

A Survey On Interactivity in Topic Models

Patrik Ehrencrona Kjellin
School of Software Engineering
Tongji University
China

Yan Liu
School of Software Engineering
Tongji University
China

Abstract—Trying to make sense and gain deeper insight from large sets of data is becoming a task very central to computer science in general. Topic models, capable of uncovering the semantic themes pervading through large collections of documents, have seen a surge in popularity in recent years. However, topic models are high level statistical tools; their output is given in terms of probability distributions, suited neither for simple interpretation nor deep analysis. Interpreting the fitted topic models in an intuitive manner requires visual and interactive tools. Additionally, some measure of human interaction is typically required for refining the output offered by such models. In the research, this area remains relatively unexplored – only recently has this aspect been receiving more attention. In this paper, the literature is surveyed as it pertains to interactivity and visualisation within the context of topic models, with the goal of finding current research trends in this area.

Keywords—topic model; latent dirichlet allocation; LDA; interactive; visualisation; IVA; survey; review

I. INTRODUCTION

With the advent of the Internet, a fundamental change has been experienced in the way we access and use information. For instance, a scientist of today wishing to research some subject may be faced with thousands of relevant articles retrieved from journals spanning numerous decades.

With this in mind, it is of increasing importance to be able to extract useful information and make sense of large collections of documents by the means of computation, a task that has come to be very central to computer science in general. *Topic models* have in recent years emerged as a powerful set of techniques for discovering the underlying semantic structure of large, unstructured collections of documents [1].

Topic models are typically Bayesian or linear algebraic models able to extract abstract topics pervading through large corpora. Through the results of such analysis, the individual documents can then in turn be organised in accordance with the themes.

Powerful as they are, topic models do suffer from some problems that may deter some users, or at the very least prevent them from reaping the full benefits of the methods. Often, the models are treated as "black box" approaches without regard for the underlying assumptions they are based on. Parameter tuning can prove difficult without a full understanding of the specific technique to be employed [2]. Additionally, the emerging topics are by no means guaranteed to be sensible to a human reader – motivating the use of human knowledge and user interaction as an additional step toward more coherent and sensible results [3].

Furthermore, the raw, numerical output of topic models may not always lend itself to easy interpretation. Interactive visual analysis in general has proven a useful tool for interpreting and gaining insight from the results of topic models in an intuitive manner. Despite the fact that topic models in general have been subject to a great deal of research in recent years, the visualisation of topic models is still a relatively unexplored area [4].

The *objective* of this paper is to survey the use of visual and interactive data analysis in conjunction with topic models in the literature. In particular, the author is interested in finding out when, how, and for what purposes interactive visual analysis have been used to enhance topic models, and in which ways visualisation can be used to interpret its results.

In Section II and III, respectively, we provide the survey methodology employed and a brief background on topic modeling techniques and interactive visual analysis. In Section IV, a survey of the literature related to interactivity and interactive visual analysis within the context of topic modeling is presented. Finally, in Section V, we conclude the survey with what we perceive to be future work in the integration of visualisation and interaction in the context of topic models.

II. SURVEY METHODOLOGY

Here we describe the methodology for finding and selecting papers for review, and the reasoning behind it.

Visualisation of fitted topic models is a relatively young field; while many topic model papers include some degree of output visualisation, it is rarely the main focus of the paper. Papers purely dealing with the subject are somewhat sparse, and to the knowledge of the author, no summarised overview of such papers exist. Additionally, we are interested in techniques and methods that not only visualise topic models, but also provide the user with some degree of interactivity.

Papers candidate for review have been found primarily through common search engines (i.e., Google) and the digital libraries of **ACM** and **IEEE**. Search terms used are various mixtures of {topic, model, IR, interactive, visualisation, visual, IVA}. Papers have then been selected based on their relevancy, as deemed by the author upon glancing over the contents. As a basic criteria for relevancy, the papers must be focused on topic modeling, while also including some aspect related to interactive visualisation, possibly incorporating human algorithm supervision.

This survey does not attempt to compare the methods surveyed against other methods of visualisation that do not

include much of an interactive component, as no such papers are reviewed. It simply attempts to create an overview of current research trends within this specific subset of visualisation techniques, as they relate to topic models.

