Hierarchical Dirichlet Process for Tracking Complex Topical Structure Evolution and Its Application to Autism Research Literature

Adham Beykikhoshk(✉), Ognjen Arandjelović,
Svetha Venkatesh, and Dinh Phung

Pattern Recognition and Data Analytics Centre, Deakin University,
Geelong, Australia
{abeyki,ognjen.arandjelovic,svetha.venkatesh,dinh.phung}@deakin.edu.au

Abstract. In this paper we describe a novel framework for the discovery of the topical content of a data corpus, and the tracking of its complex structural changes across the temporal dimension. In contrast to previous work our model does not impose a prior on the rate at which documents are added to the corpus nor does it adopt the Markovian assumption which overly restricts the type of changes that the model can capture. Our key technical contribution is a framework based on (i) discretization of time into epochs, (ii) epoch-wise topic discovery using a hierarchical Dirichlet process-based model, and (iii) a temporal similarity graph which allows for the modelling of complex topic changes: emergence and disappearance, evolution, splitting and merging. The power of the proposed framework is demonstrated on the medical literature corpus concerned with the autism spectrum disorder (ASD) – an increasingly important research subject of significant social and healthcare importance. In addition to the collected ASD literature corpus which we made freely available, our contributions also include two free online tools we built as aids to ASD researchers. These can be used for semantically meaningful navigation and searching, as well as knowledge discovery from this large and rapidly growing corpus of literature.

1 Introduction

The Autism Spectrum Disorder (ASD) is a life-long neurodevelopmental disorder with poorly understood causes on the one hand, and a wide range of potential treatments supported by little evidence on the other. The disorder is characterized by severe impairments in social interaction, communication, and in some cases cognitive abilities. Considering the social and economic burden of ASD it is unsurprising that it has been attracting an increasing amount of research attention which has resulted in a rapid growth of the relevant corpus of literature. Navigating this vast amount of data by conventional, manual means is difficult and limiting. Consequently, the potential benefit of tools based on novel data-mining and machine learning techniques is immense [1]. More meaningful ways for visualising or searching for data could provide invaluable information in

T. Cao et al. (Eds.): PAKDD 2015, Part I, LNAI 9077, pp. 550–562, 2015.
DOI: 10.1007/978-3-319-18038-0_43

clinical and administrative decision making as well as aid research, while automatic knowledge discovery would in its own right advance the understanding of the underlying phenomena (e.g. epidemiological patterns). In the present paper we describe a novel method which contributes towards this goal.

More specifically, we describe a general framework for the analysis of medical literature capable of (i) discovering the underlying topical structure, (ii) inferring the relationships between different discovered topics, and (iii) tracking the evolution of topics over time. The proposed framework uses hierarchical Dirichlet process (HDP) to extract topics automatically, and then constructs a similarity graph over them using an inter-topic similarity measure; topic evolution over time can be inferred from this graph. The effectiveness of our approach is demonstrated on the specific example of a large longitudinal data corpus of medical literature on ASD which we collected. This corpus includes more than 18,000 articles published over the course of 42 years. Another contribution is this corpus which is made publicly available.

The results we report on the collected ASD literature corpus illustrate the usefulness of our method and its ability to extract and track over time abstract topical knowledge, inferring the point at which a certain topic comes into existence, how its evolves, splits into multiple new topics or merges with the existing ones, and lastly when it ceases to exist. This is demonstrated on examples of well-known research directions in the field. Our additional contributions come in the form of two free online tools which allow researchers to (i) navigate and search the literature in a semantically meaningful manner (see www.undersdtanfigutism.tk), and (ii) understand the development and relationships between different ideas which permeate research in the domain of ASD (see http://goo.gl/Ws7V64).

2 Previous Work

In this section we review the most relevant previous work on topic modelling. We focus our attention first on latent topic models which have dominated the field in the last decade, and then on biomedical text mining, given the application domain in which our framework in evaluated in Section 4.

