Scientometrics (2010) 82:581–596
DOI 10.1007/s11192-010-0173-0

Automatic term identification for bibliometric mapping

Nees Jan van Eck · Ludo Waltman · Ed C. M. Noyons · Reindert K. Buter

Received: 11 May 2009 / Published online: 11 February 2010

Abstract A term map is a map that visualizes the structure of a scientific field by showing the relations between important terms in the field. The terms shown in a term map are usually selected manually with the help of domain experts. Manual term selection has the disadvantages of being subjective and labor-intensive. To overcome these disadvantages, we propose a methodology for automatic term identification and we use this methodology to select the terms to be included in a term map. To evaluate the proposed methodology, we use it to construct a term map of the field of operations research. The quality of the map is assessed by a number of operations research experts. It turns out that in general the proposed methodology performs quite well.

Keywords Bibliometric mapping · Term map · Automatic term identification · Probabilistic latent semantic analysis · Operations research

Introduction

Bibliometric mapping is a powerful tool for studying the structure and the dynamics of scientific fields. Researchers can utilize bibliometric maps to obtain a better understanding of the field in which they are working. In addition, bibliometric maps can provide valuable insights for science policy purposes (Noyons 1999, 2004).

Various types of bibliometric maps can be distinguished, which each visualize the structure of a scientific field from a different point of view. Some maps, for example, show relations between authors or journals based on co-citation data. Other maps show relations between words or keywords based on co-occurrence data (e.g., Rip and Courtial 1984; Peters and Van Raan 1993; Kopcsa and Schiebel 1998; Noyons 1999; Ding et al. 2001).

N. J. van Eck · L. Waltman
Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam,
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

N. J. van Eck (✉) · L. Waltman · E. C. M. Noyons · R. K. Buter
Centre for Science and Technology Studies, Leiden University, Leiden, The Netherlands
e-mail: ecknjpvan@cwts.leidenuniv.nl

The latter maps are usually referred to as co-word maps. In this paper, we are concerned with maps that show relations between terms. We refer to these maps as term maps. By a term we mean a word or a phrase that refers to a domain-specific concept. Term maps are similar to co-word maps except that they may contain any type of term instead of only single-word terms or only keywords.

When constructing a bibliometric map, one-first has to select the objects to be included in the map. In the case of a map that contains authors or journals, this is usually fairly easy. To select the important authors or journals in a field, one can usually simply rely on citation counts. In the case of a term map, things are not so easy. In most cases, it is quite difficult to select the important terms in a field. Selection of terms based on their frequency of occurrence in a corpus of documents typically yields many words and phrases with little or no domain-specific meaning. Inclusion of such words and phrases in a term map is highly undesirable for two reasons. First, these words and phrases divert attention from what is really important in the map. Second and even more problematic, these words and phrases may distort the entire structure shown in the map. Because there is no easy way to select the terms to be included in a term map, term selection is usually done manually based on expert judgment (e.g., Noyons 1999; Van Eck and Waltman 2007b). However, manual term selection has serious disadvantages as well. The most important disadvantage is that it involves a lot of subjectivity, which may introduce significant biases in a term map. Another disadvantage is that it can be very labor-intensive.

In this paper, we try to overcome the problems associated with manual selection of the terms to be included in a term map. To do so, we propose a methodology that aims to automatically identify the terms that occur in a corpus of documents. Term selection using the proposed methodology requires less involvement of domain experts than manual term selection. Consequently, we expect term maps constructed using the proposed methodology to be more objective representations of scientific fields. An additional advantage of the proposed methodology is that it makes the process of term selection less labor-intensive.

The general idea of the methodology that we propose can be explained briefly as follows. Given a corpus of documents, we first identify the main topics in the corpus. This is done using a technique called probabilistic latent semantic analysis (Hofmann 2001). Given the main topics, we then identify in the corpus the words and phrases that are strongly associated with only one or only a few topics. These words and phrases are selected as the terms to be included in a term map. An important property of the proposed methodology is that it identifies terms that are not only domain-specific but that also have a high discriminatory power within the domain of interest. This is important because terms with a high discriminatory power are essential for visualizing the structure of a scientific field. Suppose, for example, that we want to construct a term map of the field of statistics. *sample* and *chi-square test* are both statistical terms. However, *sample* is a quite general statistical term, while *chi-square test* is more specific and, consequently, more discriminatory. Because of the relatively high discriminatory power of *chi-square test*, inclusion of this term in a term map may help to reveal the structure of the field of statistics. Inclusion of *sample*, on the other hand, probably does not provide much additional insight into the structure of the field. Hence, to visualize the structure of a scientific field, terms with a high discriminatory power play an essential role.

The organization of this paper is as follows. We first provide a brief overview of the literature on automatic term identification. After discussing the literature, we propose a new methodology for automatic term identification. We then experimentally evaluate the proposed methodology, focusing in particular on its performance in the context of bibliometric mapping. Evaluation is done by applying the proposed methodology to the field

of operations research and by asking a number of experts in this field to assess the results that are obtained. We end this paper with a discussion of the conclusions of our research.

Overview of the automatic term identification literature

In this section, we provide a brief overview of the literature on automatic term identification (also known as automatic term recognition or automatic term extraction).[1] For extensive reviews of the literature, we refer to Kageura and Umino (1996), Cabré Castellví et al. (2001), Jacquemin (2001), and Pazienza et al. (2005). We note that there are almost no studies on automatic term identification in the context of bibliometric mapping. Exceptions are the work of Janssens et al. (2006), Noyons (1999), and Schneider (2006), in which automatic term identification receives some attention. In the literature discussed in the rest of this section, automatic term identification is studied for purposes other than bibliometric mapping.