III. BACKGROUND

Here we provide some brief descriptions of topic modeling techniques and interactive visual analysis in general. For a more in-depth, comprehensive view of these topics, we refer to relevant papers.

A. Topic Models in Brief

In information retrieval (IR), the general term *topic model* refers to a suite of algorithmic approaches to discovering the latent topics present in a collection of documents. Some basic vocabulary is necessary for describing the general concept of topic models. Formally,

- A word w is a basic unit of data (for instance, a string of alphanumerical characters, but topic modeling can also be applied to other domains than natural language processing)
- A document d is an ordered sequence of N words, w_1, w_2, \dots, w_N .
- A corpus is an unordered set of M documents, denoted by $D = \{d_1, d_2, \dots, d_M\}$.

Topic modeling then consists of taking a corpus D as input and computing K topics (typically in terms of multinomial distributions over the words in the vocabulary), and associating each document with the relevant topic (again, in terms of a multinomial distribution over the different emerging topics).

Early attempts at tackling this problem were however largely concerned with creating a *term-document matrix*, describing the relative frequency of words in each document $d \in D$. This method is useful for many applications, but insufficient in terms of topic modeling, as such a matrix provides little size reduction w.r.t the original corpus, and does not take into account the relationships between words within a document or documents within a corpus [5].

Further research resulted in the strictly linear algebraic approach Latent Semantic Indexing (LSI), which uses singular-value decomposition in order to significantly minimise the term-document matrix [6]. This was later on extended by Probabilistic LSI (PLSI) [7], an early generative model attempting to correct some of the statistically unsound aspects of LSI.

1) *Latent Dirichlet Allocation*: Latent Dirichlet Allocation (LDA) is a generative statistical model loosely based on earlier work on LSI and probabilistic variations thereof. LDA attempts to address some perceived shortcomings found in the previous generative model pLSI.

Namely, in pLSI, parameters to be estimated grow linearly with the size of the corpus. It also has a strong tendency for overfitting, and of even greater consequence, the model is unable to generalize topic mixtures onto previously unseen documents (not part of the training data) [2], [5]. Through correcting these problems with a truly generative model, LDA

has seen a surge in popularity and has acted like a springboard for numerous other advancements in IR.

In LDA, documents are regarded as mixtures of a finite set of K underlying topics, where the parameter K must be specified either by the user or determined through computational inference on the corpus to be analysed. Topics, in turn, are seen as multinomial distributions over the words of the vocabulary.

Inherent to LDA is the assumption that each document $w \in D$ is generated accordingly [5];

$$N \sim \text{Poisson}(\xi) \quad (1)$$

$$\theta \sim \text{Dir}(\alpha) \quad (2)$$

$$\forall n \in \{1, 2, \dots, N\} : \quad (3)$$

(a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$

(b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n .

Here, α and β are smoothing factors for document-topic and topic-term distributions, respectively. LDA is then concerned with inferring the relevant posterior distribution (i.e., given the terms present in the corpus, what are the topics) through a *latent variable model* (the latent variables being the topics).

Further details on the inner workings of LDA is not necessary for this paper, and is therefore omitted. For a more in depth description of LDA, see [5], for instance.

B. Interactive Visual Analysis

Interactive visual analysis (IVA) is a set of techniques incorporating visual analytics (VA) and user interaction in computational or statistical analysis.

Typically, IVA is employed in the task of analysing and attempting to obtain deeper insights from large and possibly complex sets of data, where certain information may not easily be extracted from looking at the data set alone.

It is particularly useful for hypotheses generation and validation, since it equips the user with tools enabling them to look at data sets in a variety of different ways and perspectives [8].

IV. TOPIC MODELS AND INTERACTIVITY

In the surveyed articles, interactivity was most commonly incorporated for addressing one of two concerns:

- 1) **Human knowledge injection.** The first use case concerns integrating human knowledge in topic models in some manner. Parameter tuning and model constraints through user interactions can enhance models in various ways. Topic models like LDA rely on parameters that, while there are methods for doing so, can not easily be estimated through computation alone. Often, some emerging topics will be nonsensical to a human user [9]. Through interactivity, a topic model can be guided towards achieving more meaningful results.

