Ontology Forecasting in Scientific Literature: Semantic Concepts Prediction based on Innovation-Adoption Priors

Conference or Workshop Item

Ontology Forecasting in Scientific Literature: Semantic Concepts Prediction based on Innovation-Adoption Priors

Amparo Elizabeth Cano-Basave[1], Francesco Osborne[2] and Angelo Antonio Salatino[2]

[1] Aston Business School, Aston University, UK
a.cano-basave@aston.ac.uk
[2] Knowledge Media Institute, Open University, UK
francesco.osborne@open.ac.uk, angelo.salatino@open.ac.uk

Abstract. The ontology engineering research community has focused for many years on supporting the creation, development and evolution of ontologies. Ontology forecasting, which aims at predicting semantic changes in an ontology, represents instead a new challenge. In this paper, we want to give a contribution to this novel endeavour by focusing on the task of forecasting semantic concepts in the research domain. Indeed, ontologies representing scientific disciplines contain only research topics that are already popular enough to be selected by human experts or automatic algorithms. They are thus unfit to support tasks which require the ability of describing and exploring the forefront of research, such as trend detection and horizon scanning. We address this issue by introducing the Semantic Innovation Forecast (SIF) model, which predicts new concepts of an ontology at time $t + 1$, using only data available at time t. Our approach relies on lexical innovation and adoption information extracted from historical data. We evaluated the SIF model on a very large dataset consisting of over one million scientific papers belonging to the Computer Science domain: the outcomes show that the proposed approach offers a competitive boost in mean average precision-at-ten compared to the baselines when forecasting over 5 years.

Keywords: Topic Evolution, Ontology Forecasting, Ontology Evolution, Latent Semantics, LDA, Innovation Priors, Adoption Priors, Scholarly Data

1 Introduction

The mass of research data on the web is growing steadily, and its analysis is becoming increasingly important for understanding, supporting and predicting the research landscape. Today most digital libraries (e.g., ACM Digital Library, PubMed) and many academic search engines (e.g., Microsoft Academic Search[3], Rexplore [21], Saffron [18]) have adopted taxonomies and ontologies for representing the domain of research areas. For example, researchers and publishers in the field of Computer Science are now well familiar with the ACM classification and use it regularly to annotate publications.

However, these semantic classifications are usually hand-crafted and thus are costly to produce. Furthermore, they grow obsolete very quickly, especially in rapidly changing fields such as Computer Science. To alleviate this task is possible to use approaches

[3] http://academic.research.microsoft.com/

for ontology evolution and ontology learning. The first task aims to extend, refine and enrich an ontology based on current domain knowledge [26, 23]. For example, an ontology of research areas should be updated regularly by including topics which emerged after the last version of the ontology was published. Ontology learning aims instead to automatically generate ontologies by analysing relevant sources, such as relevant scientific literature [20]. Nonetheless, these ontologies still reflect the past, and can only contain concepts that are already popular enough to be selected by human experts or automatic algorithms. Hence, while they are very useful to produce analytics and examine historical data, they hardly support tasks which involve the ability to describe and explore the forefront of research, such as trend detection and horizon scanning. It is thus crucial to develop new methods to allow also the identification of emerging topics in these semantic classifications.

Nonetheless, *predicting the emergence of semantic concepts*, is still a challenge. To the best of our knowledge, predicting the future iteration of a ontology and the relevant concepts that will extend it, which we refer to as *ontology forecasting*, is a novel open question.

For the particular case of scholarly data, being able to predict new research areas can be beneficial for researchers, who are often interested in emerging research areas; for academic publishers, which need to offer the most up-to-date contents; and for institutional funding bodies and companies, which have to make early decisions about critical investments.

In this paper, we address this challenge by presenting a novel framework for the prediction of new semantic concepts in the research domain, which relies on the incorporation of lexical innovation and adoption priors derived from historical data. The main contributions of this work can be summarised as follows:

1. We approach the novel task of ontology forecasting by predicting semantic concepts in the research domain;
2. We introduce two metrics to analyse the linguistic and semantic progressiveness in scholarly data;
3. We propose a novel weakly-supervised approach for the forecasting of innovative semantic concepts in scientific literature;
4. We evaluate our approach in a dataset of over one million documents belonging to the Computer Science domain;
5. Our findings demonstrate that the proposed framework offers competitive boosts in mean average precision at ten for forecasts over 5 years.

2 Related Work

The state of the art presents several approaches for identifying topics in a collection of documents and determining their evolution in time. The most adopted technique for extracting topics from a corpus is Latent Dirichlet Allocation (LDA) [4], which is is a generative statistical model that models topics as a multinomial distribution over words. LDA has been extended in a variety of ways for incorporating research entities. For example, the Author-Topic model (ATM) [24] included authorship information in the generative model. Bolelli et al. [6] extended it even further by introducing the Segmented Author-Topic model, which also takes in consideration the temporal ordering of documents to address the problem of topic evolution. In scenarios where it already exists a

taxonomy of research areas [21], it is also possible to use entity linking techniques [7] for mapping documents to related concepts. For example, the Smart Topic Miner [22], an application used by Springer Nature for annotating proceedings books, maps keywords extracted from papers to the automatically generated Klink-2 Computer Science Ontology [20] with the aim of selecting a comprehensive set of structured keywords.

