Scientometrics (2010) 82:289–306
DOI 10.1007/s11192-009-0046-6

Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications

Tom Magerman · Bart Van Looy · Xiaoyan Song

Received: 3 July 2008 / Published online: 10 June 2009

Abstract In this study, we examine and validate the use of existing text mining techniques (based on the vector space model and latent semantic indexing) to detect similarities between patent documents and scientific publications. Clearly, experts involved in domain studies would benefit from techniques that allow similarity to be detected—and hence facilitate mapping, categorization and classification efforts. In addition, given current debates on the relevance and appropriateness of academic patenting, the ability to assess content-relatedness between sets of documents—in this case, patents and publications—might become relevant and useful. We list several options available to arrive at content based similarity measures. Different options of a vector space model and latent semantic indexing approach have been selected and applied to the publications and patents of a sample of academic inventors ($n = 6$). We also validated the outcomes by using independently obtained validation scores of human raters. While we conclude that text mining techniques can be valuable for detecting similarities between patents and publications, our findings also indicate that the various options available to arrive at similarity measures vary considerably in terms of accuracy: some generally accepted text mining options, like dimensionality reduction and LSA, do not yield the best results when working with smaller document sets. Implications and directions for further research are discussed.

T. Magerman (✉) · B. Van Looy · X. Song
Centre for R&D Monitoring (ECOOM), Dekenstraat 2, 3000 Leuven, Belgium
e-mail: tom.magerman@econ.kuleuven.be

B. Van Looy
e-mail: bart.vanlooy@econ.kuleuven.be

X. Song
e-mail: xiaoyan.song@econ.kuleuven.be

T. Magerman · B. Van Looy · X. Song
Faculty of Business and Economics, Department of Managerial Economics, Strategy and Innovation, K. U. Leuven, Naamsestraat 69, 3000 Leuven, Belgium

B. Van Looy
INCENTIM, Leuven Research & Development, Naamsestraat 69, 3000 Leuven, Belgium

Keywords Text mining · Latent Semantic Analysis · Science–technology linkages · Patent–publication pairs · Author–inventor relationships

Introduction

Science and technology policies rely on a diverse set of indicators pertaining to scientific, technological and innovative activities both on a national and regional scale (e.g. NSF 2006; European Commission 2003). Such indicators are used to map and compare activities by country or region but also to analyse specific fields and domains. Both bibliometric indicators pertaining to scientific activity (publications) and technometric indicators pertaining to technology (patents) figure prominently in such analysis. At the same time, it can be noted that the majority of studies and analyses undertaken in this area focus on mere counts and/or quantitative attributes of relevant documents (e.g. number of publications, patents, number of references, number of citations); less attention is paid to analysis in which the content of the underlying documents is the focal point of attention.

In this study, 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 et al. 1986; Hinze and Grupp 1996; Glenisson et al. 2005a, b; Rabeharisoa 1992, in fuel cells; Noyons et al. 1994, in laser medicine; Schmoch 2004, in genetics; Glänzel et al. 2004, 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 (Van Looy et al. 2004, 2006; Calderini et al. 2005; Azoulay et al. 2006; Fabrizio and Diminin 2005; Murray and Stern 2005; Meyer 2006) reveals that, under certain conditions, combining scientific and technological activities yields certain beneficial effects, including scientific productivity. 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 spill-over 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.

While our previous research focused on the relevance of lexical text analysis for the purpose of domain studies by targeting one activity realm, namely patent documents (Van Dromme et al. 2006), the present analysis includes both patent and publication documents stemming from one and the same academic researcher. Contrary to domain studies, which often involve thousands of documents, the number of relevant documents under consideration is much smaller in this case. 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 document—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.

In this study, we aim to assess the accuracy of Latent Semantic Analysis based lexical text analysis techniques to construct distance measures that are well suited to grasp

similarities between patent and publication text documents. The paper is structured as follows; in the following section, we first elaborate upon current practices developed in the field of text mining and Latent Semantic Analysis. It will become apparent that different options and methods are available to arrive at similarity measures. This variety of possible approaches is then translated into the research design: for the academic inventors under study ($n = 6$), we will calculate a set of distance measures ($n = 23$) for all scientific papers 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 allow us not only 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 to small-scale applications, implying documents of a heterogeneous nature.