2.1 Latent Topic Models

An important early approach is the latent semantic indexing (LSI) [2] which remains popular. Two notable limitations of LSI are its inability to deal effectively with polysemy and to produce an explicit description of the latent space. A probabilistic improvement of LSI [3] overcomes these by explicitly characterizing the latent space with semantic topics, and by employing a probabilistic generative model that addresses the polysemy problem. Nevertheless, probabilistic LSI is prone to parameter overfitting caused by an uncontrolled growth in the number of parameters as the document corpus is increased. In addition, the necessary assignment of probabilities to documents is a nontrivial task [4].

The recently proposed latent Dirichlet allocation (LDA) method [4] overcomes the overfitting problem by adopting a Bayesian framework and a generative process at document level. While LDA has quickly become a standard tool for topic modelling, it too experiences challenges when applied on real-world data. In particular, being a parametric model the number of desired output topics has to be specified in advance. The HDP model as the nonparametric counterpart of LDA was introduced by Teh *et al.* [5] and addressed this limitation by using a Dirichlet process (DP) (as opposed to a Dirichlet distribution) as the prior on topics. Therefore, each document is modelled using an infinite mixture model, allowing the data to inform the complexity of the model and infer the number of resulting topics automatically. We discuss this model in further detail in Section 3.

Temporal Topic Modelling: A notable limitation of most models described in the previous section lies in their assumption that the data corpus is static. However, in many practical applications documents are added to the corpus in a temporal manner. Therefore their ordering has significance and at best they might be exchangeable in short time slices. As a consequence, the topical structure of the corpus changes over time. Existing work can be divided into two groups.

First, the models that hold a Markov assumption over time by discretizing and dividing it into multiple *epochs*. Then a topic model is fit to each epoch where the parameters of adjacent models are tied together [6–9]. Whilst they capture how the comprising words of a topic evolve over time, they assume the data arrives in a uniform fashion whereas in our application documents may arrive at irregular time intervals. Indeed we adopt the time desensitization from this group. However our approach diverges from those in the current literature thereafter. We do not consider the Markov assumption to obtain a model with less complexity and easier inference. Second, the models that treat the document time-stamps as an observed continuous random variable [10,11]. These models are capable of modelling the life span of a topic, but not the capturing its evolution and trajectory (i.e. split and merge). The topic model used in both groups can be parametric [6,7,10] or nonparametric [8,9]. Parametric models will still suffer from the same problem as LDA in requiring the number of topics to be specified in advance.

2.2 Biomedical Text Mining

The idea that the medical literature could be mined for new knowledge is typically attributed to Swanson [12]. For example by manually examining medical literature databases he hypothesised that dietary fish oil could be beneficial for Raynaud's syndrome patients, which was later confirmed by experimental evidence. Work that followed sought to develop statistical methods which would make this process automatic. Previous work on biomedical text mining has rather focused on (i) the tagging of names of entities such as genes, proteins,

and diseases [13], (ii) the discovery of relationships between different entities e.g. functional associations between genes [14], or (iii) the extraction of information pertaining to events such as gene expression or protein binding [15].

Most existing work on biomedical knowledge discovery is based on what may be described as traditional data mining techniques (neural networks, support vector machines etc); comprehensive surveys can be found in [15,16]. The application of state-of-the-art Bayesian methods in this domain is scarce. Amongst the notable exceptions is the work by Blei *et al.* who showed how latent Dirichlet allocation (LDA) can be used to facilitate the process of hypothesis generation in the context of genetics [17]. Arnold *et al.* used a similar approach to demonstrate that abstract topic space representation is effective in patient-specific case retrieval [18]. In their later work they introduced a temporal model which learns topic trends and showed that the inferred topics and their temporal patterns correlate with valid clinical events and their sequences [19]. Wu *et al.* used LDA for gene-drug relationship ranking [20].

