Fraunhofer ISI Discussion Papers *Innovation Systems and Policy Analysis* No. 42

ISSN 1612-1430

Karlsruhe, March 2014

Emerging Topics in Science

Subproject in the Kompetenzzentrum Bibliometrie

Carolin Michels, Achim Rettinger

Fraunhofer Institute for Systems
and Innovation Research ISI

Contents

Figures

Tables

Acknowledgments

The research underlying this paper was supported by the German Federal Ministry for Education and Research (BMBF, project number 01PQ08004D). Certain data included in this paper are derived from the Science Citation Index Expanded (SCIE), the Social Science Citation Index (SSCI), the Arts and Humanities Citation Index (AHCI), and the Index to Social Sciences & Humanities Proceedings (ISSHP) (all updated June 2012) prepared by Thomson Reuters (Scientific) Inc. (TR®), Philadelphia, Pennsylvania, USA, USA: ©Copyright Thomson Reuters (Scientific) 2012. All rights reserved.

1 Introduction

The early identification of emerging topics in science can be a helpful indicator for scientists, publication companies as well as funding agencies. Scientists can be interested in emerging topics to evaluate or direct their own research according to current trends. Companies dealing with scientific publications, i.e. publishers, scientific database administrators etc., organize their publications according to categorization schemes. These schemes have to be adjusted to recent trends in order to reflect the scientific structures and connections of scientific literature appropriately. Therefore, the monitoring of the ongoing development of emerging and vanishing topics that should be in- or excluded from the scheme is an important necessity. Furthermore, funding agencies (and also leading researchers) might include new topics in their funding programs in order to support their development at a certain university, region, etc.

Therefore, the goal of this project was an approach to identify scientific publications that deal with an emerging topic. It is based on an extended version of Latent Dirichlet Allocation (LDA). Thus, for any document set that is used as an input, topics are calculated that can be represented by the respective term-probabilities.

We perform this step for disjoint time periods. For the sake of simplicity, we use publication years since these are in most cases the best covered time specification in scientific databases. In the second step, the term-probabilities of the clusters of each period are compared to the clusters of the foregoing period. Based on these probabilities, the similarity of each pair of clusters is calculated. For each cluster of the more recent time period, the maximum value of this similarity is used to determine the preceding cluster from which the newer cluster most likely evolved. The clusters of each period are ranked according to their maximum similarity values, i.e. the probability, that they evolved at all from a former cluster. The *n* clusters with the smallest maximum similarity are chosen as emerging topic candidates, since these are the topics that are most likely to be new.

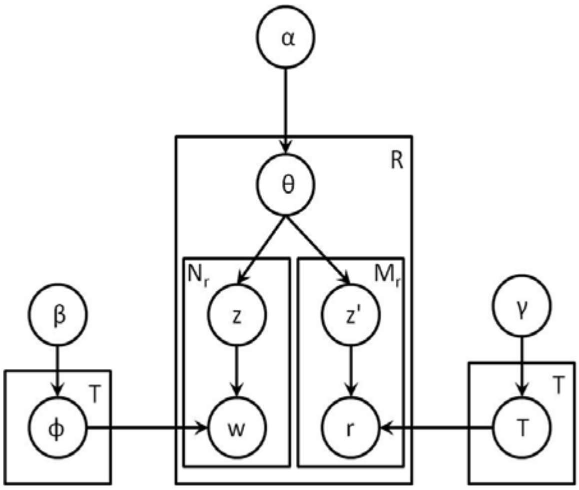
In order to evaluate our approach, we built a dataset that contained a subset of documents that were manually labeled as emerging topic documents. Therefore, we collected documents that belonged to a pre-defined emerging topic and mixed them with documents of which this information was unknown to us, i.e. they could either belong to a new topic in that specific time period that was not in our pre-defined list of emerging topics or simply belong to any other kind of topic. In the evaluation, the approach was measured according to the share in documents of emerging topics in the set of emerging topic candidates. Thus, the goal was to find as many correct emerging topic candidates as possible, while keeping the number false candidates as low as possible. Those documents that were not labeled as emerging topic documents could therefore disturb the evaluation. Therefore, the set of false candidates had to be also evaluated manually. This was a necessary evil since a complete coverage of all documents belonging to an emerging topic was not feasible as this was the purpose of the approach to be developed.

In the following, we present work that is methodological or in the application related to the approach (Section "Related Work"). We then briefly explain the structure of the dataset and the fundamental definition of the emerging topics (Sections "Dataset" and "Defintion ETs"). In Section "Proposed approach", we describe the overall approach and its two components: the LDA-based clustering and the connection of these clusters in different time periods. The Section "Evaluation" shows the results on the dataset described in the earlier sections. Finally, we give a summary and an outlook for the approach described in this paper (Section "Conclusion").