- 2) **Topic visualisation.** Visualisation of the emerging topics generated by a topic model appeared in many of the surveyed articles. Graphical tools of many varieties have proved helpful in the task of exploring and attempting to make sense of the results of topic modelling, in order to get an overarching grasp of the various topics spanning some literary corpus. IVA, in the form of allowing users to navigate the corpus and discover the relationships between topics and documents, has been shown to allow users to gain deeper insight in studies [10].

Apart from this, many different task specific measures are to be found in topic model related papers. For instance, topic modeling for the purpose of source code analysis may benefit from visual interactive analysis for displaying the relationships between actual code, requirements documents, and change logs. Here, we focus on general-purpose methods.

A. Human Knowledge Injection in Topic Models

Some researchers have attempted to improve on the results offered by topic models by correcting some of its common shortcomings through incorporating human knowledge in the process. Shortcomings of topic models identified in previous research include non-sensible and incoherent topics [11], certain terms wrongfully belonging to a topic, terms not belonging to a specific topic when they sensibly should [12], et cetera. At its heart, the problem is due to the fact that the objective function subject to optimisation in LDA does not necessarily reflect the expectations on topic quality felt on behalf of a human [13].

Many different extensions to the normal methods (LDA, in particular) have been proposed for improving the results offered by different models. One such approach is by directly incorporating domain knowledge into the model, typically in an *a priori* fashion, thereby introducing a degree of supervision to an otherwise unsupervised model.

In [3], the authors describe constrained LDA (cLDA), a framework for allowing users to add constraints to a model in order to improve it iteratively. Here, constraints are defined on the documents in terms of **must link**, indicating that two documents semantically belong to the same topic, or **cannot link**, representing the opposite.

The general process of this semi-supervised learning, outlined in Figure 1, consists of first performing LDA analysis, then presenting selected documents to the user who adds constraints based on the output, upon which a specialised constrained LDA is computed. The constraints are here encoded as *soft constraints*, which is to say they will be satisfied to some specified degree, but not necessarily fully satisfied.

Similarly, in [12] the authors implement user interaction through allowing users to add constraints to the model formulated in first-order logic (FOL). Here, the FOL constraints are similar to the **must link** and **cannot link** constraints of [3], but defined on word-pairs rather than documents. In some cases, real-time interactive knowledge injection has been applied, such as in [9], where the authors have used similar concepts as in [12] to create a framework allowing users to iteratively and interactively improve topic modeling results.

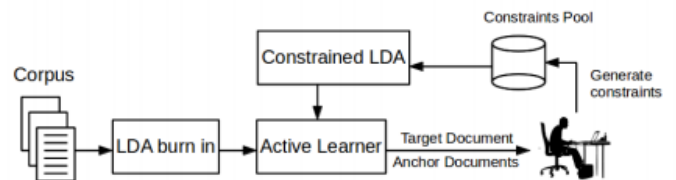


Fig. 1: Diagram from [3], outlining the general process of user-guided constrained topic models.

While the work in [12] and [9] are general-purpose solutions, many of the specialised variations of LDA which incorporate domain knowledge are custom-built, single-purpose methods. All the methods described here for semi-supervised, user guided LDA use only limited visualisation, in the form of matrices or word clouds.

B. Interactive Visualisation of Topic Models

Topic models are high level statistical tools; the raw numerical distributions produced alone are not particularly well suited for intuitive analysis [10]. The visualisation of topic models is an area previously relatively unexplored, which has come under more scrutiny in recent times. Here, some of the concepts and techniques found in the literature (summarised in Table I) are described in terms of their respective unique contributions. There are many ways of visualising topic models, however in this survey of interactive visualisations, the most common representations were found to be either graph based or matrix and text based, along with a few other novel visualisation techniques. The following subsections are organised accordingly.

1) *Matrix & Text Based:* Matrix or tabular representations are generally easily understood from a user perspective [19]. In *Termite*, the authors present a visual analysis system for quickly assessing fitted topic models computed with LDA, for the specific purpose of user-guided, iterative topic modeling. A corpus is here represented in matrix form, wherein rows correspond to words, and columns to topics. While, in its current state, *Termite* is merely a visual tool, the authors outline future work consisting of expanding it into a complete framework for user-guided, iterative topic modeling with the addition of user interaction (in terms of topic deletion and merging, model parameter adjustments, et cetera) [14].