The approaches for topic evolution can be distinguished in discriminative and generative [13]. The first ones consider topics as a distribution over words or a mixture over documents and analyse how these change in time using a variety of indexes and techniques [25]. For example, Morinaga and Yamanishi [19] employed a Finite Mixture Model to represent the structure of topics and analyse diachronically the extracted component and Mei and Zhai [16] correlated term clusters via a temporal graph model. However, these methods do not take advantage of the identification of lexical innovations and their adoption across years, but rather focus only on tracking changes in distributions of words.

The second class of approaches for topic evolution employ instead generative topic models [5] on document streams. For example, Gohr et al [11] used Probabilistic Latent Semantic Analysis and proposed a folding-in techniques for a topic adaptation under an evolving vocabulary. He et al [13] characterised the analysis of the evolution of topics into the independent topic evolution (ITE) and accumulative topic evolution (ATE) approaches. However, these models do not cater for the identification of novel topics, but rather caters for tracking change of existing ones.

In addition, some approaches aim at supporting ontology evolution by predicting extensions of an ontology. For example, Pesquita and Couto [23] introduced a method for suggesting areas of biomedical ontologies that will likely be extended in the future. Similarly Wang et al [26] proposed an approach for forecasting patterns in ontology development, with the aim of suggesting which part of an ontology will be next edited by users. Another relevant approach is iDTM (infinite dynamic topic model) [1], which studies the birth, death and evolution of topics in a text stream. iDTM can identify the birth of topics appearing on a given epoch, such topics are considered new when compared to previous epochs. In contrast to their work, our proposed model addresses the prediction of new topics in *future* epochs based on past data rather than identifying topics on the current epoch. In addition, our work is different from all previous approaches because we aim at predicting new classes (concepts) that will appear in the future representations of an ontology.

3 Language and Semantic Progressiveness in Scientific Literature

Previous work has studied the role of language evolution and adoption in online communities showing that users' conformity to innovation can impact the churn or grow of a community [9]. Inspired by this fact, we follow the intuition that language innovation and adoption could impact the generation and expiration of semantic concepts modelling a shared conceptualisation of a domain.

This section presents a motivation for predicting semantic concepts in scientific literature based on the study of the use of language in scholarly data. The following subsection 3.1 introduces the dataset used in this paper and presents an analysis of the evolution of language in the field of Computer Science during the course of 14 years in subsections 3.2 and 3.3.

3.1 Dataset Description

Our dataset comprises of a collection of research articles relevant to the Computer Science field extracted from Scopus[4], one of the largest databases of peer-reviewed literature. The full 14 years collection ranges from 1995-2008 with a total of 1,074,820 papers. Each year consists of a set of papers categorised within a semantic representation of the Computer Science domain. Such ontological representation is generated per two year-corpus starting from 1998 using the Klink-2 algorithm [20].

The Klink-2 algorithm combines semantic technologies, machine learning and knowledge from external sources (e.g., the LOD cloud, web pages, calls for papers) to automatically generate large-scale ontologies of research areas. It was built to support the Rexplore system [21] a system that integrates statistical analysis, semantic technologies and visual analytics to provide support for exploring and making sense of scholarly data. In particular, the ontology generated by Klink-2 enhances semantically a variety of data mining and information extraction techniques, and improves search and visual analytics.

The classical way to address the problem of classifying research topics has been to adopt human-crafted taxonomies, such as the ACM Computing Classification System and the Springer Nature Classification. However, the ontology created by Klink-2 presents two main advantages over these solutions. Firstly, human-crafted classifications tend to grow obsolete in few years, especially in fields such as Computer Science, where the most interesting topics are the emerging ones. Conversely, Kink-2 can quickly create a new ontology by running on recent data. Secondly, Klink-2 is able to create huge ontologies which includes very large number of concepts which do not appear in current manually created classifications. For example, the current version of the full Klink-2 Computer Science ontology includes 17 000 concepts and about 70 000 semantic relationships.

The data model of the Klink-2 ontology is an extension of the BIBO ontology which in turn builds on SKOS. It includes three semantic relations: *skos:broaderGeneric*, which indicates that a topic is a sub-area of another one (e.g., Linked Data is considered a sub-area of Semantic Web); *relatedEquivalent*, which indicates that two topics can be treated as equivalent for the purpose of exploring research data (e.g., Ontology Matching, Ontology Mapping); and *contributesTo*, which indicates that the research outputs of one topic significantly contribute to research into another (e.g., Ontology Engineering contributes to Semantic Web, but arguably it is not its sub-area).

The ontologies associated to different years were computed by feeding to Klink-2 all publications up to that year, to simulate the normal situation in which Klink-2 regularly updates the Computer Science ontology according to most recent data. Figure 1 presents general statistics of the dataset including number of articles, size of the vocabularies and number of semantic concepts per year ontology. Each paper is represented by its title and abstract. Vocabulary sizes where computed after removing punctuation, stopwords and computing Porter stemming [27]. The data presented in Figure 1 indicates that as years go by the production of scholarly articles for the Computer Science increases. Moreover, it shows that as more articles are introduced each year, novel words – not mentioned in previous years– are also appearing. When analysing the number of semantic concept over time we see that every year there is also an augmentation of the ontological concepts describing the Computer Science field. The following subsections analyse language and ontology evolution on this dataset.