2 Related Work

Latent topic modeling has been used in various application scenarios to identify topics in textual document collections. One approach to discover the hidden topics in such a dataset is LDA (Blei et al. 2003). The basic implementation of LDA derives the latent term probabilities of *k* topics in the document set so that a representation of each document as a mixture of these topics is possible. LDA has been extended to use links or references of these documents as a second latent variable (Erosheva et al. 2004).

Figure 1: The Reference LDA Model.



Source: Own illustration

In this model, there is not only a multinomial distribution over *N* vocabulary items of author names and terms in the title but also a second multinomial distribution over *M* reference items (see Figure 1 for a graphical representation of the model). For this distribution, the *k* topics are drawn independently from γ. Like in the standard LDA model, the result are the *k* topics and the corresponding latent distribution Φ, that assigns each word to each topic with probability *p*. Additionally to the standard model, the extended approach delivers the latent distribution T for assigning each reference to one of the *k* topics with probability *p*. In this work, the approach was evaluated as a mean to find the hidden topics in a time period. These topics of each time period were then compared to those of the preceding period in order to detect new emerging topics. The goal of this work was to proof that the reference extension was suitable for this specific application.

The reference LDA model was also extended and varied in the use of the references (see e.g. Dietz et al. 2007; Nallapati et al. 2008; Nallapati/Cohen 2008). Similar to this extension is an LDA model that captures tags in collaborative systems (cf. Blei/Jordan 2003; Bundschus et al. 2009).

Dietz et al. (2007) use the terms of the cited documents instead of the references themselves. A similar approach was proposed by He et al., wherein the documents and therefore their term distributions are divided in two parts: the autonomous and the inherited parts (He et al. 2009). Since one of our main assumptions is that the vocabulary for a topic evolves over time and is especially volatile in the phase of the topic's emergence, the inclusion of the terms used in cited work would rather hinder our approach. We include references because we believe that they built a common foundation, i.e. a shared ground, for an emerging topic. But the vocabulary for a new topic should be independent from the old terminology to the greatest possible extent.

Mann et al. extend the bibliometric analysis of documents to topics detected by latent topic modeling using a model called Topical N-Grams (Mann et al. 2006). They show various possibilities to automatically assess the topics gained by the topic modeling. Some of their metrics might be applicable to investigate upon emerging topics, but the high reliance on citation metrics, i.e. in particular highly cited papers or median age of citations to a topic, conflicts with the goal of this work to detect the topics with a time lag as small as possible. In our work, since no citations are used, the minimum require are comparable time periods of a field wherein in the most recent time period the emerging topics are to be detected.

Also, LDA has been extended to use authors as a further distribution (see e.g. (see e.g. Rosen-Zvi et al. 2004; Steyvers et al. 2004). In the following (and also in Michels/Rettinger 2012), we test authors not as a second distribution but as a textual input equally weighted to the other terms in the document text. In that way, the probability distribution takes into account that an author has a certain probability to work on a specific topic for which he has been known. On the other hand, the mixture model helps to not exclusively attribute one topic for each author but also allow for a division among different topics where new topics can enter any time.

Further applications of LDA in this context have been author community detection (see Liu et al. 2009; McCallum et al. 2005).

Griffiths and Steyvers already applied LDA to identify and analyze the topics in the publications of PNAS (Griffiths/Steyvers 2004). They used fixed values $\beta = 0.1$ and $\alpha = 50/k$ and varied k in order to determine the best choice. They finally used k=300 for their dataset. Since this is the most similar application of LDA, we adopt some parts of their procedure. This is explained in more detail in Section "Proposed Approach". They also used a linear trend analysis on the θ of the different years in order to determine those topics that were in particular hot or cold.

3 Dataset

In order to train and evaluate our approach with a genuine dataset, we defined and collected eight ETs in science in different development stages. We "hid" these ETs in three sets of randomized publication data, the "hay". In that way, we created a training and test set for our approach: A dataset in which the ETs were contained, but which included enough noise, i.e. other non–ET documents that made the task of finding the ET documents nearly as difficult as in a real world application.

As explained in the introduction, it was possible that the hay documents belonged to an ET that was not in our set of the eight pre-defined ETs. The training of the approach should not be disturbed by this fact since the correct hay documents, i.e. negative training examples, and the correct ET documents, i.e. positive training examples, should outweigh the false negative training examples.

Considering the evaluation, the false hay documents could indeed corrupt the results if they were included in the list of ET candidates as a result of the approach. Therefore, we had to verify manually for some metrics whether the hay documents selected by the approach as ET candidates were not indeed ET documents even when they were not in the preselected set.