3 Proposed Framework

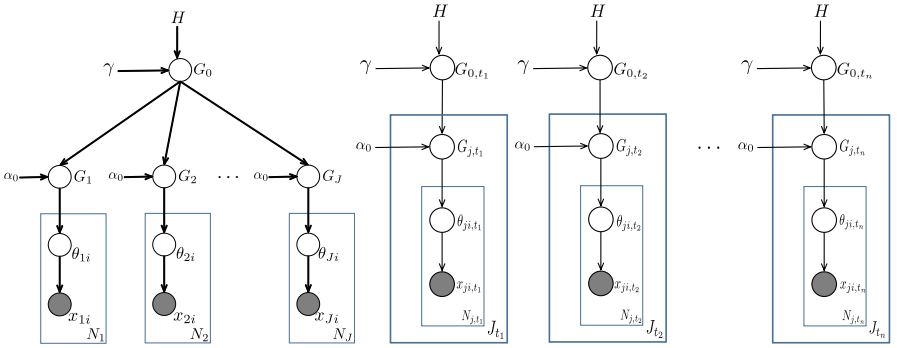
We begin this section by reviewing the relevant theory underlying HDP mixture modelling which plays the central rule in the proposed framework. Then we turn our attention to the main technical contribution of our work and explain how the HDP is employed to discover the topical content of a literature corpus and track its structural changes over time.

3.1 Hierarchical Dirichlet Process Mixture Models

Dirichlet process as the building block of Bayesian non-parametric methods allows the document collection to accommodate potentially infinite number of topics. A Dirichlet process [21] $\mathrm{DP}(\gamma, H)$ is defined as a distribution of a random probability measure G over a measure space $(\Theta, \mathcal{B}, \mu)$, such that for any finite measurable partition $(A_1, A_2, \ldots, A_r)$ of Θ the random vector $(G(A_1), \ldots, G(A_r))$ is a Dirichlet distribution with parameters $(\gamma H(A_1), \ldots, \gamma H(A_r))$. An alternative view of the DP emerges from the so-called stick-breaking process which adopts a constructive approach using a sequence of discrete draws [22]. Specifically, if $G \sim \mathrm{DP}(\gamma, H)$ then $G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$ where $\phi_k \overset{iid}{\sim} H$ and $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty}$ is the vector of weights obtained by the stick-breaking process that is $\beta_k = v_k \prod_{l=1}^{k-1}(1 - v_l)$ and $v_l \overset{iid}{\sim} \mathrm{Beta}(1, \gamma)$.

Owing to the discrete nature and infinite dimensionality of its draws, the DP is a highly useful prior for Bayesian mixture models. By associating different mixture components with atoms ϕ_k of the stick-breaking process, and assuming $x_i|\phi_k \overset{iid}{\sim} F(x_i|\phi_k)$ where $F(.)$ is the likelihood kernel of the mixing components, we can formulate the Dirichlet process mixture model (DPM). The DPM is suitable for nonparametric clustering of exchangeable data in a single group e.g. words in a document where the DPM models the underlying structure of

the document with potentially an infinite number of topics. However, many real-world problems are more appropriately modelled as comprising multiple groups of exchangeable data (e.g. a collection of documents). In such cases it is usually desirable to model the observations of different groups jointly, allowing them to share their generative clusters to remain linked. This idea is known as the "sharing statistical strength" and it is naturally obtained by hierarchical architecture in Bayesian modelling.



(a) HDP **(b)** Proposed

Fig. 1. (a) Graphical model representation of HDP. Each box represents one document whose observed data (words) is shown shaded. Unshaded nodes represent latent variables. An observed datum x_{ji} is assigned to a latent mixture component parameterized by θ_{ji}. γ and α are the concentration parameters and H is the corpus-level base measure. (b) Graphical model representation of the proposed framework. The corpus is temporally divided into t_n epochs and each epoch modelled using an HDP (outer boxes). Different epochs' HDPs share their corpus-level DP and hyperparameters.