Vector space model/Latent Semantic Analysis based text mining procedure

Quantitative linguistics dates back to at least the middle of the 19th century (see Grzybek and Kelih 2004). However, the classical theoretical work by Zipf (1949) is considered pioneering in quantitative linguistic (or text) analysis. Since the 1970s, a 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 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 and Schütze (2000) provide a comprehensive introduction to the statistical analysis of natural language, Berry (2003) provided a survey of text mining research, Leopold et al. (2004) give an overview of data and text mining fundamentals for science and technology research, and Porter and 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 et al. (1994), Bassecoulard and Zitt (2004), Leydesdorff (2004), Glenisson et al. (2005a, b) and Janssens et al. (2006).

Text mining requires a mathematical representation of textual data to describe sets of text documents in such a way that traditional data mining techniques can be used.

The vector space model (VSM) is a typical algebraic representation of text documents commonly used in information retrieval. This 'bag-of-words' approach, whereby the number of occurrences of each word in a text is counted, can be seen as a simple yet powerful representation (Salton 1968; Salton et al. 1975; Salton and 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 encountered in that document. Altogether, these document vectors 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.

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

document with as many elements as the total number of words in the global vocabulary. For words appearing in the document at hand, the value of the respective elements is equal to the number of occurrences of that word in the document. For words not appearing in the document, the respective 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 very sparse (containing mostly zeros).

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. 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 et al. 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 sharing the same stem are mapped onto the same concept or stem), but occasionally stemming will create new homonyms because of stemming errors.[1]

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 and McGill 1983), in which the raw term

[1] 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 et al. (1981), Harman (1991), Krovets (1995), and Porter (2001).

frequencies are multiplied by the inverse document frequency (IDF) for that term; this results in upgrading the impact of relatively rare terms when calculating distance measures:

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

with

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

and n the number of documents.

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 and 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 and 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).[2]

Dimensionality reduction and Latent Semantic Analysis

Natural language text is noisy due to inconsistencies, typographical errors, author's style, choice of words, and so on; it is further complicated by phenomena such as synonymy and polysemy (words with multiple meanings). Latent Semantic Analysis (LSA) constructs a concept-by-document matrix by using a low-rank approximation of the term-by-document matrix, combining dimensions or terms into 'concepts' (Deerwester et al. 1990). The rank lowering is expected to merge dimensions associated with terms that have similar meanings, increasing the power to really grasp meaningful relations between documents. LSA presumes some underlying latent semantic structure in the data and uses statistical techniques to approximate this latent structure. In the context of text indexing, LSA is also referred to as Latent Semantic Indexing (LSI).

LSI builds upon semantic similarity and hence uses proximity models such as clustering, factor analysis and multidimensional scaling (see Carroll and 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 and van Rijsbergen 1971) and factor analysis (Atherton and Borko 1965; Borko and Bermick 1963; Ossorio 1966); LSI builds further on these factor analysis techniques.

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

In practice, LSI is implemented by using Singular Value Decomposition. A theorem by Eckart and Young (1936) states that the rank-k approximation provided by the Singular Value Decomposition (SVD) is the closest rank-k approximation:

$$\left\|A - A_{k(SVD)}\right\|_2 = \min_{rank(B) \leq k} \|A - B\|_2 = \sqrt{\sigma_{k+1}^2 + \ldots + \sigma_n^2}.$$

The actual dimensionality reduction is then realized by truncating the SVD decomposition:

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

with A the original term-by-document matrix, Σ a diagonal matrix of singular values ($\sigma_1^2 > \sigma_2^2 > \ldots > \sigma_n^2$), and U, V containing orthogonal columns of left and right singular vectors, so that only its k largest singular values and corresponding dimensions of U and V are retained:

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

Thus, instead of working in the original m-dimensional vector space of the term-by-document matrix $A^{m \times n}$, we are now able to work with the k-dimensional right singular vectors $V^{k \times n}$ of the so-called concept-by-document space resulting from the SVD.

The k dimensions in the reduced space, or concept space, are now no longer mere words or stems, as in the original vector space, but linear combinations of such linguistic terms or stems. Therefore, the basic unit of analysis has become not just a mere word but a word-and-its-context, a concept (hence, the denomination 'concept space'). Typically, a few 100 dimensions (Landauer and Dumais 1997) seem to work best. However, the best choice for k might be database dependent, as suggested by Berry and Browne (1999). In practice, LSI has proven to partly deal with the synonymy problem and, to a lesser extent, with the polysemy problem.