As explained above, the dataset was divided into three subsets, each representing a scientific discipline. We assume that an end-user would only be interested in the emerging topics in (his) specific disciplines and would therefore run the algorithm on a dataset consisting of these disciplines. Nonetheless, the approach would work just as well on a dataset consisting of multiple disciplines and should thus also be able to detect topics that emerge from a mixture of disciplines. But for illustrative and performance reasons, we decided for 3 separate haystacks. Our set of ETs made this also possible, because it could be clearly divided upon the three disciplines.

The training of the approach was performed with all available information for the publication years 2000 and 2001, while the evaluation encompassed the following years. Thus, we adjusted all parameters with information that was not used in the evaluation. In the final application, the end-user should first select a scientific discipline he wants to analyze and then to adjust with some test runs the parameters himself in order to get the best possible results on his specific dataset. Thus, he would make some test runs of the approach with varying parameters and evaluate and compare (excerpts of) the ET candidate lists before deciding for the best results.

4 Definition ETs

The collected ETs can be described by the following labels:

1. Biodiesel from waste/cooking oil
2. DNA decoding
3. Energy harvesting
4. HVDC
5. Life-logging
6. Location based services
7. User-generated content
8. Wearable devices

The topics were chosen and evaluated according to their novelty and their growth characteristics in scientific publications. Some of them started before the year 2000 and a pure bibliometric analysis based on their publication and citation numbers showed quite different scientific activity and reception of those topics. Nonetheless, all topics had a peak in publication and citation numbers in the respective period and thus seemed to be "emerging" in this time.

For each ET, a keyword search was performed on Thomson Reuters' Web of Science. In this keyword search, all publications with a publication year of 2000 or later were included.

The keyword query was expressed as broad as possible to collect the maximum number of documents for each ET. Therefore, keywords were used that might by trend emit too much documents instead of too few. This strategy was chosen because a vocabulary could not have been developed for a topic at an early stage. For instance, the topic "Energy harvesting" is described in the scientific publications in our dataset by, among others, the following terms: "self-powered (sensors)", "human-powered mobile computing", "vibration power generator system", "scavenging energy", "piezoelectricity energy" and "ambient energy".

Queried text fields were title, keywords and abstract of the publications. The argumentation for this strategy goes hand in hand with that for the broader keyword search. For an ET, the usage of specific terms describing it is not necessarily restricted to the document title or the keyword field. Also the topic might not be represented sufficiently in just one of these fields. E.g. the title of one document in the dataset for "Biodiesel from waste/cooking oil" had the title "Towards producing a truly green biodiesel". A fully automatic document selection would exclude such a document from the dataset. On the other hand, if the search was based solely on the title, the document would not be identified as a possible match. A closer look at the abstract revealed that the document deals indeed (among other topics) with a production of biodiesel that "uses waste vegetable oil". Thus, a manual selection of the documents had to be conducted after the keyword search.

As a final step to gather as many documents as possible for each ET, we queried those documents that were cited at least by at least 1 percent of the documents in the ET data set collected so far. These documents were again verified manually before they were added to the respective ET.

As a summary of the collection of the ETs, we can divide the procedure in 4 steps:

1. Keyword search
2. Manual selection of documents found by keyword search
3. Search for documents cited by at least 1% of the documents in the set so far
4. Manual selection of documents found by citation search

Of course, step 3 and 4 could be repeated various times to ensure that even those documents that were cited by those documents added in the last iteration of step 4 were considered as possible supplement. Since the changes made in the first iteration, we stopped the procedure at this point. Table 1 gives an overview over the number of documents found in each step and the final size of the ETs.

Table 1: Total size of ETs.

ET	Documents found in step 1	Size of ET after step 2	Documents found in step 3	Size of ET after step 4
biodiesel	1,575	150	437	170
dna decoding	364	72	2,675	73
energy harvesting	1,273	943	81	1,009
hvdc	1,339	805	22	819
life-logging	51	25	56	34
location based services	889	310	28	320
user-generated content	229	109	40	123
wearable devices	268	116	57	117

Source: Web of Science, own calculations.

By conducting step 3 and 4 only once, we might miss some more recent documents that could not be cited by enough documents in the set built after step 2, but these documents are negligible for two reasons. First, the vocabulary should be more fixed for these recent documents, thus the majority of them should be found in step 1. Secondly, this step is conducted in order to identify documents that precede the documents found so far and thus influence or redefine a) the point of time of the emergence of the topic and b) the documents that should be found by the approach at that point in time. For the ETs Energy harvesting, HVDC and Biodiesel some documents (8, 10 and 2) were found for publication years before 2000, but since these are still single points that built – in our opinion – merely the foundation for the emergence of the topics later, we neglected these documents for the remainder of this paper. The analyses of publication and citation numbers confirmed our assumption that the emergence of the topic was indeed in the analyzed time frame.