[4] Scopus, https://www.elsevier.com/solutions/scopus

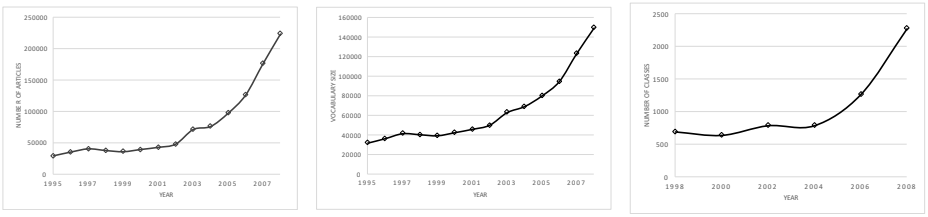

Fig. 1: From left to right, a) number of articles per year, b) vocabulary size per year, c) number of classes per year.

3.2 Linguistic Progressiveness

Language innovation in a corpus refers to the introduction of novel patterns of language which do not conform to previously existing patterns [9]. Changes in time on the use of lexical features within a corpus characterise the language evolution of such corpus. To characterise such changes, here we first generate a language model – probability distribution over sequences of words [15]– per year. For this analysis we use the Katz back-off smoothing language model [14]. This model estimates the conditional probability of a word given the number of times such word has been seen in the past.

To analyse differences in language models between consecutive years we use the perplexity metric. Perplexity is commonly used in Natural Language Processing to evaluate how well a language model predicts an unseen test set [8]. To analyse changes in language patterns for consecutive years we : 1) obtained the language model for year t (lm_t) then; 2) we computed perplexity comparing lm_t to the unseen corpus at $t+1$.

Perplexity predicts word-error rate well when only in-domain training data is used, but poorly when out-of-domain text is added [8]. Figure 2, left, shows that for the Computer Science domain perplexity increases as time goes by. Therefore, language models representing language patterns trained in previous years provide poor predictions when tested on future datasets, indicating that language models can become outdated.

To analyse the impact of lexical innovation in language model changes, we perform a progressive analysis based on *lexical innovation* and *lexical adoption*. Let D_t be the collection of papers from corpus at year t. Let V_t be the vocabulary of D_t; we define a *lexical innovation* in D_t, LI_t, as the set of terms appearing in V_t, which were not mentioned in V_{t-1} [5]. We also define a *lexical adoption* in D_t, LA_t, as the set of terms appearing in LI_t which also appear in V_{t+1}. Figure 2, right, shows that while the number of novel words in Computer Science is high in consecutive years, only few of these words are adopted.

Based on these two metrics we introduce the ***linguistic progressiveness metric***, LP_t as the ratio of *lexical adoption* and *lexical innovation*, i.e., $LP_t = \frac{|LA_t|}{|LI_t|}$. The higher the adoption of innovative terms the more progressive the language used in a domain. In Figure 3, left, the data indicates that the Computer Science domain has had a tendency towards being linguistically progressive. The following subsection studies the impact of innovation and adoption on semantic concepts in temporally consecutive ontologies of a domain.

[5] Notice that we are following a one step memory approach, further historical data could be used in future research

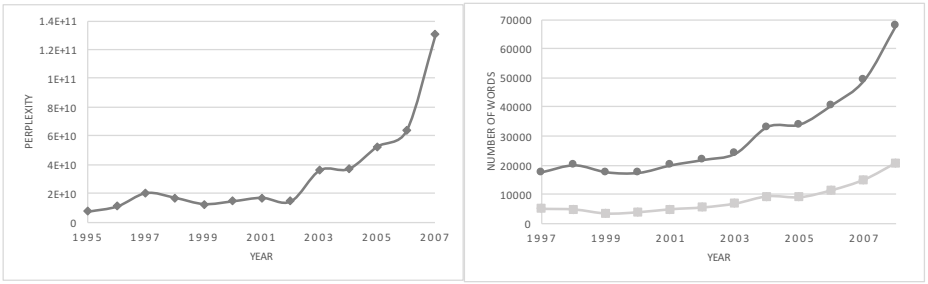

Fig. 2: From left to right, a) Language Models's perplexity per year; b) Number of new words per year (●), number of adopted words per year(■)

3.3 Semantic Progressiveness

Ontology evolution refers to the maintenance of an ontological structure by adapting such structure with new data from a domain [28]. Such adaptation can result in both the generation or expiration of an ontology's concepts and properties. Hence the introduction of new classes that better describe the conceptualisation of a domain can be considered to be a semantic innovation. In this subsection we analyse the introduction of new concepts to an ontological per consecutive year.

Let (D_t, O_t) represent a tuple where D_t is a collection of articles belonging to year t and O_t is the corresponding ontology representation computed with Klink-2 over the D_t collection. Let CI_t be the conceptual innovation in D_t, which we define as the set of concepts appearing in O_t, which were not mentioned in O_{t-1}. Also let CA_t be the conceptual adoption in D_t, which consists on the set of concepts in CI_t that also appear in O_{t+1}. Based on these definitions we introduce the ***semantic progressiveness metric***, CP_t, as the ratio of conceptual adoption and conceptual innovation, i.e., $CP_t = \frac{|CA_t|}{|CI_t|}$.