We first discuss the notions of unithood and termhood (for the original definitions of these notions, see Kageura and Umino 1996). We define unithood as the degree to which a phrase constitutes a semantic unit. Our idea of a semantic unit is similar to that of a collocation (Manning and Schütze 1999). Hence, a semantic unit is a phrase consisting of words that are conventionally used together. The meaning of the phrase typically cannot be fully predicted from the meaning of the individual words within the phrase. We define termhood as the degree to which a semantic unit represents a domain-specific concept. A semantic unit with a high termhood is a term. To illustrate the notions of unithood and termhood, suppose that we are interested in statistical terms. Consider the phrases *many countries*, *United States*, and *probability density function*. Clearly, *United States* and *probability density function* are semantic units, while *many countries* is not. Hence, the unithood of *United States* and *probability density function* is high, while the unithood of *many countries* is low. Because *United States* does not represent a statistical concept, it has a low termhood. *probability density function*, on the other hand, does represent a statistical concept and therefore has a high termhood. From this it follows that *probability density function* is a statistical term.

In the literature, two types of approaches to automatic term identification are distinguished, linguistic approaches and statistical approaches. Linguistic approaches are mainly used to identify phrases that, based on their syntactic form, can serve as candidate terms. Statistical approaches are used to measure the unithood and termhood of phrases. In many cases, linguistic and statistical approaches are combined in a single hybrid approach.

Most terms have the syntactic form of a noun phrase (Justeson and Katz 1995; Kageura and Umino 1996). Linguistic approaches to automatic term identification typically rely on this property. These approaches identify candidate terms using a linguistic filter that checks whether a sequence of words conforms to some syntactic pattern. Different researchers use different syntactic patterns for their linguistic filters (e.g., Bourigault 1992; Dagan and Church 1994; Daille et al. 1994; Justeson and Katz 1995; Frantzi et al. 2000).

Statistical approaches to measure unithood are discussed extensively by Manning and Schütze (1999). The simplest approach uses frequency of occurrence as a measure of unithood (e.g., Dagan and Church 1994; Daille et al. 1994; Justeson and Katz 1995). More advanced approaches use measures based on, for example, (pointwise) mutual information

[1] A more elaborate overview of the literature can be found in an earlier version of this paper (Van Eck et al. 2008).

(e.g., Church and Hanks 1990; Damerau 1993; Daille et al. 1994) or a likelihood ratio (e.g., Dunning 1993; Daille et al. 1994). Another statistical approach to measure unithood is the C-value (Frantzi et al. 2000). The NC-value (Frantzi et al. 2000) and the SNC-value (Maynard and Ananiadou 2000) are extensions of the C-value that measure not only unithood but also termhood. Other statistical approaches to measure termhood can be found in the work of, for example, Drouin (2003) and Matsuo and Ishizuka (2004). In the field of machine learning, an interesting statistical approach to measure both unithood and termhood is proposed by Wang et al. (2007).

Methodology

In this section, we propose a three-step methodology for automatic term identification. An overview of the proposed methodology is provided in Fig. 1. Consider some domain or some scientific field, and suppose that we want to identify terms that belong specifically to this domain or this field. Our methodology assumes the availability of a corpus that is partitioned into a number of segments, each of which is concerned with a particular topic or a particular combination of topics within the domain of interest. Such a corpus may for example consist of a large number of documents or abstracts. In the first step of our methodology, a linguistic filter is applied to the corpus in order to identify noun phrases. In the second step, the unithood of noun phrases is measured in order to identify semantic units. In the third and final step, the termhood of semantic units is measured in order to identify terms. Termhood is measured as the degree to which the occurrences of a semantic unit are biased towards one or more topics. Compared with alternative approaches to automatic term identification, such as the ones discussed in the previous section, the innovative aspect of our methodology mainly lies in the third step, that is, in the measurement of termhood. We now discuss the three steps of our methodology in more detail.

Step 1: Linguistic filter

In the first step of our methodology, we use a linguistic filter to identify noun phrases. We first assign to each word occurrence in the corpus a part-of-speech tag, such as noun, verb, or adjective. The appropriate part-of-speech tag for a word occurrence is determined using a part-of-speech tagger developed by Schmid (1994, 1995). We use this tagger because it has a good performance and because it is freely available for research purposes.[2] In addition to a part-of-speech tag, the tagger also assigns a so-called lemma to each word occurrence in the corpus. The lemma assigned to a word occurrence is the root form (or the stem) of the word. The words *function* and *functions*, for example, both have *function* as their lemma. In all further stages of our methodology, we use the lemmatized corpus instead of the original corpus. In this way, differences between, for example, uppercase and lowercase letters and singular and plural nouns are ignored.

After the corpus has been tagged and lemmatized, we apply a linguistic filter to it. The filter that we use identifies all word sequences that meet the following three criteria:

1. The sequence consists of nouns and adjectives only.
2. The sequence ends with a noun.

[2] See http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

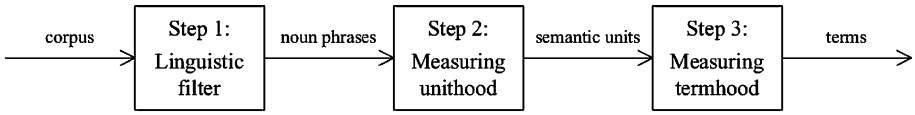

Fig. 1 Overview of the proposed methodology

3. The sequence occurs at least a certain number of times in the corpus (ten times in the experiment discussed later on in this paper).