The ETs were aggregated in three sets, where each set contained only topics of the same field[1]. Therefore, the ETs were aggregated in the following sets:

- Engineering: Energy harvesting, HVDC, Biodiesel from waste/cooking oil
- Molecular Biology – Genetics: DNA decoding
- Computer Science: User-generated content, Location based services, Wearable devices , Life-logging

To build the "haystacks" on which the approach was tested, we added random documents ("hay") of the same field to each of these sets. The number of random documents added equals a tenth of the number of documents in each year for the specific field. These 10% were also cleaned, so that all publications that did not have at least 3 references or had an empty title or one containing less than 6 characters were sorted out. This results in a much smaller dataset but one that is better suited for evaluation purposes, since there are less outliers that result from the database coverage or low quality.

Thus, the number of documents in the haystack mirrors the actual growth of the subject category itself. Since our previous analysis of the ETs showed that they all of them emerged between the years 2000-2007, we restricted the haystacks to this period. Table 2 gives an overview of the size and composition of each haystack. Table 3 shows the relation of ET to hay documents in the datasets.

[1] Based on the journal allocation to the 22 fields defined for the *Essential Science Indicators*^{SM}, http://sciencewatch.com/about/met/journallist/, retrieved April 10th 2012.

Table 2: Overview of Haystacks for the Years 2000 to 2007

	Haystack 1 (Engineering)	Haystack 2 (Molecular Biology - Genetics)	Haystack 3 (Computer Sciences)
ET documents	782	49	276
Hay	45,920	23,734	17,251
Total Size	46,702	23,783	17,527

Source: Web of Science, own calculations.

Table 3: Percentage of ET documents in the haystacks.

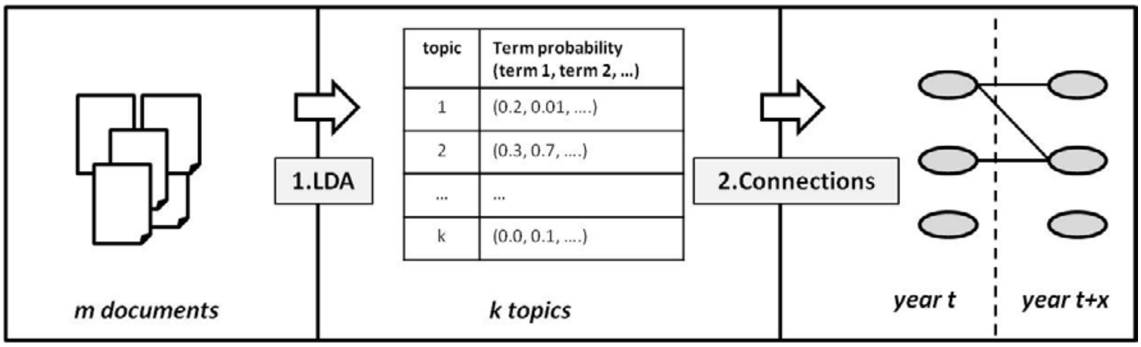
Year	2000	2001	2002	2003	2004	2005	2006	2007
Haystack 1	0.79%	1.25%	1.26%	1.07%	1.18%	1.55%	2.81%	2.75%
Haystack 2	0.21%	0.10%	0.13%	0.34%	0.36%	0.13%	0.16%	0.21%
Haystack 3	0.38%	0.56%	0.80%	1.50%	1.54%	1.71%	1.61%	2.74%

Source: Web of Science, own calculations.

5 Proposed approach

The approach starts with a set of publications in which those documents belonging to an ET are supposed to be found. First, the publications are clustered based on the terms used in the textual parts and their reference lists. This results in clusters of documents that have a common term distribution and share references in their reference lists. Thus, each document is assigned to a cluster with a probability *p* based on its term and reference distribution. A cluster is represented by its term distribution Φ and its reference distribution T. Figure 2 gives an overview of the complete process and exemplarily shows how the *k* topics found in the document set are represented by their term distribution Φ .Documents (in the original set or a new set) can be assigned to the topics according to the terms and references they use, but for the second step, the establishment of connections between topics of different time periods, only the distributions are used as they are. Thus, a topic is still represented solely by its vocabulary (Φ) and its connectivity to former topics (T) and not by actual instantiations.

Figure 2: Overview of the ET candidate selection process.



topic	Term probability (term 1, term 2, …)
1	(0.2, 0.01, ….)
2	(0.3, 0.7, ….)
…	…
k	(0.0, 0.1, ….)