Similarity or distance calculation

The similarity measure typically used in information retrieval applications (Berry and Browne 1999) is the cosine similarity measure. 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. This no longer holds for vectors in the reduced concept space after SVD, since vector elements may be negative, resulting in a concept-by-document space V^{kxn} that is no longer positive semi-definite, and 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 and Ribeiro-Neto 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).

Before moving to our research design, we wish to stress that the proposed VSM/LSA methodology is only one—albeit commonly used—method for text content based similarity deduction.[3]

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 and van Raan 1994; Bassecoulard and Zitt 2004, for some meso and macro level application of lexical patent and publication coupling).

Moreover, while text mining may be relevant for natural language documents, only titles and abstracts are widely and easily available for patents and publications (large sets of full-text documents are difficult or expensive to obtain), and especially patent abstracts rarely read as natural language.

Moreover, as the previous section 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 VSM and 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

Six academic inventors from the Catholic University of Leuven—four from the medical faculty and two from the engineering faculty—were taken as a starting point. All WO, EPO and USPTO patents were downloaded from MicroPatent where the six professors appear 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 extracted. Next, all publications of these professors appearing in the 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 contained 467 documents.

[3] Other methods e.g. do not rely on semantic representation like LSA but use semantic topic models based on generative models (probabilistic inference models, topic models and probabilistic latent semantic indexing—see e.g. Wong and Yao 1995; Hofmann 1999; Blei et al. 2003).

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. Stop-word removal using the SMART stop-word list was applied before indexing, as well as stemming using the Snowball analyser (Porter stemmer). 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 propriety 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 unique 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 include both TF-IDF weighting and no weighting (using the raw term frequencies) in our model.

The literature also suggests that LSI using SVD can improve significantly the performance of the distance measures. 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 dimensionality of SVD in our research design. Normally, a set of documents is indexed and weighted as a whole, and LSI 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 LSI: index all documents of all academic inventors together and perform weighting and LSI on the global, unified vocabulary of all six academic inventors, or index the documents separately for each academic inventor and perform weighting and LSI on the case-specific vocabulary of the respective academic inventor. The individual or case-specific approach holds the promise that the weighting and LSI 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; hence, only low k-values can be used for the SVD rank k approximation.

We will include both the global unified vocabulary weighting and LSI and local case-specific weighting and LSI 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 et al. 2005a, b). For the global unified vocabulary approach, the highest rank k approximation possible is 467 (the total number of documents of 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 LSI 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). Table 1 contains an overview of the obtained measures.

SVD 0 means no LSI has been performed, while SVD 30 means that an SVD rank 30 approximation is used for LSI (only relevant for local case document indexing), being the number of documents of the academic inventor.

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 1 lists 24 combinations, only 23 measures are in fact implied. Indeed, when neither weighting nor LSI are applied, global unified document indexing and individual case document indexing yield the same distance measures for the set of relevant documents.

All distances between all seed patents and all target publications were calculated using these different distance measure variants. It should be remembered that we are only interested in the relation between patents and publications within the separate document sets of the respective academic inventors. Hence, we only calculate the distance between the patents of an academic inventor and all publications of the same academic inventors, and not the distance between patents and publications of different academic inventors. In practice, this means that we have 23 different distance measure calculations for 2,345 different patent-publication pairs.

We deliberately decided not to apply more pre-processing tasks, like compound term and collocation detection, because we wanted to keep the processing simple and

Table 1 Overview of measures

Measure	Index	Weighting	SVD	Measure	Index	Weighting	SVD
1	U	NO	0	15	C	N0	0
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	20	C	TI	0
7	U	NO	300	21	C	TI	5
8	U	TI	0	22	C	TI	10
9	U	TI	5	23	C	TI	20
10	U	TI	10	24	C	TI	30
11	U	TI	20				
12	U	TI	30				
13	U	TI	100				
14	U	TI	300				

automated. These more advanced pre-processing tasks almost always imply more human involvement and manual attention. In this setting, we wanted to try out if a simple automatic approach would work.

Comparative analysis of distance measures

Differences in measure characteristics

An overview of the obtained descriptive statistics of all measures can be found in Appendix 1. It is clear that one group of measures displays a highly skewed distribution (measures with no or high values of SVD), while other measures are far less skewed (measures with low values of SVD). Figure 1 contains distribution examples of some representative measures for all patent-publication and patent–patent pairs.