Assuming an English language corpus, the first two criteria ensure that all identified word sequences are noun phrases. Notice, however, that our filter does not identify all types of noun phrases. Noun phrases that contain a preposition, such as the phrase *degree of freedom*, are not identified (for a discussion of such noun phrases, see Justeson and Katz 1995). We emphasize that the choice of an appropriate linguistic filter depends on the language of the corpus. The filter that we use works well for the English language but may not be appropriate for other languages. For all noun phrases that are identified by our linguistic filter, the unithood is considered in the second step of our methodology.

Step 2: Measuring unithood

In the second step of our methodology, we measure the unithood of noun phrases. Unithood is only relevant for noun phrases consisting of more than one word. For such noun phrases, unithood determines whether they are regarded as semantic units. The main aim of the second step of our methodology is to get rid of noun phrases that start with uninteresting adjectives such as *first*, *many*, *new*, and *some*.

The most common approach to measure unithood is to determine whether a phrase occurs more frequently than would be expected based on the frequency of occurrence of the individual words within the phrase. This is basically also the approach that we take. To measure the unithood of a noun phrase, we first count the number of occurrences of the phrase, the number of occurrences of the phrase without the first word, and the number of occurrences of the first word of the phrase. In a similar way as Dunning (1993), we then use a so-called likelihood ratio to compare the first number with the last two numbers. We interpret this likelihood ratio as a measure of the unithood of the phrase. In the end, we use a cutoff value to determine which noun phrases are regarded as semantic units and which are not. (In the experiment discussed later on in this paper, noun phrases are regarded as semantic units if the natural logarithm of their likelihood ratio is less than -30.) For all noun phrases that are regarded as semantic units (which includes all single-word noun phrases), the termhood is considered in the third step of our methodology.

Step 3: Measuring termhood

In the third step of our methodology, we measure the termhood of semantic units. As mentioned earlier, we assume that we have a corpus that is partitioned into a number of segments, each of which is concerned with a particular topic or a particular combination of topics within the domain of interest. A corpus segment may for example consist of a document or an abstract, or it may consist of the set of all documents or all abstracts that appeared in a journal during a certain period of time. We use the following mathematical notation. There are K semantic units of which we want to measure the termhood. These units are denoted by $u_1,\ldots,u_K$. The corpus is partitioned into I segments, which are denoted

by $s_1,\ldots,s_I$. The number of occurrences of semantic unit u_k in corpus segment s_i is denoted by n_{ik}. Finally, there are J topics to be distinguished. These topics are denoted by $t_1,\ldots,t_J$.

The main idea of the third step of our methodology is to measure the termhood of a semantic unit as the degree to which the occurrences of the unit are biased towards one or more topics. We first discuss an approach that implements this idea in a very simple way. We assume that there is a one-to-one relationship between corpus segments and topics, that is, each corpus segment covers exactly one topic and each topic is covered by exactly one corpus segment. Under this assumption, the number of corpus segments equals the number of topics, so $I = J$. To measure the degree to which the occurrences of semantic unit u_k, where $k \in \{1,\ldots,K\}$, are biased towards one or more topics, we use two probability distributions, namely the distribution of semantic unit u_k over the set of all topics and the distribution of all semantic units together over the set of all topics. These distributions are denoted by, respectively, $P(t_j \mid u_k)$ and $P(t_j)$, where $j \in \{1,\ldots, J\}$. Assuming that topic t_j is covered by corpus segment s_j, the distributions are given by

$$P(t_j|u_k) = \frac{n_{jk}}{\sum_{j'=1}^{J} n_{j'k}} \tag{1}$$

and

$$P(t_j) = \frac{\sum_{k=1}^{K} n_{jk}}{\sum_{j'=1}^{J} \sum_{k=1}^{K} n_{j'k}}. \tag{2}$$

The dissimilarity between the two distributions indicates the degree to which the occurrences of u_k are biased towards one or more topics. We use the dissimilarity between the two distributions to measure the termhood of u_k. For example, if the two distributions are identical, the occurrences of u_k are unbiased and u_k most probably does not represent a domain-specific concept. If, on the other hand, the two distributions are very dissimilar, the occurrences of u_k are strongly biased and u_k is very likely to represent a domain-specific concept. The dissimilarity between two probability distributions can be measured in many different ways. One may use, for example, the Kullback–Leibler divergence, the Jensen–Shannon divergence, or a chi-square value. We use a somewhat different measure. Based on this measure, the termhood of u_k is calculated as

$$\text{termhood}\,(u_k) = \sum_{j=1}^{J} p_j \log p_j, \tag{3}$$

where 0 log 0 is defined as 0 and where

$$p_j = \frac{P(t_j|u_k)/P(t_j)}{\sum_{j'=1}^{J} P(t_{j'}|u_k)/P(t_{j'})}. \tag{4}$$

It follows from (4) that $p_1,\ldots,p_J$ define a probability distribution over the set of all topics. In (3), termhood (u_k) is calculated as the negative entropy of this distribution. Notice that termhood (u_k) is maximal if $P(t_j \mid u_k) = 1$ for some j and that it is minimal if $P(t_j \mid u_k) = P(t_j)$ for all j. In other words, termhood (u_k) is maximal if the occurrences of u_k are completely biased towards a single topic, and termhood (u_k) is minimal if the occurrences of u_k do not have a bias towards any topic.