Source: Own illustration.

The clustering is explained in more detail in section "LDA". It is conducted for each set of documents of each publication year in the dataset separately. Thus, only documents of the same year can be clustered together and we gain a separate set of clusters for each observed publication year.

In the second step, clusters of different years are compared based on their term distribution Φ and their reference distribution T.

In the final application, the first step would be conducted for each year in the dataset, while the second step, the establishing of the connections between clusters of one year with foregoing years and the selection of ET candidates, is applied for the most recent year only, since the end user would only be interested in the ET at the cutting edge.

5.1 LDA

Figure 1 gives an overview of the clustering approach. In the first step, an LDA approach is applied with *k*=500 topics. Our LDA approach was extended for the separate usage of references in the topic model.[2]

The set of vocabulary items used for the term distribution Φ could be collected from different parts of the documents. We tested the single usage and all possible combinations of the following fields:

- Title
- Abstract
- Keywords
- Authors

[2] The basic LDA algorithm that was extended was an implementation by (Heinrich 2008). Gibbs Sampling was applied with 1,000 iterations so that convergence should be reached (Griffiths 2002; Griffiths/Steyvers 2004).

In this model, the author names were used simply as other terms that could appear in the document like any other term. An extension of the existing model for the usage a third multinomial distribution for the author names as well is thinkable.

Preprocessing of the documents was performed before applying the LDA model. All stopwords were removed and the remaining terms were stemmed.[3]

Even though common terms would not disturb the LDA algorithm, they were eliminated after the preprocessing for performance reasons. Therefore, varying values for a threshold for the term appearance t_o will be tested. All terms occurring in more than t_o percent of the documents are eliminated from the vocabulary list. For instance, a value of t_o =50% removes all terms that appeared in more than 50% of the documents of that particular year. We used a fixed value of t_o = 50% in the following experiments.

Furthermore all terms and references that were used by only one document in a timer period were excluded. This improved the overall performance with no negative implications for the clustering since a term/reference that appeared in only one document could not give implications for similar documents. It might be the case that a term/reference of one year was excluded in this way even when it was used in another year and thus could have influenced the connection calculation. But we assume that.

5.2 Connections

For each pair of topics c_1 and c_2 of two time periods t_1 and t_2 we calculate the similarity value sim(c_1,c_2).

Connections are established between clusters if their similarity value exceeds the threshold t_c and the cluster of the more recent year has no higher similarity value with any other older cluster.

The similarity value is based on the similarity of the term and reference vectors of the clusters only. Note, that for two clusters of the same period, both Φs and Ts would contain the same terms since they were calculated with a common LDA model. This does not necessarily hold for clusters of different time periods. Thus, for two clusters c_1 and c_2 and their corresponding term distribution Φ_1 and Φ_2, primarily the union of both term sets has to be calculated. Then, the term vectors of both clusters can be determined so that entries at the same position of the vectors correspond to the same term.

There are different possibilities to calculate the similarity between two clusters based on these distributions:

[3] Porter Stemmer (1980).

The similarity for each pair of clusters is calculated as the cosine similarity between their term vectors. The values in the term vectors correspond to the probabilities in Φ. The same can be done for vectors consisting of the term and the reference probabilities in Φ and T.

The similarity between the distributions is calculated using the chi-square test to decide whether both distributions derive from the same distribution or not. Similarity of both clusters (or in this case more the distance) is then represented by the P-value.

A linear regression is calculated for both distributions, e.g. for the term distributions alone this would result in $\Phi_1 = a + b*\Phi_2$. After that, the hypothesis that "a = 0 and b=1" is tested and again the cluster similarity equals the P-value.

In previous work, for each cluster, the cluster from a previous year with the highest similarity value is selected. If this similarity exceeds a threshold *t*, a topical connection between both clusters is assumed. If no similarity is higher than *t,* the cluster seems not to continue any foregoing topics content-wise and thus no connection to any of those clusters can be found. The cluster seems to deal with a completely new topic and is thus listed to the end user in the list of ET clusters.

In this project, the connections between clusters were ranked according to their similarity value. Then, instead of a fixed threshold t for the similarity, we consider the *n* connections with the smallest similarity value as ETs. Therefore, *n* determines the number of ET candidates the end user would see.

Both distributions, Φ and T, are used for calculating the similarity values with varying weightings, i.e. the similarity between two clusters c_1 and c_2 is $sim(c_1,c_2) = w_\Phi * sim_\Phi(c_1,c_2) + (1- w_\Phi)* sim_T(c_1,c_2)$ where $w_\Phi \in [0;1]$. First experiments are conducted with $w_\Phi = 0$ and $w_\Phi = 1$ to decide whether the references could help to improve the similarity calculation or not.