Amongst different ways of linking group-level DPMs, HDP [5] offers an interesting solution whereby base measures of group-level DPs are drawn from a corpus-level DP. In this way the atoms of the corpus-level DP (i.e. topics in our case) are shared across the documents. Formally, if $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_J\}$ is a document collection where $\mathbf{x}_j = \left\{x_{j1}, \ldots, x_{jN_j}\right\}$ is the j-th document comprising N_j words, each document is modelled with a DPM $G_j|\alpha_0, G_0 \overset{iid}{\sim} \mathrm{DP}(\alpha_0, G_0)$ where its DP prior is further endowed by another DP $G_0|\gamma, H \sim \mathrm{DP}(\gamma, H)$. This is illustrated schematically in Figure 1a. Since the base measure of G_j is drawn from G_0, it takes the same support as G_0. Also the parameters of the group-level mixture components, θ_{ji}, share their values with the corpus-level DP support on $\{\phi_1, \phi_2, \ldots\}$. Therefore G_j can be equivalently expressed using the stick-breaking process as $G_j = \sum_{k=1}^{\infty} \pi_{jk}\delta_{\phi_k}$ where $\boldsymbol{\pi}_j|\alpha_0, \gamma \sim \mathrm{DP}(\alpha_0, \gamma)$[5]. The posterior for θ_{ji} has been shown to follow a Chinese restaurant franchise process which can be used to develop inference algorithms based on Gibbs sampling [5].

3.2 Modelling Topic Evolution Over Time

In this section we show how the described HDP-based model can be applied to the analysis of temporal topic changes in a longitudinal data corpus. We begin by dividing the literature corpus by time into multiple *epochs*. Each epoch is then modelled separately using an HDP. Different epochs' models share their hyperparameters and the corpus-level base measure. Hence if n is the number of epochs, we obtain n sets of topics $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{t_1}, \ldots, \boldsymbol{\theta}_{t_n}\}$ where $\boldsymbol{\theta}_t = \{\theta_{1,t}, \ldots, \theta_{K_t,t}\}$ is the set of topics that describe epoch t, and K_t their number (which is inferred automatically, as described previously). This is illustrated in Figure 1b. In the next section we describe how given an inter-topic similarity measure the evolution of different topics across epochs can be tracked.

3.3 Measuring Topics Similarity

Our goal now is to track changes in the topical structure of a data corpus over time. The simplest changes of interest include the emergence of new topics, and the disappearance of others. More subtly, we are also interested in how a specific topic changes – how it evolves over time in terms of the contributions of different words it comprises, as well as how it splits into new topics or merges with the existing ones. Clearly this information can provide valuable insight into the refinement of ideas and findings in the scientific community, effected by new research and accumulating evidence.

The key idea behind our approach stems from the observation that while topics may change significantly over time, by their very nature their change between successive epochs is limited. Therefore we infer the continuity of a topic in one epoch by relating it to all topics in the immediately subsequent epoch which are sufficiently similar to it under some similarity measure. This can be seen to lead naturally to a similarity graph representation whose nodes correspond to topics and whose edges link those topics in two epochs which are related. Formally, the weight of the directed edge that links $\phi_{j,t}$, the j-th topic in epoch t, and $\phi_{k,t+1}$ is set equal to $\rho(\phi_{j,t}, \phi_{k,t+1})$ where ρ is an appropriate similarity measure. Given that in our HDP-based model each topic is represented by a probability distribution, suitable similarity metrics include the Jaccard similarity, the Jenson-Shannon divergence, and the L_2-norm.

A conceptual illustration of a similarity graph is shown in Figure 2a. It shows three consecutive time epochs $t-1, t$, and $t+1$ and a selection of topics in these epochs. Graph edge weight i.e. inter-topic similarity is encoded by varying the thickness of the corresponding line connecting two nodes – a thicker line signifies more similar topics. We use a threshold to eliminate automatically weak edges, retaining only the edges which correspond to sufficiently similar topics in adjacent epochs. It can be seen that this readily allows us to detect the disappearance of a particular topic, the emergence of new topics, as well as the splitting or merging of different topics:

Emergence If a node does not have any edges incident to it, the corresponding topic is taken as having emerged in the associated epoch (e.g. ϕ_{j+2} at time t in Figure 2a).

Disappearance If no edges originate from a node, the corresponding topic is taken to vanish in the associated epoch (e.g. ϕ_j at time t in Figure 2a).

Splitting If more than a single edge originates from a node, the corresponding topic is understood as being split into multiple topics in the next epoch (e.g. ϕ_i is split into ϕ_j and ϕ_{j+1} in Figure 2a).