M1 is the measure with neither weighting nor SVD; M3 is a measure with no weighting and low SVD (SVD 10) performed on the global unified document set; M13 is a measure with weighting and medium SVD (SVD 100) performed on the global unified document set; and M24 is a measure with weighting and medium SVD (SVD 30) performed on the local case document set. The lines represent the number of patent–publication and patent–patent pairs having distances within the range indicated on the X-axis (distance buckets of 0.1). The measure with low SVD (M3) 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. It seems that low SVD 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 inspect descriptive statistics. In a next step, we compared the calculated similarity measures with similarity ratings obtained from independent ratings.

Assessing accuracy of measures

In order to compare and asses the accuracy of the different measures, all implied patent–paper pairs have been rated independently by two researchers. We opted for two researchers for each individual case (all patent–publication pairs of all patents of one academic inventor) in order to be sure that this independent assessment was carried out in a

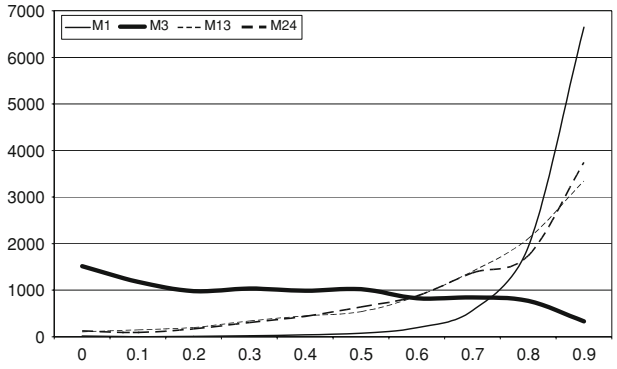

Fig. 1 Distribution examples

consistent manner. In total, five different researchers—all active and experienced in the field of science and technology studies—have been involved in this exercise for all six academic inventors.

Each researcher 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 pair were compared and Kappa scores—indicating between-subject consistency—were calculated. In the case of two assessments differing greatly (highly related versus unrelated), both assessors reviewed their assessments repeatedly, but this process did not always result in a modification of one or both scores. After this iteration, Kappa scores were obtained ranging from 0.62 to 0.90, signalling satisfactory and even excellent levels of consistency (average for the six academic inventors: 0.83).

In a next step, these assessments are used as the independent variable in an Anova analysis. 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. 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

Table 2 Accuracy levels obtained for different measures under study

Index	Weighting	SVD	Mean	Std. deviation	N
Case	NO	0	0.401	0.293	16
		5	0.247	0.257	16
		10	0.321	0.254	16
		20	0.362	0.274	16
		30	0.379	0.270	16
	TF-IDF	0	0.459 (3)	0.288	16
		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
Unified	NO	0	0.401	0.293	16
		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
	TF-IDF	0	0.489 (1)	0.301	16
		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

relatedness. Table 2 provides an overview of the average R^2 obtained for the measures under study. The higher the observed R^2, the more calculated similarity measures coincide with the independent expert assessments.

An inspection of Table 2 immediately reveals considerable differences between different measures. Measures coinciding most with independent assessment scores imply either high SVD values ($n = 300$) 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 with no SVD or 'high' levels of SVD ($n = 30$). Note that, in this case-based indexing approach, SVD values of 30 can also be considered high, given the size of case-based document sets. Differences with less performing combinations are highly significant ($p < 0.0001$). Better performing measures share the characteristic that they are relatively modest in terms of information reduction. Applying no SVD by definition implies refraining from reducing the initial word space: 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 2 reveals, SVD performs worst under all circumstances, 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 2 already illustrates the deterioration, scrutinizing specific pairs really reveals the impact of parameter choices. Appendix 2 contains the title and abstract of one patent document and two publications (co-)authored by an inventor under study. On reading these documents, it becomes apparent that one publication is 'highly related' while the other is 'unrelated'. Table 3 provides a detailed insight with respect to the distances obtained under different conditions.

Note that low values indicate similarity—with a zero value indicating complete similarity—while values approaching 1 signal no relatedness at all. As Table 3 clarifies,

Table 3 Impact of specific text mining choices on obtained measures

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 similarity measures			Obtained measures		Assessment
Index	Weighting	SVD	PUB 1 (highly related)	PUB 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

applying an SVD solution with a limited number of dimensions ($n = 5$) 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. At the same time, the two other examples included in Table 3 (unified thesaurus, SVD 300 and case-based thesaurus, SVD 30) also strongly 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 set up of a vector space model and how to proceed when calculating similarity measures, affect considerably the outcomes obtained.