5.3 Evaluation

Projected: After application of proposed approach, we calculate for varying years the following evaluation metrics:

1. Precision = (ET-Documents in ET candidate list)/(documents in ET candidate list)
2. Recall = (ET-Documents in ET candidate list)/(ET-Documents in dataset)
3. If not fixed number *n* is used: Absolute value for size of ET candidate list (manageable by end user?)

Furthermore we assess manually the cluster quality in the ET candidate list.

5.4 Parameter estimation

To estimate the parameters for the clustering and the cluster connection without corrupting the later evaluation, we restricted the tests to those documents of the years 2000 and 2001. Both years were clustered with our approach. Thereafter, the clusters in the year 2001 were connected to those in 2000. The clustering was evaluated by the number of clusters that consisted of documents of the ETs. For this purpose, we distinguish between ET, mixed and hay clusters. While the ET and hay clusters only contain documents from a ET or no ET respectively, the mixed clusters did not separate these classes. Desirable was a high number of pure (i.e. ET and hay) clusters.

For the evaluation of the connections, the number of hay and ET clusters without a connection was calculated. For ET and mixed clusters, no connection to any previous (hay) cluster was preferred, since this corresponded to the correct identified ET candidates. Hay clusters with no connection were false candidates and thus to avoid.

In the LDA approach, α, β and γ as well as the number of topics *k* were controllable. Since it was not possible to test all combinations of all parameters due to time constraints, we set *k*=500 and α=0.5. In spot tests, these parameters were corroborated. *k* was chosen after some initial tests with fixed parameters α, β and γ to evaluate which number of topics resulted in an appropriate number of clusters With α close to 0, we would have asked for one topic per document, but we still wanted to allow for the mixture models to represent the subsidiary topics. This left us with the parameters β and γ for the multinomial distributions. We tested all combinations for both parameters having values between 0.1 and 0.9. Table 4 shows the results for these experiments in the number of ET and mixed clusters for dataset 3. According to these results, we set β=0.4 and γ=0.2.

Table 4: Number of ET/Mixed Clusters for Dataset 3 in the Year 2001.

β \ γ	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0.1	**1/5**	0/6	**1/5**	**1/ 5**	0/6	0/5	0/6	0/6	**1/5**
0.2	0/6	0/5	0/6	0/6	**1/5**	0/6	0/6	0/6	0/6
0.3	0/6	**1/5**	0/6	0/6	0/6	0/6	0/6	0/6	0/6
0.4	0/5	**1/5**	0/6	0/6	0/6	**1/5**	**1/5**	0/5	**1/5**
0.5	0/6	0/6	0/6	0/6	0/6	0/6	0/6	0/6	0/6
0.6	0/6	0/6	0/6	0/6	0/6	**1/ 5**	0/6	0/6	0/6
0.7	0/6	0/6	0/6	0/6	0/6	0/6	0/6	0/6	0/6
0.8	0/6	0/6	0/6	0/6	0/6	0/6	0/6	0/6	0/6
0.9	**1/5**	**1/5**	0/6	0/6	0/6	0/6	0/6	**1/5**	0/6

Source: Web of Science, own calculations.

The fact that we found one ET at most seems a bit disappointing but bearing in mind that this corresponds to identifying 1 out of 6 ET documents in a set of 1642 documents in 2001 these results seem very promising. These figures also show that most mixed clusters consisted of only one ET document mixed with hay documents.

Next, we wanted to determine the similarity calculation and threshold for the connection of the clusters of different time periods. First, we had to create a training set for this purpose: We calculated randomized clusters for each dataset for the years 2000 and 2001 with varying (randomized) sizes. The only condition in this clustering was that hay and ET documents were separated in this clustering. Then, we built pairs of clusters of the different years (2000 and 2001). If the pair consisted of an ET cluster in the more recent year, the connection to the older cluster was assumed to be false and correct otherwise (that is, if the more recent cluster was a hay cluster and thus based on former work). Instead of trying to detect topical relations, the connections between these clusters were supposed to measure only the novelty of the more recent cluster, i.e. if it could be separated from all the other former clusters or not. Thus, supposing connections for all pairs of hay clusters was a necessary and valid step in order to create a feasible training set. A random sample of 8,421 cluster pairs with false connections and 32,975 pairs with true connections was drawn to train the similarity calculation. This is how we derived a linear regression using aforementioned features and the threshold *t* to determine whether a connection between two clusters was correct or not.