Merging If more than a single edge is incident to a node, the topics of the nodes from which the edges originate are understood as having merged together to form a new topic (e.g. ϕ_i and ϕ_{i+1} merge to form ϕ_{j+1} in Figure 2a).

4 Experimental Evaluation

Having introduced the main technical contribution of our work we now illustrate its usefulness on the example of ASD literature analysis, and describe additional contributions in the form of two free online tools that we developed to aid ASD researchers.

4.1 Data Collection

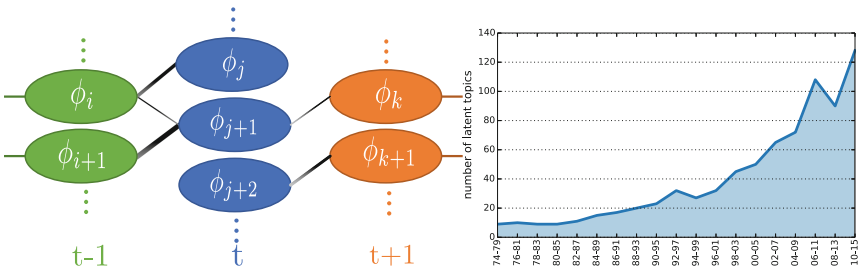
fe To the best of our knowledge there are no publicly available corpora of ASD-related medical literature. Hence we collected a comprehensive dataset ourselves that we describe its collection methodology and the pre-processing of data we performed to extract standard features used for text analysis.

Raw Data Collection: We used the PubMed search engine that allows users to access the United States National Library of Medicine for abstracts and references of life science and biomedical scholarly articles. We assumed a paper is related to ASD if the term "autism" is present in its title or abstract, and collected only papers written in English. The earliest publication fitting our criteria is that by Kanner [23], and we collected all matching publications up to the final one indexed by PubMed on 24th July 2014, yielding a corpus of 20,138 publications. We discarded the 1,946 which do not have an abstract indexed, ending with the total of 18,192 papers in our dataset. We used the abstracts text to evaluate our method.

Data Pre-processing: Following the standard practice in text processing literature we applied soft lemmatization on the abstracts in our dataset, using the freely available WordNet tool [24]. No stemming was performed to avoid potential distortion of words which is sometimes effected by heuristic rules used by stemming algorithms. After lemmatization and the removal of so-called stop words, we obtained 1.9 million terms in the entire corpus when repetitions are counted, and 37,278 unique terms. We construct the vocabulary for our method by selecting the subset of the most frequent unique terms which explain 90% of the energy of the corpus, which resulted in a 3,738 term vocabulary.

4.2 Proposed Method Implementation

We divided the 42 year timespan of our data corpus into overlapping five year epochs, with a two year lag between consecutive epochs, resulting in 18 epochs in total. The topics of each epoch were then extracted as described in Section 3.2 and their dynamics inferred as per Section 3.3. The number of latent topics of different epoch is plotted in Figure 2b. Notice the exponential rise in the number of topics which mirrors the exponential increase in the number of publications over time in our dataset. This increasing interest in ASD can be illustrated by the observation that in 2013 there are five times as many publications as in 2000. For our inter-topic similarity described in Section 3.3 we adopted the use of the well-known Jaccard similarity; this similarity measure was used to obtain all results reported in this section. Lastly, Gibbs sampling was used for HDP inference, implemented in Python 2.7, with hyperparameter resampling as described by Teh *et al.* [5].



(a) Topic similarity graph **(b)** Number of topics per epoch

Fig. 2. (a) Conceptual illustration of the proposed similarity graph that models topic dynamics over time. A node corresponds to a topic in a specific epoch; edge weights are equal to the corresponding topic similarities. (b) As the document corpus grows so does the number of topics needed to model its latent structure.

4.3 Case Study 1: ASD and Genetics

While the exact aetiology of the ASD is still poorly understood, the existence of a significant genetic component is beyond doubt [25]. Work on understanding complex genetic factors affecting the development of autism, which possibly involve multiple genes which interact with each other and the environment, is a major theme of research and as such a good case study on which the usefulness of the proposed method can be illustrated.