Figure 3, right, shows that the ontologies extracted for the Computer Science domain indicate a tendency to be less semantically progressive. A tendency towards a lower semantic progressiveness can be understood as a tendency towards having a more stable representation of the domain. Notice that the semantic progressiveness metric do not account for churn of semantic concepts but focuses only of innovation and adoption.

Both linguistic and semantic progressiveness characterise the rate of change on the language and semantic conceptualisations used in a research field over the years. This constant evolution of a scientific area motivates us to study the prediction of semantic concepts that will likely enhance the current semantic representation of a research domain. The following section introduces our proposed model for forecasting concepts appearing on an ontology based on historical data.

4 Framework for Forecasting Semantic Concepts based on Innovation-Adoption Priors

The proposed framework relies on the representation of an ontology's class as a topic word distribution. Learning topic models from text-rich structured data has been successfully used in the past [3, 10, 2]. Our proposed framework focuses on the task defined as

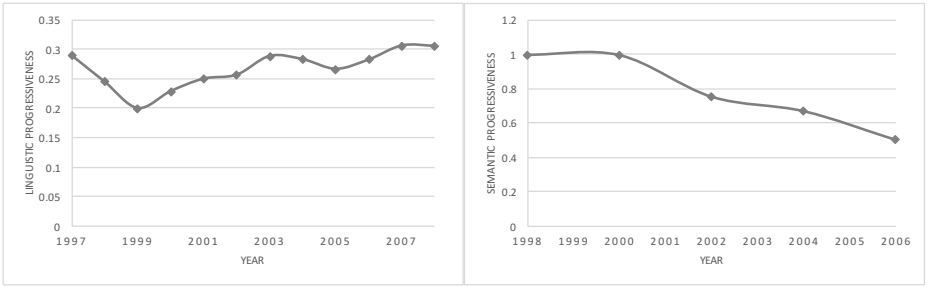

Fig. 3: From left to right, a) Linguistic progressiveness per year, b) Semantic progressiveness per year.

follows: *Given a set of documents at year t and a set of historical priors, forecast topic word distributions representing new concepts in the ontology* O_{t+1}.

The proposed framework breaks down into the following phases: 1) Predicting new semantic concepts with the Semantic Innovation Forecast (SIF) model; 2) Incorporating innovation priors; 3) Inferring topics with SIF; 4) Matching predicted topics to the forecast year's semantic concepts' gold standard

The overall pipeline is depicted in Figure 4.

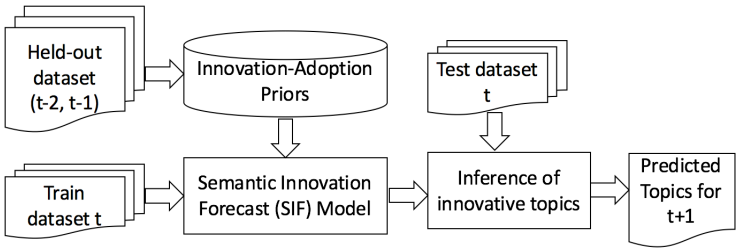

Fig. 4: Pipeline of the proposed framework for predicting semantic concepts using innovation/adoption priors.

4.1 Semantic Innovation Forecast (SIF) model

We propose a weakly-supervised approach for forecasting innovative concepts based on lexical innovation-adoption priors. We introduce the Semantic Innovation Forecast (SIF) model which forecasts future semantic concepts in the form of topic-word distributions. The proposed SIF model favours the generation of innovative topics by considering distributions that enclose innovative and adopted lexicons based on word priors computed from historical data.

Assume a corpora consisting of a collection of documents grouped by consecutive years. Let a corpus of documents written at year t be denoted as $D_t = \{d_1, d_2, ..d_{D_d}\}$.

Let each document be represented as a sequence of N_d words denoted by $(w_1, w_2, ..., w_{N_d})$; where each word in a document is an element from a vocabulary index of V_t.

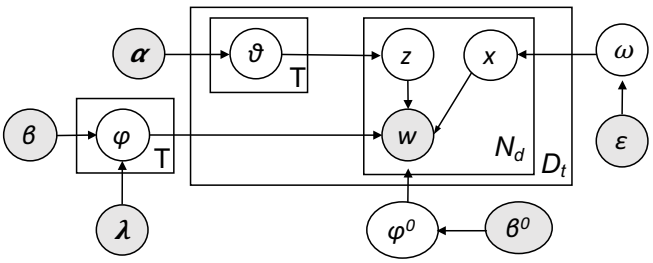

Fig. 5: Semantic Innovation Forecasting Model

We assume that when an author writes an article, she first decides whether the paper will be innovative or will conform to existing work. In the proposed generative model we consider that if a paper is innovative then a topic is drawn from an innovation specific topic distribution θ. In such case each word in the article is generated from either the background word distribution ϕ_0 or the multinomial word distribution for the innovation-related topics ϕ_z.