Conclusions, discussion, limitations, and directions for further research

In this study, we applied and validated a set of existing 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 using a vector space model and latent semantic indexing text mining approach.

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 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. For relatively small datasets, options that respect the potential richness of the underlying data yield better results: either one opts for no SVD or SVD with 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 easier in practice. LSA seems not to redeem its promise to deal with synonymy and polysemy problems in our setting; all measures involving SVD perform worse than the ones without applying SVD. 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 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[4]). Improving accuracy levels might be feasible by further broadening the set of pre-processing options. For instance, when inspecting several patent–paper pairs, 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 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.

Acknowledgments The authors would like to express their gratitude to Julie Callaert and Mariette Du Plessis for their contribution to the independent assessment of the patent–paper pairs and Frizo Janssens for useful methodological comments and suggestions. We also wish to thank the participants of the Triple Helix Conference (Singapore, May 2007) for their helpful remarks in response to a previous version of this work, and two anonymous reviewers for their valuable comments.

Appendix 1

See Table 4.

Appendix 2: Title and abstract of one patent document and two publications (highly related and unrelated) authored by the 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 fibre bundle, demonstrating the potential to process fully biodegradable composite materials. Qualitative analysis suggests that a reasonably strong interface between the natural fibres 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 moulded 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

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

Table 4 Basic distribution descriptions of all measures

Index	Weighting	SVD	M	Mean	Std Dev	Min	Max	Range	Median	Low Q	Upp Q	Q Range	Kurt	Skew
U	NO	0	1	0.912	0.101	0.000	1.000	1.000	0.941	0.885	0.976	0.091	16.724	–3.160
U	NO	5	2	0.314	0.265	0.000	1.163	1.163	0.250	0.083	0.487	0.404	–	0.712
U	NO	10	3	0.421	0.277	0.000	1.172	1.172	0.404	0.171	0.643	0.472	–	0.217
U	NO	20	4	0.538	0.267	0.000	1.237	1.237	0.568	0.334	0.755	0.420	–	–
U	NO	30	5	0.618	0.258	0.000	1.195	1.195	0.660	0.445	0.830	0.385	–	–
U	NO	100	6	0.797	0.187	0.000	1.189	1.189	0.850	0.710	0.933	0.224	1.880	–
U	NO	300	7	0.875	0.139	0.000	1.042	1.042	0.917	0.832	0.966	0.133	7.845	–
U	TI	0	8	0.949	0.086	0.000	1.000	1.000	0.974	0.943	0.991	0.048	37.174	–
U	TI	5	9	0.124	0.134	0.000	1.179	1.179	0.083	0.037	0.167	0.130	15.002	3.050
U	TI	10	10	0.306	0.269	0.000	1.230	1.230	0.234	0.072	0.487	0.415	–	0.795
U	TI	20	11	0.448	0.295	0.000	1.241	1.241	0.438	0.174	0.697	0.523	–	0.151
U	TI	30	12	0.533	0.294	0.000	1.186	1.186	0.567	0.272	0.795	0.523	–	–
U	TI	100	13	0.775	0.220	0.000	1.364	1.364	0.838	0.675	0.938	0.263	1.083	–
U	TI	300	14	0.899	0.145	0.000	1.084	1.084	0.948	0.873	0.982	0.109	11.657	–
C	NO	5	16	0.433	0.277	0.000	1.350	1.350	0.400	0.196	0.652	0.456	–	0.316
C	NO	10	17	0.604	0.262	0.000	1.354	1.354	0.626	0.410	0.813	0.403	–	–
C	NO	20	18	0.727	0.220	0.000	1.187	1.187	0.775	0.605	0.894	0.288	0.322	–
C	NO	30	19	0.784	0.197	0.000	1.167	1.167	0.834	0.695	0.926	0.231	1.725	–
C	TI	0	20	0.960	0.077	0.000	1.000	1.000	0.982	0.957	0.993	0.035	55.526	–
C	TI	5	21	0.326	0.290	0.000	1.503	1.503	0.230	0.076	0.537	0.461	–0.369	0.818
C	TI	10	22	0.542	0.298	0.000	1.209	1.209	0.566	0.287	0.801	0.514	–	–
C	TI	20	23	0.703	0.256	0.000	1.280	1.280	0.767	0.530	0.910	0.380	–	–
C	TI	30	24	0.785	0.222	0.000	1.355	1.355	0.844	0.675	0.953	0.278	1.113	–1.169

Index Union (U); Case (C), *Weighting* No weighting (NO), TF-IDF weighting (TI), *SVD* SVD reduction, *M* Measure identification number, *Mean* Mean distance between patent and publication for all patent–publication pairs, *Std Dev* Standard deviation of distance, *Min/Max* Minimum/maximum distance, *Range* Range between minimum and maximum distance, *Median* Median, *Low Q/Upp Q* Lower/upper quartile, *Q Range* Quartile range, *Kurt*: kurtosis, *Skew* skewness

composite material comprising gluten-coated fibre, 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 moulded 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.