As mentioned, LDA requires several input parameters, the smoothing variables β , α and the number of topics K . There are no strict guidelines for setting these parameters; tuning

TABLE I: Summary of Surveyed Papers

Type	Papers
Matrix & text based	Termite [14], A. Chaney, et al. [10], The Topic Browser [15], H. Yuening, et al. [9], Y. Yang, et al. [3], D. Andrzejewski, et al. [12]
Graph based	Topicnets [16], ParallelTopics [17], LDAVis [18], LDAexplore [19], TopicPanorama [20], S. Rönqvist, et al. [4], Hierarchie [21]
Time Visualisation	TextFlow [22], TIARA [23], ThemeRiver [24], RoseRiver [25]

simple fashion. Users may explore the topics of the sunburst chart in terms of keywords, through selecting individual slices.

3) *Changes Over Time*: Beyond simply visualising fitted topic models statically, recently, plenty of research has been conducted on the visualisation of topics changing over time [22]. Examples of early such methods include *The-meRiver* [24], where topic evolution over time is displayed in terms of a metaphorical river (see Figure 3 for example) made up of smaller streams (topics). Set against a time line, the river provides users with an intuitive overview of how a corpus has changed and at which point specific topics are more or less pervasive in the associated documents. In *TIARA* [23], a tool that resembled the work of [24] in terms of visuals, a river is similarly used as metaphor for the changing of topics over time. However, *TIARA* also includes a rich set of interactive tools for further analysis; users may zoom in on selected topics or topic segments for further analysis. Additionally, by selecting some keyword in the river view, a user can retrieve relevant documents for further examination.

In [22], *TextFlow* was introduced using a novel approach for LDA output analysis. Here, in contrast to previous research, topics are not only displayed as they progress over time (again, in terms of a river), but the splitting and merging of topics is also captured in the visualisation. It is also highly interactive, allowing users to discover what causes the birth, death, splitting and merging of topics throughout the time period of the associated corpus. *Roseriver* [25] further builds upon the work in [22], using a similar river-flow visual representation, but employs a hierarchical topic model in order to better describe large corpus, and for providing users with different overview levels as desired.

V. CONCLUSION & FUTURE WORK

Topic models have seen a surge in popularity in recent years and have provided a new way of discerning useful information from big, complex sets of data, with applications in several different fields.

Recently, much of the effort put into researching topic models, as has been summarised in this survey, is focused on providing users with tools for interacting with and visualising topic models, both in order to improve results in terms of topic coherence and sensibility, and also for allowing users to fully comprehend and benefit from the model outputs.

Many of the attempts at visualising the results of topic models are not limited to simple visualisation, they also provide varying degrees of user interaction with demonstrably improved results [15], suggesting that IVA may play a central role in making these models more available and intuitive to end users.

Research suggests that different representations may aid in different tasks and lead to discovery at different levels [14]. Whereas an overarching graph or matrix based topic overview may provide a deeper understanding of the corpus as a whole, other visuals displaying word relatedness and topic-topic or topic-document relationships may aid in providing other forms of insight or discovery. Currently, most studies focus on a specific level or representation. In future work, several of the proposed representations could be integrated for a more comprehensive view.

Future work in this area may also include more comprehensive frameworks in terms of combining the interactive elements of semi-supervised LDA described in Section IV-A with interactive visual aids for output analysis – both have demonstrable value in terms of increased usability.

ACKNOWLEDGMENT

This work was supported by the Innovation Project at Tongji University Graduate Education (2014JYJG016).