The approach discussed above relies on the assumption of a one-to-one relationship between corpus segments and topics. For most corpora, this assumption is probably not very realistic. For example, if each segment of a corpus consists of a single document or a

single abstract, there will most likely be some segments that are concerned with more or less the same topic. Or the other way around, if each segment of a corpus consists of a set of documents or abstracts that all appeared in the same journal, there will most likely be some segments (particularly segments corresponding to multidisciplinary journals) that are concerned with more than one topic. Below, we extend our approach in such a way that it no longer relies on the assumption of a one-to-one relationship between corpus segments and topics.

Identifying topics

In order to allow for a many-to-many relationship between corpus segments and topics, we make use of probabilistic latent semantic analysis (PLSA) (Hofmann 2001). PLSA is a quite popular technique in machine learning, information retrieval, and related fields. It was originally introduced as a probabilistic model that relates occurrences of words in documents to so-called latent classes. In the present context, we are dealing with semantic units and corpus segments instead of words and documents, and we interpret the latent classes as topics.

When using PLSA, we first have to determine an appropriate value for the number of topics J. This value is typically much smaller than both the number of corpus segments I and the number of semantic units K. In this paper, we manually choose a value for J. PLSA assumes that each occurrence of a semantic unit in a corpus segment is independently generated according to the following probabilistic process. First, a topic t is drawn from a probability distribution $P(t_j)$, where $j \in \{1,\ldots,J\}$. Next, given t, a corpus segment s and a semantic unit u are independently drawn from, respectively, the conditional probability distributions $P(s_i \mid t)$, where $i \in \{1,\ldots,I\}$, and $P(u_k \mid t)$, where $k \in \{1,\ldots,K\}$. This then results in the occurrence of u in s. It is clear that, according to the generative process assumed by PLSA, the probability of generating an occurrence of semantic unit u_k in corpus segment s_i equals

$$P(s_i, u_k) = \sum_{j=1}^{J} P(t_j) P(s_i \mid t_j) P(u_k \mid t_j). \tag{5}$$

The probabilities $P(t_j)$, $P(s_i \mid t_j)$, and $P(u_k \mid t_j)$, for $i = 1,\ldots,I$, $j = 1,\ldots,J$, and $k = 1,\ldots,K$, are the parameters of PLSA. We estimate these parameters using data from the corpus. Estimation is based on the criterion of maximum likelihood. The log-likelihood function to be maximized is given by

$$L = \sum_{i=1}^{I} \sum_{k=1}^{K} n_{ik} \log P(s_i, u_k). \tag{6}$$

We use the EM algorithm discussed by Hofmann (1999, Sect. 3.2) to perform the maximization of this function.[3] After estimating the parameters of PLSA, we apply Bayes' theorem to obtain a probability distribution over the topics conditional on a semantic unit. This distribution is given by

[3] A MATLAB implementation of this algorithm is available on request.

$$P(t_j \mid u_k) = \frac{P(t_j)P(u_k \mid t_j)}{\sum_{j'=1}^{J} P(t_{j'})P(u_k \mid t_{j'})}. \quad (7)$$

In a similar way as discussed earlier, we use the dissimilarity between the distributions $P(t_j \mid u_k)$ and $P(t_j)$ to measure the termhood of u_k. In this case, however, $P(t_j \mid u_k)$ is given by (7) instead of (1) and $P(t_j)$ follows from the estimated parameters of PLSA instead of being given by (2). We again use (3) and (4) to calculate the termhood of u_k.

Experimental evaluation

In this section, we experimentally evaluate our methodology for automatic term identification. We focus in particular on the performance of our methodology in the context of bibliometric mapping.

Application to the field of operations research

We apply our methodology to the field of operations research (OR), also known as operational research. The OR field was chosen because some of us have some background in this field and because we have easy access to a number of OR experts who can help us with the evaluation of our results. We note that sometimes a distinction is made between OR on the one hand and management science on the other hand (e.g., Eto 2000, 2002). For our purpose, however, such a distinction is not important. In this paper, the term OR therefore also includes management science.

We start with a discussion of how we put together our corpus. We first selected a number of OR journals. This was done based on the subject categories of Thomson Reuters. The OR field is covered by the category *Operations Research & Management Science*. Since we wanted to focus on the core of the field, we selected only a subset of the journals in this category. More specifically, a journal was selected if it belongs to the category *Operations Research & Management Science* and possibly also to the closely related category *Management* and if it does not belong to any other category. This yielded 15 journals, which are listed in the first column of Table 1. We used the database of the Centre for Science and Technology Studies, which is similar to the Web of Science database of Thomson Reuters, to retrieve all documents, except those without an abstract, that were published in the selected journals between 2001 and 2006. For each journal, the number of documents retrieved from the database is reported in the second column of Table 1. Of each of the documents retrieved, we included the title and the abstract in our corpus.