The formula for calculating connections between previous and recent clusters was also trained on the cluster results gained by the clustering step with β=0.4 and γ=0.2. Table 5 shows the number of hay, ET, mixed clusters in the different datasets that were contained in the dataset and were therefore considered in the connection step. The number of clusters that were not connected to any cluster from a previous year is given in the last column.

Table 5: Number of ET/Mixed/Hay Clusters for the Year 2001.

	Overall (ET/mixed/hay)			Not connected (ET/mixed/hay)		
Dataset1	0	52	448	0	1	7
Dataset2	0	3	496	0	0	2
Dataset3	1	5	471	1	0	51

Source: Web of Science, own calculations.

6 Results

The test set for evaluating our approach was set up with the remaining years of the haystacks, i.e. the years 2002 to 2007. The documents for these years were clustered and connected to the clusters of the foregoing years. The clusters for the years 2000 and 2001 were calculated for this purpose anew, but are not as such taken into account for the evaluation. Of course, their quality also influences the number of connections made for the following years, so that they indirectly affect the overall results. Previous tests suggested that the number of connected ET and mixed clusters is relatively stable in terms of cluster changes in other years.

Table 6 shows the results for the three datasets in terms of the number of ET, mixed and hay clusters. All in all, the total number of clusters varied between 483 and 500. Most of the mixed clusters contained at most one ET document. In rare cases, two ET documents were clustered together and in only 46 of the 802 mixed clusters in total for the evaluation period two ET were represented. This corroborates our assumption, that ET documents in the early development stage are not connected yet by vocabulary or similar and thus are outliers for the dataset as a whole.

Table 6: Number of ET/Mixed/Hay Clusters after Applying the Proposed Approach.

Year	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**
Test Set 1	0/59/441	0/48/452	0/50/450	0/76/424	0/145/355	0/157/343
Test Set 2	0/4/495	0/9/488	0/11/489	0/4/495	0/4/496	0/7/493
Test Set 3	1/10/479	0/21/479	0/32/466	0/44/455	0/41/459	0/80/420

Source: Web of Science, own calculations.

We investigated the connectedness of these clusters with clusters from previous years (Table 7). Since the training for the ET candidate selection was rather restrictive, the set of ET candidates is relatively small. The one pure ET cluster that was found by our approach in dataset 3 was also not connected to any other cluster in 2000 or 2001. Also, one mixed cluster was separated from the other clusters in the dataset 1. Thus, for 2002, the approach had a hit rate of approx. 5 %, reducing the user effort of inspecting 11,331 for the year 2002 documents to merely having a look at 81. In that way, two out of 5 detectable ET for 2002 could be found. Even though the number of ET candidates would be handlebar for further manual inspection, the approach fails in detecting ETs after 2002. In total, an end-user would have to inspect 112 of 81,051 documents in the years 2002 to 2007.

Table 7: Number of ET/Mixed/Hay Clusters in the ET Candidate Lists.

Year	2002	2003	2004	2005	2006	2007
Test Set 1	0/1/10	0/0/0	0/0/0	0/0/0	0/0/0	0/0/0
Test Set 2	0/0/4	0/0/4	0/0/2	0/0/1	0/0/2	0/0/2
Test Set 3	1/0/27	0/0/14	0/0/0	0/0/3	0/0/1	0/0/2

Source: Web of Science, own calculations.

Therefore, one main finding of this work is that our approach definitely performs best for the year 2002. Reasons for this could be that we trained our approach on the year 2001 and the characteristics for the dataset changed significantly over time. Another problem was probably the configuration of the dataset, as a manual investigation of the ET candidate documents showed that the selection of the ET documents was highly corrupted by other "outlier" documents. These documents might consist of short and ambiguous titles or have an insufficient or misleading reference list that led to false results in the extended LDA. Also, our dataset selection did not necessarily lead to the full inclusion of previous work of "established" topics (in contrast to emerging ones). Thus, a connection to a former instance of a topic might not be found even though it exists and the respective (hay) cluster would be included in the ET candidate list. Another problem that we encountered was that even a document labeled as ‘hay’ could represent an ET, even if this ET was not in our list of ETs. In dataset 3 in particular, the hay clusters selected as ET candidates contained novel approaches which might indeed be a seed document for a new topic.

7 Conclusion

We presented an approach for building clusters of scientific publications according to the information given by their title, authors and reference list and for selecting a list of ET candidates from these clusters. The approach was trained with information from the years 2000 and 2001 and evaluated for the years 2002 to 2007, which correspond to the years in our dataset in which ETs actually emerge. The information that would have to be processed manually could be reduced to less than 0.14 % of the dataset.