The generative process for SIF is as follows:

- Draw $\omega \sim$ Beta(ϵ), $\varphi^0 \sim$ Dirichlet(β^0), $\varphi \sim$ Dirichlet(β).
- For each topic z draw $\phi_z \sim$ Dirichlet$(\lambda \times \beta_z^T)$.
- For each document $m \in \{1..D\}$,
  - Choose $\theta_m \sim$ Dirichlet(α)
  - For each word $n \in \{1..N_d\}$ in document m,
    - draw $x_{m,n} \sim$ Bernoulli(ω);
    - if $x_{m,n} = 0$,
      - draw a word $w_{m,n} \sim$ Multinomial(φ^0);
    - if $x_{m,n} = 1$,
      - draw a topic $z_{m,n} \sim$ Multinomial(θ),
      - draw a word $w_{m,n} \sim$ Multinomial$(\varphi_{z_{m,n}})$.

The SIF model can be considered as an adaptation of a smoothed LDA [4], where we have added a per token latent random variable x which acts as a switch. If $x = 0$, words are generated from a background distribution, which accumulates words common to conformer articles. While if $x = 1$, words are sampled from the topic-specific multinomial ϕ_z. Moreover, SIF encodes word priors generated from historical data, such priors encapsulate innovation and adoption polarity in the matrix λ and are explained in more detail in the following Subsection.

4.2 Incorporating Innovation-Adoption Priors

Word priors enable us to have a preliminary or prior model of the language related to a topic of interest in the absence of any other information about this topic. A word prior

is a probability distribution that expresses one's belief about a word's relevance to, in this case, being characteristic of innovative topics, when no other information about it is provided. Since the aim is to discover new semantic concepts, we propose to use lexical innovation and lexical adoption as indicators of lexicons characterising innovative word distributions.

The procedure to generate such ***innovation-adoption priors*** is as follows; to compute priors for a SIF model at time t we make use of two vocabularies, the one at year $t-1$ and $t-2$. From these vocabularies we identify innovative (at $t-2$) and adopted (at $t-1$) lexicons as described in Subsection 3.2. The union of these lexicons constitute a vocabulary of size K. Then for each term $w \in \{1, ..K\}$ in this vocabulary we assign it a weight. We experimented with different weights and we found an optimum when assigning 0.7 if $w \in LI_{t-2}$ and 0.9 if $w \in LA_{t-1}$. This setting favours adoption over innovation since innovative words may not necessarily be embraced by the Computer Science community in the future. This weighted vocabulary constitutes the innovation priors λ.

Compared to the original LDA model [4] in SIF we have added a dependency link of ϕ on the vector λ of size K. Therefore we use innovation priors as supervised information and modify the topic-word Dirichlet priors for innovation classification.

4.3 SIF Inference

We use Collapsed Gibbs Sampling [12] to infer the model parameters and topic assignments for a corpus at year $t+1$ given observed documents at year t. Such sampling estimates empirically the target distribution. Let the index $t = (m, n)$ denote the n_{th} word in document m and let the subscript $-t$ denote a quantity which excludes data from the n_{th} word position in document m, the conditional posterior of x_t is:

$$P(x_t = 0|\mathbf{x}_{-t}, \mathbf{z}, \mathbf{w}, \beta^0, \epsilon) \propto \frac{\{N^0_m\}_{-t} + \epsilon}{\{N_m\}_{-t} + 2\epsilon} \times \frac{\{N^0_{w_t}\}_{-t} + \beta^0}{\sum_{w'}\{N_{w'}\}_{-t} + V\beta^0}, \quad (1)$$

where N^0_m denotes the number of words in document m assigned to the background component, N_m is the total number of words in document m, $N^0_{w_t}$ is the number of times word w_t is sampled from the background distribution.

$$P(x_t = 1|\mathbf{x}_{-t}, \mathbf{z}, \mathbf{w}, \beta, \epsilon) \propto \frac{\{N^s_m\}_{-t} + \epsilon}{\{N_m\}_{-t} + 2\epsilon} \times \frac{\{N^s_{w_t}\}_{-t} + \beta}{\sum_{w'}\{N_{w'}\}_{-t} + V\beta}, \quad (2)$$

where N^s_m denotes the number of words in document m sampled from the topic distribution, $N^s_{w_t}$ is the number of times word w_t is sampled from the topic specific distributions.

The conditional posterior for z_t is:

$$P(z_t = j|\mathbf{z}_{-t}, \mathbf{w}, \alpha, \beta) \propto \frac{N_{d,j}^{-t} + \alpha_j}{N_d^{-t} + \sum_j \alpha_j} \cdot \frac{N_{j,w_t}^{-t} + \beta}{N_j^{-t} + V\beta}, \quad (3)$$

where N_d is the total number of words in document d, $N_{d,j}$ is the number of times a word from document d has been associated with topic j, N_{j,w_t} is the number of times word w_t appeared in topic j, and N_j is the number of words assigned to topic j.

When the assignments have been computed for all latent variables, then we can estimate the model parameters $\{\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\varphi^0}, \boldsymbol{\omega}\}$. For our experiments we set the symmetric prior $\epsilon = 0.5$, $\beta_0 = \beta = 0.01$. We learn the asymmetric prior $\boldsymbol{\alpha}$ directly from the data using maximum-likelihood estimation [17] and updating this value every 40 iterations during the Gibbs sampling. In our experiments we run the sampler for 1000 iterations, stopping once the log-likelihood of the learning data has converged under the learning model.