After putting together the corpus, we applied our methodology for automatic term identification. In the first step of our methodology, the linguistic filter identified 2662 different noun phrases. In the second step, the unithood of these noun phrases was measured. 203 noun phrases turned out to have a rather low unithood and therefore could not be regarded as semantic units. Examples of such noun phrases are *first problem*, *good use*, and *optimal cost*. The other 2459 noun phrases had a sufficiently high unithood to be regarded as semantic units. In the third and final step of our methodology, the termhood of these semantic units was measured. To do so, each title-abstract pair in the corpus was treated as a separate corpus segment. For each combination of a semantic unit u_k and a corpus segment s_i, it was determined whether u_k occurs in s_i ($n_{ik} = 1$) or not ($n_{ik} = 0$). Topics were identified using PLSA. This required the choice of the number of topics J. Results for

Table 1 Overview of the selected journals

Journal	Number of documents	Coverage (%)
European Journal of Operational Research	2705	97.2
Journal of the Operational Research Society	830	96.9
Management Science	726	98.9
Annals of Operations Research	679	95.3
Operations Research Letters	458	93.0
Operations Research	439	97.7
Naval Research Logistics	327	98.5
Omega-International Journal of Management Science	277	97.1
Interfaces	257	98.4
Journal of Operations Management	211	98.1
Journal of the Operations Research Society of Japan	158	96.8
Asia–Pacific Journal of Operational Research	140	99.3
OR Spectrum	140	97.9
RAIRO-Operations Research	92	93.5
Military Operations Research	53	98.1
Total	7492	97.0

various numbers of topics were examined and compared. Based on our own knowledge of the OR field, we decided to work with $J = 10$ topics. The output of our methodology consisted of a list of 2459 semantic units together with their termhood values. For the interested reader, this list is available online.[4]

Evaluation based on precision and recall

The evaluation of a methodology for automatic term identification is a difficult issue. There is no generally accepted standard for how evaluation should be done. We refer to Pazienza et al. (2005) for a discussion of the various problems. In this paper, we evaluate our methodology in two ways. We first perform an evaluation based on the well-known notions of precision and recall. We then perform a second evaluation by constructing a term map and asking experts to assess the quality of this map. Since our methodology for automatic term identification is intended to be used for bibliometric mapping purposes, we are especially interested in the results of the second evaluation.

We first discuss the evaluation of our methodology based on precision and recall. The main aim of this evaluation is to compare the performance of our methodology with the performance of two simple alternatives. One alternative is a variant of our methodology. This variant assumes a one-to-one relationship between corpus segments and topics, and it therefore does not make use of PLSA. The other alternative is a very simple one. It uses frequency of occurrence as a measure of termhood.

In the context of automatic term identification, precision and recall are defined as follows. Precision is the number of correctly identified terms divided by the total number of identified terms. Recall is the number of correctly identified terms divided by the total number of correct terms. Unfortunately, because the total number of correct terms in the

[4] See http://www.neesjanvaneck.nl/term_identification/.

OR field is unknown, we could not calculate the true recall. This is a well-known problem in the context of automatic term identification (Pazienza et al. 2005). To circumvent this problem, we defined recall in a slightly different way, namely as the number of correctly identified terms divided by the total number of correct terms within the set of all semantic units identified in the second step of our methodology. Recall calculated according to this definition provides an upper bound on the true recall. However, even using this definition of recall, the calculation of precision and recall remained problematic. The problem was that it is very time-consuming to manually determine which of the 2459 semantic units identified in the second step of our methodology are correct terms and which are not. We solved this problem by estimating precision and recall based on a random sample of 250 semantic units. The first two authors of this paper, who both have some knowledge of the OR field, independently determined which of these 250 units are correct terms and which are not. Units on which the authors did not agree were discussed until agreement was reached.

To identify terms, we used a cutoff value that determined which semantic units were regarded as terms and which were not. Semantic units were regarded as terms if their termhood value was greater than the cutoff value. Obviously, a lower cutoff value leads to a larger number of identified terms and, consequently, to a higher recall. However, a lower cutoff value usually also leads to a lower precision. Hence, there is a trade-off between precision and recall. By varying the cutoff value, the relation between precision and recall can be obtained. In Fig. 2, the graphs labeled *PLSA* and *No PLSA* show this relation for, respectively, our methodology and the variant of our methodology that does not make use of PLSA. The third graph in the figure shows the relation between precision and recall for the approach based on frequency of occurrence. It is clear from the figure that our methodology outperforms the two simple alternatives. Except for very low and very high levels of recall, our methodology always has a considerably higher precision than the variant of our methodology that does not make use of PLSA. The low precision of our methodology for very low levels of recall is based on a very small number of incorrectly identified terms and is therefore insignificant from a statistical point of view. The approach based on frequency of occurrence has a very bad performance. For almost all levels of

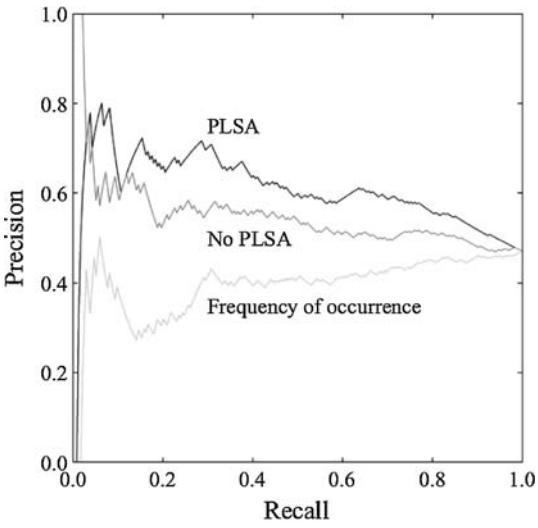

Fig. 2 The relationship between precision and recall for our methodology and for two simple alternatives

recall, the precision of this approach is even lower than the precision that would have been obtained if terms had been identified at random. Unfortunately, there is no easy way to compare the precision/recall performance of our methodology with that of other approaches proposed in the literature. This is due to the lack of a generally accepted evaluation standard (Pazienza et al. 2005). We refer to Cabré Castellví et al. (2001) for an overview of some precision/recall results reported for other approaches.