We started by identifying the topic of interest as that with the highest probability of the terms "gene" or "genetic" conditioned on the topic, and tracing it back in time to the epoch in which it originated. This led to the discovery of the relevant topic in the epoch spanning the period 1986–1991. Figure 4 shows the evolution of this topic from 1992 revealed by our method (due to space constraints only the most significant parts of the similarity graph are shown;

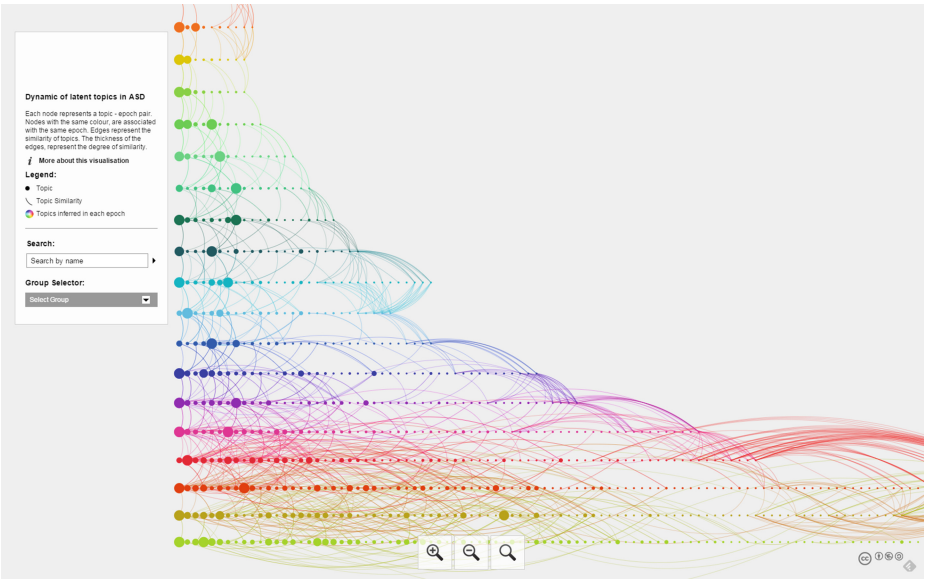


Fig. 3. Interactive similarity graph analysis tool (see http://goo.gl/Ws7V64). Word clouds of a few topics are shown for illustration. Nodes and links between them represent respectively topics in particular epochs and their similarities.

minor changes to the topic before 1992 are also omitted for clarity, as indicated by the dotted line in the figure). Each topic is labelled with its first few dominant terms. The following interpretation of our findings is readily apparent. Firstly, in the period 1992–1997, the topic is rather general in nature. Over time it evolves and splits into topics which concern more specific concepts (recall that such splitting of topics cannot be captured by any of the existing methods). For example by the epoch 2002–2007 the single original topic has evolved and split into four topics which concern:

- the relationship between mutations in the gene `mecp2` (essential for normal functioning of neurone), and mental disorders and epilepsy (it is estimated that one third of ASD individuals also have epilepsy),
- gene alternations, for example the duplication of `15q11--13` and deletion of `16p11.2` both of which are associated with ASD,
- genetic linkage association analysis and heritability of autism, and
- observational work on autistic twins and probands with siblings on the spectrum.

Our framework also allows us to look 'back' in time. For example, by examining the topics that the 1992 genetics topic originate from we discovered that the topic evolved from the early concept of "infantile ASD" (originated by Kanner [23]).