5 Experimental Setup

Here we present the experimental set up used to assess the SIF framework. We evaluate the accuracy of SIF in a semantic-concept forecasting task.

We perform this task by applying our framework on the dataset described in Section 3.1. Each collection of documents per year is randomly partitioned into three independent subsets contains respectively 20%, 40% and 40% of the documents. For a given document collection at year t, the 20% partition represents a held-out dataset used to derive innovation priors (Dp_t) ; while the other two partitions represent the training ($Dtrain_t$) and testing sets($Dtest_t$).

5.1 Forecasting with SIF

To forecast semantic concepts for a corpus at year $t+1$, we assume no information from $t+1$ is known at the time of the forecast. We train a SIF model on year t with $Dtrain_t$ using innovative priors computed on the held-out datasets for the two previous years: Dp_{t-1} and Dp_{t-2}. Then using the trained model on year t we perform inference over $Dtest_t$ and consider this output to be the forecast for concepts aiming to match those in CI_{t+1} (concept innovation at $t+1$, see Subsection 3.3). The output of this last step is a set of topics that are effectively sets of word distributions, which we use to compare against our gold standard.

5.2 Gold Standard

We build our gold standard by generating a one-topic model per semantic-concept appearing in CI_{t+1}. This is performed by applying the standard LDA model [4] over the test dataset for documents belonging to each concept at year $t+1$.

Table 1 shows some examples of the gold standard computed for each innovative semantic concept of each year. The one-topic model representation of a semantic-concept provides a word distribution, which can be compared against the ones generated with SIF.

Year	Semantic Concept	Top 10 LDA words
2000	anthropomorph robot	robot, control, humanoid, human,anthropomorph,mechan, system ,design,skill,method
2002	context-free-grammar	languag,grammar, model, context-fre, system,algorithm,gener,method,show,paper
2004	video-stream	video, stream, network,rate,system,applic,adapt,bandwidth,packet,internet
2006	3d-reconstruct	reconstruct,imag,model,algorithm,structur,camera,point,surfac,data,base
2008	open-access	access,open,research,journal,repositori,publish,articl,develop,data,institut

Table 1: Examples of semantic concepts' gold-standard. For a given year, we present a semantic concept and an extract of the word distribution representing such concept. Each distribution is derived from a one-topic standard LDA model computed from documents belonging to such concept. Words are presented stemmed, weights assigned to each word are omitted in this example.

5.3 Baselines

We compare SIF against four baselines. For a year t forecasting for year $t+1$:

1. ***LDA Topics*** (LDA); referring to word distributions weighted by latent topics extracted from the training $Dtrain_t$. This setting makes no assumption over innovative/adopted lexicons. It outputs a collection of n topics per training set, which are compared against the gold standard.
2. ***LDA Innovative Topics*** (LDA-I); computes topics based on documents containing at least one word appearing in LI_t.
3. ***LDA Adopted Topics*** (LDA-A); computes topics based only on documents containing at least one word appearing in LA_t.
4. ***LDA Innovation/Adoption Topics*** (LDA-IA): this baseline filters documents based on words appearing λ_t.

Baselines 2-4 represent three strong baselines, which consider innovative and adopted lexicons.

5.4 Estimating the Effectiveness of SIF

To estimate the effectiveness of SIF we consider how similar the predicted semantic concepts for $t+1$ are from the reference gold standard concepts for that year. To this end we based the similarity scores using the cosine similarity metric [15]. This metric ranges from 0 (no similarity exists between compared vectors) to 1 (the compared vectors are identical), therefore scoring a similarity higher than 0.5 indicates that the compared vectors are similar.

To compute this similarity metric we used the word vector representation of a predicted topic and of the topics generated for that year's gold standard. Therefore when forecasting for $t+1$ we computed the cosine similarity between the predicted candidate topic x and each of the topic y in CI_{t+1}, keeping as matches the similar ones.

We evaluated the semantic concept forecast task as a ranked retrieval task, where the appropriate set of forecast concepts are given by the top retrieved topic distributions. To measure the effectiveness on this task we used the Mean Average Precision (MAP) metric [15], a standard metric for evaluating rank retrieval results. For our experiments we computed MAP@10 to measure the mean of precision scores obtained from the top 10 predicted topics ranked based on topic-word distributions. The higher the word weights assigned on a topic the higher in the rank the topic is within the set of predicted topics.

6 Experimental Results and Evaluation

In this section we report the experimental results obtained for the semantic concept forecasting task. SIF and LDA require defining the number of topics to extract before applying on the data [6]. For our experiments we considered a fixed number of 100 topics, making no assumption on the expected number of new concepts appearing on the forecast year. These 100 topics are ranked based on topic-word distributions. The evaluation is done over the top 10 forecast topics using MAP@10.

Results in all experiments are computed using 2-fold cross validation over 5 runs of different random splits of the data to evaluate results' significance. Statistical significance is done using the T-test. The evaluation consists in assessing the following:

1) Measure and compare SIF against the proposed baselines introduced in subsection 5.3.
2) Investigate whether the proposed SIF approach effectively forecasts future semantic concepts.