Evaluation using a term map

We now discuss the second evaluation of our methodology for automatic term identification. This evaluation is performed using a term map. The evaluation therefore focuses specifically on the usefulness of our methodology for bibliometric mapping purposes.

A term map is a map, usually in two dimensions, that shows the relations between important terms in a scientific field. Terms are located in a term map in such a way that the proximity of two terms reflects their relatedness as closely as possible. That is, the smaller the distance between two terms, the stronger their relation. The aim of a term map usually is to visualize the structure of a scientific field.

In order to evaluate our methodology, we constructed a term map of the OR field. The terms to be included in the map were selected based on the output of our methodology. It turned out that, out of the 2459 semantic units identified in the second step of our methodology, 831 had the highest possible termhood value. This means that, according to our methodology, 831 semantic units are associated exclusively with a single topic within the OR field. We decided to select these 831 semantic units as the terms to be included in the term map. This yielded a coverage of 97.0%, which means that 97.0% of the title-abstract pairs in the corpus contain at least one of the 831 terms to be included in the term map. The coverage per journal is reported in the third column of Table 1.

The term map of the OR field was constructed using a procedure similar to the one used in our earlier work (Van Eck and Waltman 2007b). This procedure relies on the association strength measure (Van Eck and Waltman 2009) to determine the relatedness of two terms, and it uses the VOS technique (Van Eck and Waltman 2007a) to determine the locations of terms in the map. Due to the large number of terms, the map that was obtained cannot be shown in this paper. However, a simplified version of the map is presented in Fig. 3. This version of the map only shows terms that do not overlap with other more important terms. The complete map showing all 831 terms is available online.[5] A special computer program called VOSviewer (Van Eck and Waltman in press) allows the map to be examined in full detail. VOSviewer uses colors to indicate the different topics that were identified using PLSA.

The quality of the term map of the OR field was assessed by five experts. Two of them are assistant professor of OR, one is associate professor of OR, and two are full professor of OR. All experts are working at Erasmus University Rotterdam. We asked each expert to examine the online term map and to complete a questionnaire. The questionnaire consisted of one multiple-choice question and ten open-ended questions. The main results of the questionnaire are discussed below. The full results are available on request.

In the multiple-choice question, we asked the experts to indicate on a five-point scale how well the term map visualizes the structure of the OR field. Four experts answered that the map visualizes the structure of the field quite well (the second highest answer on the five-point scale). The fifth expert answered that the map visualizes the structure of the field

[5] See http://www.neesjanvaneck.nl/term_identification/.

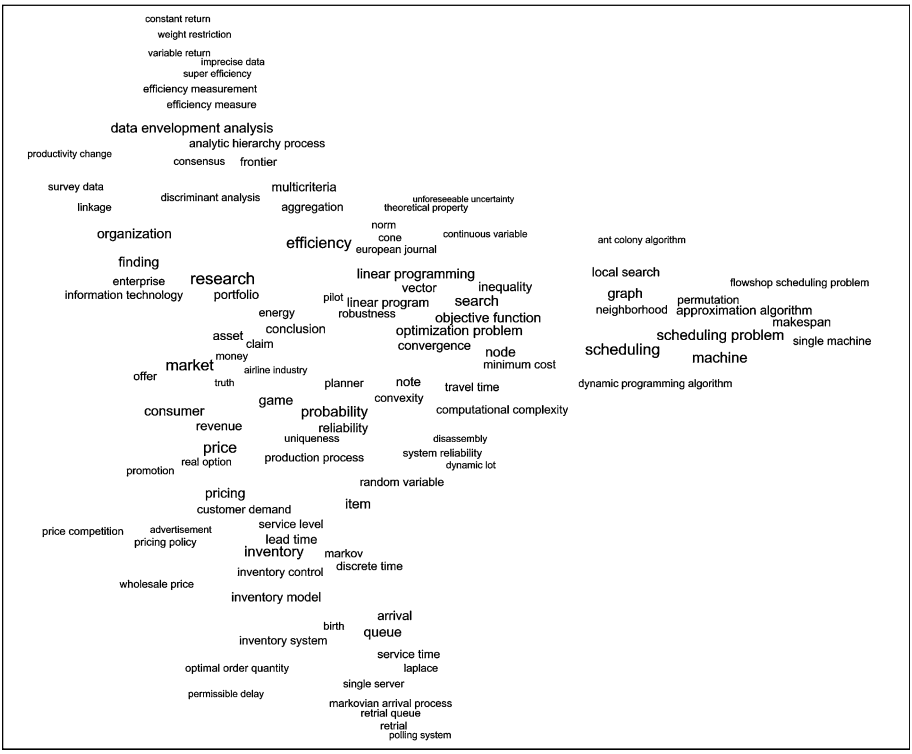

Fig. 3 Simplified version of the term map of the OR field

very well (the highest answer on the five-point scale). Hence, overall the experts were quite satisfied with the map. The experts could also easily explain the global structure of the map, and for them the topics shown in the map (indicated using colors) generally had an obvious interpretation. We also asked the experts whether the map showed anything unexpected to them. One expert answered that he had not expected scheduling related terms to be located at the boundary of the map. Two other experts turned out to be surprised by the prominent position of economics related terms such as *consumer*, *price*, *pricing*, and *revenue*. None of these three experts regarded the unexpected results as a weakness of the map. Instead, two experts stated that their own perception of their field may not have been correct. Hence, it seems that these experts may have learned something new from the map.