References

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

Azoulay, P., Ding, W., & Stuart, T. (2006). *The impact of academic patenting on the rate, quality and direction of (public) research*. NBER Working Paper No. 11917. Cambridge MA: National Bureau of Economic Research.

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* (pp. 665–694). Dordrecht: Kluwer Academic Publishers.

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

Berry, M. W. (Ed.). (2003). *Survey of text mining*. New York: Springer.
Berry, M. W., & Browne, M. (1999). *Understanding search engines: Mathematical modeling and text retrieval*. Philadelphia: Society for Industrial and Applied Mathematics.
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.
Calderini, M., Franzoni, C., & Vezzulli, A. (2005). *If star scientists do not patent: An event history analysis of scientific eminence and the decision to patent in the academic world*. CESPRI Working Paper No. 169.
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* (Vol. 31, pp. 607–649). Palo Alto, CA: Annual Reviews, Inc.
Courtial, J. P. (1994). A coword analysis of Scientometrics. *Scientometrics, 31*(3), 251–260.
Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.
Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, I*, 211–218.
Engelsman, E. C., & van Raan, A. F. J. (1994). A patent based cartography of technology. *Research Policy, 23*, 1–26.
European Commission. (2003). Third European Report on S&T Indicators.
Fabrizio, K. R., & DiMinin, A. (2005). *Commercializing the laboratory*: *Faculty patenting and the open science environment*. Working paper.
Glänzel, W., et al. (2004). *Biotechnology: An analysis of patents and publications*. Report Steunpunt O&O Statistics (www.steunpuntoos.be).
Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005a). Combining full-text and bibliometric information in mapping scientific disciplines. *Information Processing & Management, 41*(6), 1548–1572.
Glenisson, P., Glänzel, W., & Persson, O. (2005b). Combining full-text and bibliometric indicators: A pilot study. *Scientometrics, 63*(1), 163–180.
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). Hew effective is suffixing? *Journal of the American Society for Information Science, 42*, 7–15.
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.
Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the Twenty-Second Annual International SIGIR Conference* (pp. 50–57). New York: ACM Press.
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.
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*, 211–240.
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* (pp. 187–213). Dordrecht: 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.
Manning, C. D., & Schütze, H. (2000). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

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.
Meyer, M. (2006). Knowledge integrators or weak links? An exploratory comparison of patenting researchers with their non-inventing peers in nano-science and technology. *Scientometrics, 68*(3), 545–560.
Moens, M. F. (2006). *Information extraction: Algorithms and prospects in a retrieval context (The Information Retrieval Series 21)*. New York: Springer.
Murray, F. & Stern, S. (2005). *Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis*. NBER Working Paper No. 11465. Cambridge, MA: National Bureau of Economic Research.
National Science Foundation (NSF). (2006). Science and Engineering Indicators.
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* (pp. 587–612). Dordrecht: Kluwer Academic Publishers.
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* (pp. 45–72). Berlin: Springer.
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 Automatic Indexing. *Communications of the ACM, 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* (pp. 9–22). Boston: Butterworths.
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* (pp. 717–731). Dordrecht: Kluwer Academic Publishers.
Sparck Jones, K. (1971). *Automatic keyword classification for information Retrieval*. London: Buttersworth.
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., 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 Rijsbergen, C. J., Robertson, S. E., & Porter, M. F. (1980). *New models in probabilistic information retrieval*. London: British Library (British Library Research and Development Report, No. 5587).
Vandromme, D., Magerman, T., Song, X., Van Looy, B., Hoskens, M., Glenisson, P., Thijs, B., Vertomme, J., De Moor, B., & Duflou, J. (2006). *A comparative analysis of distance measures and text mining methods supporting domain studies*. Paper presented at the Ninth STI indicator conference, Leuven, 2006.
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.