6.1 Semantic Concept Forecast Results

Table 2 presents MAP results for SIF and the four baselines. The first three columns of Table 2 shows: i) the year in which the model was trained; ii) the year from where the innovative priors were derived for that setting; iii) the year for which semantic concepts are forecast .

All baselines except LDA offer competitive results. LDA achieves a poor average result of 16% over the 5 forecast years. For the predictions of 2002 and 2004, LDA fails to generate concepts matching those from the gold standard. This is expected since LDA alone do not make assumptions over linguistic innovation and adoption, therefore it's unlikely that the LDA-based generated topic based on past data will predict future concepts. However, pre-filtering documents containing either innovative lexicons, adopted lexicons or both appear instead to have a positive effect in the forecasting task.

YEAR FORECAST	YEAR TRAINED	YEAR PRIOR	SIF	LDA	LDA-A	LDA-I	LDA-IA
2000	1999	1997-1999	0.7031	0.125	0.4761	0	0.408
2002	2001	1999-2001	0.8750	0	0.8227	0.6428	0.7486
2004	2003	2001-2003	0.9060	0	0.5822	0.5726	0.6347
2006	2005	2003-2005	0.8755	0.3069	0.7853	0.8385	0.6893
2008	2007	2005-2006	0.988	0.398	0.681	0.5661	0.7035
AVG			**0.8695**$\star$	0.1659	0.6694	0.524	0.6368

Table 2: MAP@10 for SIF and baselines. The number of topics is set to 100 for all five models. The value highlighted in bold corresponds to the best results obtained in MAP@10. A $\star$ denotes that the MAP@10 of SIF significantly outperforms the baselines. Significance levels: $p - value < 0.01$.

[6] The data generated in the evaluation are available on request at http://technologies.kmi.open.ac.uk/rexplore/ekaw2016/OF/

In particular, the use of LDA-A over LDA-I gives a boost on MAP of 14.54%, indicating that *adopted words features are better predictors of innovative semantic concepts*. LDA-A also improves in average upon the LDA-IA baseline with a boost of 3%. The proposed SIF model however outperforms significantly all four baselines with an average boost: over LDA of 70%; over LDA-A of 20%; over LDA-I of 34%; over LDA-IA of 23% (significant at $p < 0.01$). We could have expected LDA-IA to achieve closer results to SIF, since it is computed on documents filtered using both innovative and adopted lexicons. However, LDA-IA do not assign any preference over distributions of words containing either of such lexicons. In contrast, SIF takes innovation priors as a weighting strategy to build a prior model of language which is potentially used in future semantic concepts. The model is learnt over the full training set allowing to make use of both documents containing innovative and adopted lexicons and otherwise. The above results show the effectiveness of SIF for semantic concept forecasting over the baselines.

Table 3 presents examples of SIF's predicted topics that obtained a match in the forecast year's gold-standard (GS). While SIF do not forecast a specific name for the new semantic concept, the information provided by the word distribution gives context to the predicted concept. Table 3 presents top 10 words for the forecast SIF and GS representation however similarity computations where made using the whole topic-word representations. When comparing the SIF prediction vs the GSs we observe very close matches in 2000-2006 while for 2008 it is interesting to observe the appearance of words such as islam, victim, terror which don't match the top 10 of the corresponding GS (notice however they may appear in the further topic-word representation of the GS), however the word hate within the GS gives a insight of the use of mechatronics in violence-related scenarios.

2000		2002		2004		2006		2008	
Wireless Network		Asynchronous Transfer Mode		Image Threedimension		Cryptography		Mechatronics	
SIF	GS	SIF	GS	SIF	GS	SIF	GS	SIF	GS
control	control	network	network	activ	model	method	model	robot	robot
system	system	service	servic	function	algorithm	structur	method	model	model
propos	propos	system	applic	show	function	data	algorithm	base	propos
network	applic	mobil	system	result	data	protocol	system	perform	simul
servic	network	protocol	mobil	image	result	secur	data	simul	process
data	servic	wireless	protocol	respons	image	inform	process	islam	mechan
time	commun	rout	base	effect	measure	signatur	scheme	time	control
perform	compu	perform	perform	patient	cell	authenti	user	control	applic
distribut	manag	packet	algorithm	clinic	structure	detec	protocol	applic	dynam
traffic	schem	control	packet	visual	patient	attack	secur	victim	hate
protocol	mobil	scheme	control	brain	surfac	sequenc	inform	terror	best

Table 3: Examples of semantic concepts forecast with SIF for each year. The second row describes the semantic concept matching the predicted topic obtained with SIF. SIF columns presents top 10 words extracted from the word distribution of the SIF topic prediction. GS columns present top 10 words extracted from the one-topic LDA distribution.

7 Conclusions and Future Work

This work focused on the task of *semantic concept forecasting*, which aims at predicting classes which will be added to an ontology at time $t+1$ when only information up to time

t is available. To approach this task we proposed the concepts of linguistic and semantic progressiveness, and introduced a strategy to encode lexical innovation and adoption as innovation priors. Based on these concepts we introduced the Semantic Innovation Forecast Model (SIF), which is a generative approach relying on historical innovation priors for the prediction of word distributions characterising a semantic concept.