REFERENCES

- [1] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012. [Online]. Available: <http://doi.acm.org/10.1145/2133806.2133826>
- [2] T.-H. Chen, S. W. Thomas, and A. E. Hassan, "A survey on the use of topic models when mining software repositories," *Empirical Software Engineering*, pp. 1–77, 2015. [Online]. Available: <http://dx.doi.org/10.1007/s10664-015-9402-8>
- [3] Y. Yang, S. Pan, D. Downey, and K. Zhang, "Active learning with constrained topic model," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore, Maryland, USA: Association for Computational Linguistics, June 2014, pp. 30–33. [Online]. Available: <http://www.aclweb.org/anthology/W/W14/W14-3104>
- [4] S. Rönqvist, X. Wang, and P. Sarlin, "Interactive visual exploration of topic models using graphs," in *The Eurographics Conference on Visualization (EuroVis)*, R. S. Laramee and M. Chen, Eds. Eurographics, 2014.
- [5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: <http://dl.acm.org/citation.cfm?id=944919.944937>
- [6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, vol. 41, no. 6, pp. 391–407, 1990.
- [7] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '99. New York, NY, USA: ACM, 1999, pp. 50–57. [Online]. Available: <http://doi.acm.org/10.1145/312624.312649>
- [8] S. Oeltze, H. Doleisch, H. Hauser, and G. Weber, "Interactive visual analysis of scientific data," Tutorial at the IEEE VisWeek 2012, October 2012. [Online]. Available: <http://visweek.org/visweek/2012/tutorial/interactive-visual-analysis-scientific-data>
- [9] Y. Hu, J. Boyd-Graber, and B. Satinoff, "Interactive topic modeling," in *Association for Computational Linguistics*, 2011.
- [10] A. J.-B. Chaney and D. M. Blei, "Visualizing topic models," in *ICWSM, J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, Eds. The AAAI Press*, 2012. [Online]. Available: <http://dblp.uni-trier.de/db/conf/icwsml/icwsml2012.htmlChaneyB12>
- [11] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 100–108. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1857999.1858011>
- [12] 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*, ser. IJCAI'11. AAAI Press, 2011, pp. 1171–1177. [Online]. Available: <http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-200>
- [13] J. Boyd-Graber, J. Chang, S. Gerrish, C. Wang, and D. Blei, "Reading tea leaves: How humans interpret topic models," in *Neural Information Processing Systems (NIPS)*, 2009.

- [14] J. Chuang, C. D. Manning, and J. Heer, "Termite: Visualization techniques for assessing textual topic models," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, ser. AVI '12. New York, NY, USA: ACM, 2012, pp. 74–77. [Online]. Available: <http://doi.acm.org/10.1145/2254556.2254572>
- [15] M. J. Gardner, J. Lutes, J. Lund, and J. Hansen, "The topic browser: An interactive tool for browsing topic models," 2010.
- [16] B. Gretarsson, J. O'Donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth, "Topicnets: Visual analysis of large text corpora with topic modeling," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 2, pp. 23:1–23:26, Feb. 2012. [Online]. Available: <http://doi.acm.org/10.1145/2089094.2089099>
- [17] W. Dou, X. Wang, R. Chang, and W. Ribarsky, "Paralleltopics: A probabilistic approach to exploring document collections," in *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, Oct 2011, pp. 231–240.
- [18] C. Sievert and K. E. Shirley, "Ldavis: A method for visualizing and interpreting topics," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014. [Online]. Available: <https://github.com/cpsievert/LDAvis>
- [19] A. Ganesan, K. Brantley, S. Pan, and J. Chen, "Ldaexplore: Visualizing topic models generated using latent dirichlet allocation," *CoRR*, vol. abs/1507.06593, 2015. [Online]. Available: <http://arxiv.org/abs/1507.06593>
- [20] S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo, "Topicpanorama: A full picture of relevant topics," in *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, Oct 2014, pp. 183–192.
- [21] A. Smith, T. Hawes, and M. Myers, "Hiearchie: Visualization for hierarchical topic models," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore, Maryland, USA: Association for Computational Linguistics, June 2014, pp. 71–78. [Online]. Available: <http://www.aclweb.org/anthology/W/W14/W14-3111>
- [22] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong, "Textflow: Towards better understanding of evolving topics in text," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2412–2421, Dec 2011.
- [23] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian, "Interactive, topic-based visual text summarization and analysis," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ser. CIKM '09. New York, NY, USA: ACM, 2009, pp. 543–552. [Online]. Available: <http://doi.acm.org/10.1145/1645953.1646023>
- [24] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "Themeriver: visualizing thematic changes in large document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 9–20, Jan 2002.
- [25] W. Cui, S. Liu, Z. Wu, and H. Wei, "How hierarchical topics evolve in large text corpora," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2281–2290, Dec 2014.
- [26] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested chinese restaurant process," in *Advances in Neural Information Processing Systems*. MIT Press, 2004.