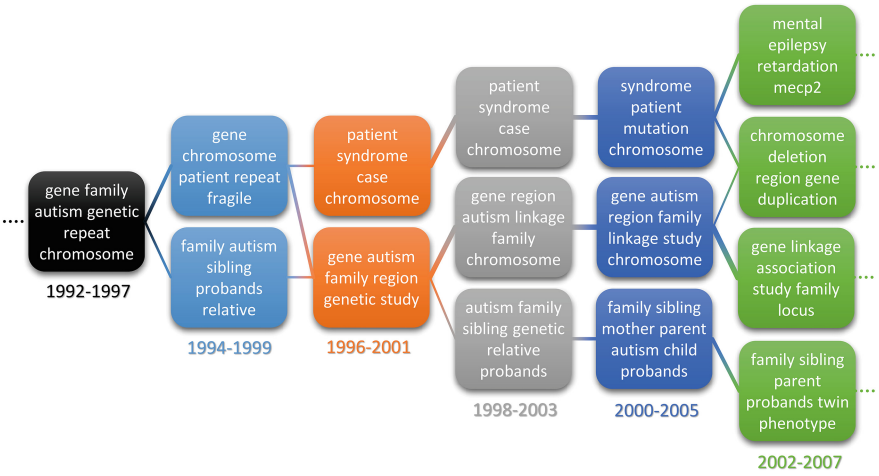


Fig. 4. Dynamics of the topic most closely associated with the concept of "genetics". A few dominant words are shown for each topic (shaded boxes).

4.4 Case Study 2: ASD and Vaccination

For our second case study we chose to examine research on the relationship between ASD development and vaccination. This subject has attracted much attention both in the research community, as well as in the media and the general public. The controversy was created with the publication of the work by Wakefield [26] which reported epidemiological findings linking MMR vaccination and the development of autism and colitis. Despite the full retraction of the article following the discovery that it was fraudulent, and numerous subsequent studies who failed to show the claimed link, a significant portion of the general public remains concerned with the issue.

As in the previous example, we begun by identifying the topic with the highest probability of the terms "vaccine" and "vaccination" conditioned on the topic, and tracing it back to the epoch in which it first emerged. Again, a single topic was readily identified, in the epoch spanning the period 1996–2001. Notice that this is consistent with the publication date of the first relevant publication by Wakefield [26]. The evolution of the topic is illustrated in Figure 5 in the same way as in the previous section. It can be seen that the original topic concerned the subjects initially brought to attention such as "measles", "vaccine", and "autism". In the subsequent epoch, when the original claim was still thought to have credibility, the topic evolves and splits into numerous others mirroring research directions taken by various researchers. Following this period and the revelations of its fraudulence, the topic assumes mainly single-threaded evolution, at times incorporating various originally separate ideas. For example observe the independent emergence of the term "mercury". Though initially unrelated to it this topic merges with the topic that concerns vaccination which can be explained by the widely publicized thiomersal (vaccine preservative)

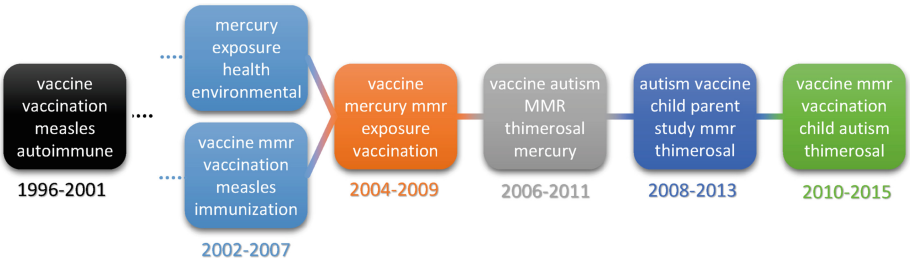


Fig. 5. Dynamics of the topic most closely associated with the concept of "vaccination". Notwithstanding the rejection of any link between vaccination and autism, this topic remains active albeit in a form which evolved over time.

controversy (again note that such merging of topics cannot be captured by the existing methods). Although rejected by the medical community due to a lack of evidence, this topic can be seen as persisting to date.

4.5 Topic Browser

A topic model can be seen as a dimensionality reduction framework that reduces documents into a topic space. This transformation of data can provide powerful insight and allow for the browsing of documents in a more subject-specific, semantic manner. For example by describing documents in the topic space, documents most related to a particular topic of interest can be readily identified and retrieved. To provide this functionality to the research community interested in ASD we used the framework described in this paper to model the entire literature corpus we collected, and built a website to facilitate free and ready use of our model and data. Researchers can use our online tool to browse topics, annotate them, and navigate through publications by topic. The website is available at http://www.understandingautism.tk.