The experts also indicated some weak points of the term map. Some of these points were related to the way in which the terms shown in the map were selected. Other points were of a more general nature. The most serious criticism on the results of the automatic term identification concerned the presence of a number of rather general terms in the map. Examples of such terms are *claim*, *conclusion*, *finding*, *item*, and *research*. There were three experts who criticized the presence of terms such as these. We agree with these experts that some of the terms shown in the map are too general. Although the number of such terms is not very large, we consider it is highly desirable to get rid of them. To achieve this, further improvement of our methodology for automatic term identification is necessary. We will come back to this below.

Another point of criticism concerned the underrepresentation of certain topics in the term map. There were three experts who raised this issue. One expert felt that the topic of supply chain management is underrepresented in the map. Another expert stated that he had expected the topic of transportation to be more visible. The third expert believed that the topics of combinatorial optimization, revenue management, and transportation are underrepresented. It seems likely that in many cases the perceived underrepresentation of topics was not due to our methodology for automatic term identification but was instead caused by the way in which the corpus used by our methodology was put together. As discussed earlier, when we were putting together the corpus, we wanted to focus on the core of the OR field and we therefore only included documents from a relatively small number of journals. This may for example explain why the topic of transportation is not clearly visible in the map. Thomson Reuters has a subject category *Transportation Science & Technology*, and it may well be that much transportation related OR studies are published in journals that belong to this category (and possibly also to the category *Operations Research & Management Science*). The corpus that we put together does not cover these journals and hence may contain only a small portion of the transportation related OR studies. It is then not surprising that the topic of transportation is difficult to see in the map.

The remaining issues raised by the experts are of a more general nature, and most likely these issues would also have been raised if the terms shown in the term map had been selected manually. One of the issues had to do with the character of the OR field. When asked to divide the OR field into a number of smaller subfields, most experts indicated that there are two natural ways to make such a division. On the one hand, a division can be made based on the methodology that is being used, such as decision theory, game theory, mathematical programming, or stochastic modeling. On the other hand, a division can be made based on the area of application, such as inventory control, production planning, supply chain management, or transportation. There were two experts who noted that the term map seems to mix up both divisions of the OR field. According to these experts, one part of the map is based on the methodology-oriented division of the field, while the other part is based on the application-oriented division. One of the experts stated that he would be interested to see an explicit separation of the methodology and application dimensions.

A final issue, which was raised by two experts, had to do with the more detailed interpretation of the term map. The experts pointed out that sometimes closely related terms are not located very close to each other in the map. One of the experts gave the terms *inventory* and *inventory cost* as an example of this problem. In many cases, a problem such as this is probably caused by the limited size of the corpus that was used to construct the map. In other cases, the problem may be due to the inherent limitations of a two-dimensional representation. The best solution to this kind of problems seems to be not to show individual terms in a map but to only show topics (e.g., Noyons and Van Raan 1998; Noyons 1999). Topics can then be labeled using one or more representative terms.

Conclusions

In this paper, we have addressed the question how the terms shown in a term map can be selected without relying extensively on the judgment of domain experts. Our main contribution consists of a methodology for automatic identification of terms in a corpus of documents. Using this methodology, the process of selecting the terms to be included in a term map can be automated for a large part, thereby making the process less labor-intensive and less dependent on expert judgment. Because less expert judgment is required,

the process of term selection also involves less subjectivity. We therefore expect term maps constructed using our methodology to be more objective representations of scientific fields.

We have evaluated our methodology for automatic term identification by applying it to the OR field. In general, we are quite satisfied with the results that we have obtained. The precision/recall results clearly indicate that our methodology outperformed two simple alternatives. In addition, the quality of the term map of the OR field constructed using our methodology was assessed quite positively by five experts in the field. However, the term map also revealed a shortcoming of our methodology, namely the incorrect identification of a number of general noun phrases as terms. We hope to remedy this shortcoming in future work.

Finally, we would like to place the research presented in this paper in a broader perspective. As scientific fields tend to overlap more and more and disciplinary boundaries become more and more blurred, finding an expert who has a good overview of an entire domain becomes more and more difficult. This poses serious difficulties for any bibliometric method that relies on expert knowledge. Term mapping is one such method. Fortunately, advanced computational techniques from fields such as data mining, machine learning, statistics, and text mining may be used to take over certain tasks in bibliometric analysis that are traditionally performed by domain experts (for an overview of various computational techniques, see Leopold et al. 2004). The research presented in this paper can be seen as an elaboration of this idea in the context of term mapping. We acknowledge, however, that our research is only a first step towards fully automatic term mapping. To produce accurate term maps, the output of our methodology for automatic term identification still needs to be verified manually and some amount of expert knowledge is still required. In future work, we intend to take even more advantage of the possibilities offered by various kinds of computational techniques. Hopefully, this allows the dependence of term mapping on expert knowledge to be reduced even further.

