Although literature suggests that academic patenting has no negative impact on the quantity and quality of the scientific output of involved academic inventors, doubts still raise whether this trend might cause a shift in the orientation of research, and especially in a shift to more – potentially profitable – applied research away from basic research, also potentially jeopardizing scientific and technological development in the long run.

Finally, the absence of an anti-commons effect does not imply that we have reached the end of the patent-publication debate. On the contrary, we still need a far better understanding of the many, often multidimensional, spillovers that involvement of scientists in both patent and publication activities can bring and generate. These spillovers do not only occur at the material level, but also at the immaterial, cognitive level. Understanding them and linking them to the performance of scientists in setting and advancing their research agendas, remains a question of primary importance. A better insight into these substantive relationships, both at the personal level and at the institutional level, can indeed only improve our understanding of the effective and fruitful management of scientific activity.

9.4 Limitations and directions for further research on the potential threat of an anti-commons effect

These results definitely are an invitation to further examine the joint effects of patenting and publishing activities by scientists. However, our current approach also has limitations.

The first point relates to corroborating and consolidating the robustness of the text mining methodology that was deployed, as well as a further, independent, confirmation of the optimal identification algorithm. Although we strongly believe in the precision or accuracy of the method, as mentioned already in the conclusions of the methodological part, doubts arise when it comes to the recall or exhaustiveness of the method, resulting in an undersampling of patent-publication pairs in our study. 584 patent-publication pairs for all biotechnology patents and (*WOS*-covered) publications from 1991 to 2008 is low compared to 169 paired publications found by Murray & Stern (2007) in *Nature Biotechnology* in the period 1997-1999. Although relaxing our criteria to identify patent-publication pairs and using these additional patent-publication pairs in our analyses does not undermine our findings, it is worth to find out why we are missing patent-publication pairs compared to the manual method of Murray & Stern (we only find 9 Nature Biotechnology publications paired to a patent for the same period). An inverse search approach, in which we first match patents and publications based on inventor/author name matching followed by a text mining approach to assess content similarity and eliminate false matches because of homonyms, might reduce recall, without jeopardizing high precision levels compared to traditional name matching techniques, and is definitely worth trying.

A second point of attention is the difficulty to control for the heterogeneity in the large set of biotechnology publications compared to the small set of publications that are part of a patent-publication pair. This makes it difficult to distinguish underuse because of a potential anti-commons effects from general qualitative differences when observing differences in citation counts (especially because there is a bias for paired publications as we can expect to find more publications of higher quality in that group – i.e. publications valuable enough to justify costs and efforts to apply for a patent). Murray & Stern (2007) use a natural experiment based on the non-disclosure of USPTO patent filings prior to 2001 to grasp direct anti-commons effects. Besides this approach, we control for general quality differences by matching paired and non-paired publications on journal and volume year, assuming journal impact factors are a good indication of the average quality and citation rate of published publications. However, robustness of results would benefit from additional matching criteria or control variables to further

rule out general quality differences. E.g. sector and country of affiliations of publications, and citation network information might be included. Another interesting and feasible experiment is to compile a dataset of all publications of all authors having a publication paired with a patent, and look for differences in citation rates between the paired and non-paired publications of the same author.

Another point of attention that arises is the one of generalization towards other fields of 'techno-scientific' economic activity. Can we substantiate the current findings in technology domains such as materials or in other fields? And can we corroborate and consolidate the robustness of the text mining methodology that was deployed, as well as a further, independent, confirmation of the optimal identification algorithm. A last point pertains to the continuous cross-validation of the results obtained with our method with the results obtained by sets of patent-publication pairs that have been constructed manually by experts, like the Murray & Stern dataset mentioned before.

Besides previous points, disentangling patent-publication pairs by their nature deserves more attention. In line with Czarnitzki, Glänzel & Hussinger (2009) we did already look at the differences between patent-publication pairs with an academic or non-profit patentee compared to those with an academic or non-profit inventor, but additional research is needed to get more insight in the dynamics and heterogeneity of patentees and publishers.

Finally we also want to remark that citation patterns are only one aspect of the diffusion of knowledge and follow-on research and therefore comparison of citation patterns is not the ultimate method to shed a light on the anti-commons issue. First of all not all IPR-protected scientific discoveries will be published in the form of a scientific publication, and this kind of 'privatized' scientific knowledge remains unobserved in our method based on the identification of patent-publication pairs. On the other hand there are many reasons to cite previous work in a scientific publication – not necessarily implying knowledge diffusion or follow-on research – and knowledge diffusion is more than just citing previous work. We also must bear in mind that individual scientists feeling blocked by a patent for their own research have different means to deal with this threat without necessarily jeopardizing their follow-on research. They can also try

to circumvent or invent around, or even to ignore the threat if they believe that chances of litigation are small. These adaptation strategies will be highly situational and for instance different for immaterialized knowledge that is easy to reproduce, research tools, or materialized knowledge like material transfer agreements. The role of research exemption should be mentioned here too, turning the debate on academic patenting in a complex story that cannot by grasped by one study or indicator. More fundamental insights are needed on the barriers scientists encounter with potentially blocking patents and the way they deal with it. We still need a far better understanding of the many, often multidimensional, spillovers that involvement of scientists in both patent and publication activities can bring and generate. These spillovers do not only occur at the material level, but also at the immaterial, cognitive level. Understanding them and linking them to the performance of scientists in setting and advancing their research agendas, remains a question of primary importance. A better insight into these substantive relationships, both at the personal level and at the institutional level, can indeed only improve our understanding of the effective and fruitful management of scientific activity.

Doctoral dissertations from the Faculty of Business and Economics, see:
http://www.kuleuven.ac.be/doctoraatsverdediging/archief.htm.