In SIF each semantic concept is represented as a distribution of words obtained from the one-topic model of the collection of documents belonging to such concept. To this end we applied the proposed approach on a very large dataset belonging to the Computer Science domain, consisting of over one million papers on the course of 14 years. Our data analysis included the introduction of two novel metrics namely the linguistic and semantic progressiveness; which gave insights on the semantic trends in the Computer Science domain. Our experiments indicate that adopted lexicon are better predictors for semantic classes. Our experimental results also proof that the proposed approach is useful for the innovative semantic concept forecasting task. The SIF model outperforms the best baseline LDA-A showing an average significant boost of 23%.

To the best of our knowledge this is the first approach to address the ontology forecasting task in general and in particular the first one in addressing the prediction of new semantic concepts. We believe that research on the prediction of semantic concepts in particular and in general the forecast of changes in an ontology can be beneficial to different areas of research not limited to the study of scholarly data. For the future, we plan to keep working on the integration between explicit and latent semantics, improve further the performance of our approach and introduce graph-structure information into the model. We also intend to use this approach for detecting innovative authors and forecast topic trends.

Acknowledgements We would like to thank Elsevier BV and Springer DE for providing us with access to their large repositories of scholarly data.

References

1. A. Ahmed and E. Xing. Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. *Uncertainty in Artificial Intelligence*, 2010.
2. D. Andrzejewski, X. Zhu, M. Craven, and B. Recht. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pages 1171–1177. AAAI Press, 2011.
3. V. Bicer, T. Tran, Y. Ma, and R. Studer. Trm — learning dependencies between text and structure with topical relational models. In *Proceedings of the 12th International Semantic Web Conference - Part I*, ISWC '13, pages 1–16, New York, NY, USA, 2013. Springer-Verlag New York, Inc.
4. N. A. Y. Blei, D. M. and M. I. Jordan. Latent dirichlet allocation. In *J. Mach. Learn. Res. 3*, pages 993–1022, 2003.
5. L. Bolelli, c. Ertekin, and C. L. Giles. Topic and trend detection in text collections using latent dirichlet allocation. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 776–780. Springer-Verlag, 2009.
6. L. Bolelli, S. Ertekin, D. Zhou, and C. L. Giles. Finding topic trends in digital libraries. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '09, pages 69–72, New York, NY, USA, 2009. ACM.

7. R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *EACL*, volume 6, pages 9–16, 2006.
8. S. Chen, D. Beeferman, and R. Rosenfeld. Evaluation metrics for language models. 1998.
9. C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts. No country for old members: user lifecycle and linguistic change in online communities. In *In Proceedings of the 22nd international conference on World Wide Web, WWW 13*, page 307318, 2013.
10. H. Deng, J. Han, B. Zhao, Y. Yu, and C. X. Lin. Probabilistic topic models with biased propagation on heterogeneous information networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1271–1279, New York, NY, USA, 2011. ACM.
11. A. Gohr, A. Hinneburg, R. Schult, and M. Spiliopoulou. Topic evolution in a stream of documents. In *In SDM*, pages 859–872, 2009.
12. T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):52285235, 2004.
13. Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles. Detecting topic evolution in scientific literature: How can citations help? In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 957–966, New York, NY, USA, 2009. ACM.
14. S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *In IEEE Trans. on Acoustics, Speech, and Signal Processing*, 1987.
15. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
16. Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 198–207. ACM, 2005.
17. T. Minka. Estimating a Dirichlet distribution. Technical report, 2003.
18. F. Monaghan, G. Bordea, K. Samp, and P. Buitelaar. Exploring your research: Sprinkling some saffron on semantic web dog food. In *Semantic Web Challenge at the International Semantic Web Conference*, volume 117, pages 420–435. Citeseer, 2010.
19. S. Morinaga and K. Yamanishi. Tracking dynamics of topic trends using a finite mixture model. In *10th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.
20. F. Osborne and E. Motta. Klink-2: integrating multiple web sources to generate semantic topic networks. In *The 14th International Semantic Web Conference*, 2015.
21. F. Osborne, E. Motta, and P. Mulholland. Exploring scholarly data with rexplore. In *The 13th International Semantic Web Conference*, 2013.
22. F. Osborne, A. Salatino, A. Birukou, and E. Motta. Automatic classification of springer nature proceedings with smart topic miner. 2016.
23. C. Pesquita and F. M. Couto. Predicting the extension of biomedical ontologies. *PLoS Comput Biol*, 8(9):e1002630, 2012.
24. M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.
25. Y.-H. Tseng, Y.-I. Lin, Y.-Y. Lee, W.-C. Hung, and C.-H. Lee. A comparison of methods for detecting hot topics. *Scientometrics*, 81(1):73–90, 2009.
26. H. Wang, T. Tudorache, D. Dou, N. F. Noy, and M. A. Musen. Analysis and prediction of user editing patterns in ontology development projects. *Journal on data semantics*, 4(2):117–132, 2015.
27. P. Willett. The porter stemming algorithm: then and now. *Program*, 40(3):219–223, 2006.
28. F. Zablith, G. Antoniou, M. d'Aquin, G. Flouris, H. Kondylakis, E. Motta, D. Plexousakis, and M. Sabou. Ontology evolution: a process-centric survey.