Acknowledgements We thank Rommert Dekker, Moritz Fleischmann, Dennis Huisman, Wilco van den Heuvel, and Albert Wagelmans for their help with the evaluation of the term map of the OR field.

References

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In: *Proceedings of the 14th conference on computational linguistics* (pp. 977–981). Morristown, NJ: Association for Computational Linguistics.

Cabré Castellví, M. T., Estopà Bagot, R., & Vivaldi Palatresi, J. (2001). Automatic term detection: A review of current systems. In: D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Eds.), *Recent advances in computational terminology* (pp. 53–87). Amsterdam: John Benjamins.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*(1), 22–29.

Dagan, I., & Church, K. (1994). TERMIGHT: Identifying and translating technical terminology. In: *Proceedings of the 4th conference on applied natural language processing* (pp. 34–40). Morristown, NJ: Association for Computational Linguistics.

Daille, B., Gaussier, É., & Langé, J.-M. (1994), Towards automatic extraction of monolingual and bilingual terminology. In: *Proceedings of the 15th conference on computational linguistics* (pp. 515–521). Morristown, NJ: Association for Computational Linguistics.

Damerau, F. J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management, 29*(4), 433–447.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing and Management, 37*(6), 817–842.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology, 9*(1), 99–115.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61–74.

Eto, H. (2000). Authorship and citation patterns in operational research journals in relation to competition and reform. *Scientometrics, 47*(1), 25–42.

Eto, H. (2002). Authorship and citation patterns in Management Science in comparison with operational research. *Scientometrics, 53*(3), 337–349.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal of Digital Libraries, 3*(2), 117–132.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In: K. B. Laskey & H. Prade (Eds.), *Proceedings of the 15th conference on uncertainty in artificial intelligence* (pp. 289–296). San Fransisco, CA: Morgan Kaufmann.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42*(1–2), 177–196.

Jacquemin, C. (2001). *Spotting and discovering terms through natural language processing*. Cambridge, MA: MIT 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.

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering, 1*(1), 9–27.

Kageura, K., & Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology, 3*(2), 259–289.

Kopcsa, A., & Schiebel, E. (1998). Science and technology mapping: A new iteration model for representing multidimensional relationships. *Journal of the American Society for Information Science, 49*(1), 7–17.

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* (pp. 187–213). Dordrecht: Kluwer Academic Publishers.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools, 13*(1), 157–169.

Maynard, D., & Ananiadou, S. (2000). Identifying terms by their family and friends. In: *Proceedings of the 18th conference on computational linguistics* (pp. 530–536). Morristown, NJ: Association for Computational Linguistics.

Noyons, E. C. M. (1999). *Bibliometric mapping as a science policy and research management tool*. PhD thesis, Leiden University.

Noyons, E. C. M. (2004). Science maps within a science policy context. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 237–255). Dordrecht: Kluwer Academic Publishers.

Noyons, E. C. M., & Van Raan, A. F. J. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science, 49*(1), 68–81.

Pazienza, M. T., Pennacchiotti, M., & Zanzotto, F. M. (2005). Terminology extraction: An analysis of linguistic and statistical approaches. In: S. Sirmakessis (Ed.), *Knowledge mining: Proceedings of the NEMIS 2004 final conference* (pp. 255–279). Berlin Heidelberg: Springer.

Peters, H. P. F., & Van Raan, A. F. J. (1993). Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling. *Research Policy, 22*(1), 23–45.

Rip, A., & Courtial, J. P. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics, 6*(6), 381–400.

Schmid, H. (1994), Probabilistic part-of-speech tagging using decision trees. In: *Proceedings of the international conference on new methods in language processing* (pp. 44–49).

Schmid, H. (1995), Improvements in part-of-speech tagging with an application to German. In: *Proceedings of the ACL SIGDAT workshop* (pp. 47–50).

Schneider, J. W. (2006). Concept symbols revisited: Naming clusters by parsing and filtering of noun phrases from citation contexts of concept symbols. *Scientometrics, 68*(3), 573–593.

Van Eck, N. J., & Waltman, L. (2007a). VOS: A new method for visualizing similarities between objects. In: H.-J. Lenz & R. Decker (Eds.), *Advances in data analysis: Proceedings of the 30th annual conference of the German classification society* (pp. 299–306). Berlin Heidelberg: Springer.

Van Eck, N. J., & Waltman, L. (2007b). Bibliometric mapping of the computational intelligence field. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 15*(5), 625–645.

Van Eck, N. J., & Waltman, L. (2009). How to normalize co-occurrence data? An analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology*, *60*(8), 1635–1651.

Van Eck, N. J., & Waltman, L. (in press). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*.

Van Eck, N. J., Waltman, L., Noyons, E. C. M., & Buter, R. K. (2008). *Automatic term identification for bibliometric mapping*. Technical Report ERS-2008-081-LIS, Erasmus University Rotterdam, Erasmus Research Institute of Management.

Wang, X., McCallum, A., & Wei, X. (2007), Topical N-grams: Phrase and topic discovery, with an application to information retrieval. In: *Proceedings of the 7th IEEE international conference on data mining* (pp. 697–702). Washington, DC: IEEE Computer Society.