We experienced problems because the actual relevance of the "hay" documents in the train and test set was still unknown. A hay document still could represent an ET that was not included in our list when creating these artificial haystacks. It could also be an outlier due to a deviant title, reference list etc. For instance, in the ET candidate list for dataset 3, there were documents (or 1-instance clusters) that dealt with automatically recognizing traffic signs and extracting validation rules from microbiological data. Such documents might indeed represent a topic that might have been established as a new scientific topic. Or they could simply represent a novel application for established methodologies. A clear trend in the ET candidate list for dataset 3 could be identified for grid systems which was then a newly established scientific topic.

Our evaluation showed that even-though we included citations as a second non-textual feature for LDA, the approach was still prone to disruptions based on insufficient or misleading information in both titles and reference lists. An improvement of the clustering step in future work with further extensions of LDA would also influence and probably improve the connection step. Nonetheless, this step could be further improved with a more specific analysis of the individual connections made. Since our results were worse for more recent years, the connection step could be influenced by the sheer number of connection candidates, since for each year approximately 500 cluster candidates are added.

Another factor for future work would be to investigate upon the fact that the performance for dataset 3 was significantly better than for the other datasets.

All in all, the results suggest that the composition of our approach and its implementation are promising, but further investigations have to be made in terms of extensibility and appropriate evaluation environments.

8 References

Blei, D.M./Jordan, M.J. (2003): Modeling annotated data, Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 127-134. Toronto, Canada.

Blei, D.M./Ng, A.Y./Jordan, M.J. (2003): Latent Dirichlet Allocation, Journal of Machine Learning Research, 3, 993-1022.

Bundschus, M./Yu, S./Tresp, V./Rettinger, A./Dejori, M. (2009): Hierarchical Bayesian Models for Collaborative Tagging Systems, IEEE International Conference on Data Mining. Miami, Florida.

Dietz, L./Bickel, S./Scheffer, T. (2007): Unsupervised Prediction of Citation Influences, Proceedings of the 24th International Conference on Machine Learning 2007. Corvallis, Oregon: Oregon State University.

Erosheva, E./Fienberg, S./Lafferty, J. (2004): Mixed Membership Models of Scientific Publications, Proceedings of the National Academy of Sciences of the United States of America, 101 (Suppl. 1), 5229 - 5227. Washington, D.C.: National Academy of Sciences.

Griffiths, T.L. (2002): Gibbs sampling in the generative model of Latent Dirichlet Allocation, Technical Report. Stanford, California: Stanford University.

Griffiths, T.L./Steyvers, M. (2004): Finding scientific topics, Proceedings of the National Academy of Sciences 101 (Suppl. 1), 5228-5235.

He, Q./Bi, C./Pei, J./Qiu, B.P.M./Giles, C.L. (2009): Detecting topic evolution in scientific literature: how can citations help?, Proceedings of the 8th ACM Conference on Information and Knowledge Management (CIKM 2009). Hong Kong, China: CIKM 2009.

Heinrich, G. (2008): Parameter estimation for text analysis. Darmstadt: Fraunhofer IGD.

Liu, Y./Niculescu-Mizil, A./Gryc, W. (2009): Topic-Link LDA: Joint Models of Topic and Author Community, Proceedings of the 26th International Conference on Machine Learning. Montreal, Canada.

Mann, G.S./Mimno, D./McCallum, A. (2006): Bibliometric Impact Measures Leveraging Topic Analysis, JCDL '06: Proceedings of the Joint Conference on Digital Libraries. Chapel HIll, North Carolina.

McCallum, A./Corrada-Emmanuel, A./Wang, X. (2005): Topic and Role Discovery in Social Networks, Computer Science Department Faculty Publication Series (Paper 3). Amherst, Massachusetts: University of Massachusetts.

Michels, C./Rettinger, A. (2012): The Tell-Tale Title: How to Track Topics Over Time with a Two-Step Approach, Proceedings of the 17th International Conference on Science and Technology Indicators (STI). Montreal, Quebec, Canada.

Nallapati, R./Ahmed, A./Xing, E.P./Cohen, W.W. (2008): Joint Latent Topic Models for Text and Citations, Proceedings of The Fourteen ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). Las Vegas, Nevada.

Nallapati, R./Cohen, W. (2008): Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs, International Conference for Weblogs and Social Media. Seattle, Washington.

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

Rosen-Zvi, M./Griffiths, T./Steyvers, M./Smyth, P. (2004): The Author-Topic Model for Authors and Documents, Proceedings of the 20th conference on Uncertainty in artificial intelligence, 487–494. ARlington, Virginia: AUAI Press.

Steyvers, M./Smyth, P./Rosen-Zvi, M./Griffiths, T. (2004): Probabilistic author-topic models for information discovery, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, Washington.