5 Conclusions

We described a novel framework for temporal modelling of the topical structure of a longitudinal document corpus. Our approach consists of discretizing time into overlapping epochs, modelling the static topic structure within each epoch using an HDP, and tracking the evolution of topics over time using an inter-topic similarity measure. The resultant similarity graph captures relationships between topics in different epochs and allows for the automatic inference of the time of emergence and disappearance of topics, their evolution over time, merging and splitting. The power of the proposed general framework was demonstrated on the example of ASD-related medical literature. On two case studies which concern two important research issues in ASD literature we demonstrated that our method extracts meaningful topics and their temporal changes. A novel data corpus and free online tools are made freely available to researchers.

References

1. Beykikhoshk, A., Arandjelovic, O., Phung, D., Venkatesh, S., Caelli, T.: Data-mining twitter and the autism spectrum disorder: A pilot study (2014)
2. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. JASIS **41**, 391–407 (1990)
3. Hofmann, T.: Probabilistic latent semantic indexing. SIGIR, 50–57 (1999)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. JMLR **3**, 993–1022 (2003)
5. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. Journal of the American Statistical Association 101 (2006)
6. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: ICML, pp. 113–120 (2006)
7. Wang, C., Blei, D., Heckerman, D.: Continuous time dynamic topic models. In: UAI, pp. 579–586 (2008)
8. Ren, L., Dunson, D.B., Carin, L.: The dynamic hierarchical Dirichlet process. In: ICML, pp. 824–831 (2008)
9. Zhang, J., Song, Y., Zhang, C., Liu, S.: Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora. In: SIGKDD, pp. 1079–1088 (2010)
10. Wang, X., McCallum, A.: Topics over time: a non-Markov continuous-time model of topical trends. In: SIGKDD, pp. 424–433 (2006)
11. Dubey, A., Hefny, A., Williamson, S., Xing, E.P.: A nonparametric mixture model for topic modeling over time. In: SDM, pp. 530–538 (2013)
12. Swanson, D.R.: Undiscovered public knowledge. Library Quarterly **56**, 103–118 (1986)
13. Settles, B.: ABNER: an open Source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics **21**, 3191–3192 (2005)
14. Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pander, A., Chinnaiyan, A.M.: A cancer microarray database and integrated data-mining platform. Neoplasia **6**, 1–6 (2004)
15. Simpson, M.S., Demner-Fushman, D.: Biomedical text mining: a survey of recent progress. In: Mining Text Data, pp. 465–517 (2012)
16. Kumar, V.D., Tipney, H.J.: Biomedical Literature Mining. Springer (2014)
17. Blei, D.M., Franks, K., Jordan, M.I., Mian, I.S.: Statistical modeling of biomedical corpora: mining the Caenorhabditis genetic center bibliography for genes related to life span. BMC Bioinformatics **7**, 250 (2006)
18. Arnold, C.W., El-Saden, S.M., Bui, A.A., Taira, R.: Clinical case-based retrieval using latent topic analysis. AMIA **2010**, 26 (2010)
19. Arnold, C.W., Speier, W.: A topic model of clinical reports. SIGIR, pp. 1031–1032 (2012)
20. Wu, Y., Liu, M., Zheng, W., Zhao, Z., Xu, H.: Ranking gene-drug relationships in biomedical literature using latent Dirichlet allocation. In: Pacific Symposium on Biocomputing, pp. 422–433 (2012)
21. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. The Annals of Statistics, 209–230 (1973)
22. Sethuraman, J.: A constructive definition of Dirichlet priors. Technical report, DTIC Document (1991)
23. Kanner, L.: Irrelevant and metaphorical language in early infantile autism. American Journal of Psychiatry **103**, 242–246 (1946)

24. Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K.: WordNet: An online lexical database. Int. J. Lexicograph **1**, 235–244 (1990)
25. Miles, J.H.: Autism spectrum disorders - a genetics review. Nature **13**, 278–294 (2011)
26. Wakefield, A.J., Murch, S.H., Anthony, A.: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. The Lancet, 637–641 (1998) (retracted)