International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307, Volume-2, Issue-6, January 2013

Data Visualization for University Research Papers

Alpa K. Oza

Abstract- Quite many publications are being published either in form of Theses, essays or Research papers at various levels of scientists, research scholars or Ph.D students. This is a big jargon. They are required to be segregated under various Topics. Topic modeling is a set of tool that provides a solution. Topic modeling discovers a hidden thematic structure in collection of documents. Topic models are high level statistical tools. A user must scrutinize numerical distribution to understand and explore their results. Latent Dirichlet Allocation LDA has been used to generate automatically topics of text corpora and also to subdivide the corpus words among those topics. Topic models also fall in the same line of functioning. This model (topic model) has proven remarkably powerful for information retrieval tasks. Information visualization technologies when used in conjunction with data mining and text analyses tools can be of great value for various types of tasks. For this reason various visualizations have been designed. Quite laborious work has been done and still being labored at various levels of scholars. Here our aim is to present a brief description to the topical method of visualization under data mining.

Keywords- Topic Models, Text Visualization, Visual analysis, Text, Statistical model

I. INTRODUCTION

In Gujarat Technological University (GTU) number of research papers published by students of M.E., M.Tech. and Ph.D theses increases. Growth in research papers increases the complexity in analyzing. The chronology of research paper published is clear but tracing the lineage of decision-making may be more opaque. The topical similarity also increases complexity between university departments in terms of their published papers. Consider the visualizations in Fig1, which depict "topical similarity" [1].

Topical Similarity provides one means of identifying which disciplines are sharing information. Because each dissertation is associated with one or more departments, the content of these dissertations was seen as a reasonable basis for inferring whether two departments are working on the same content as seen through the words in their published dissertations. Thus explored various text-derived similarity

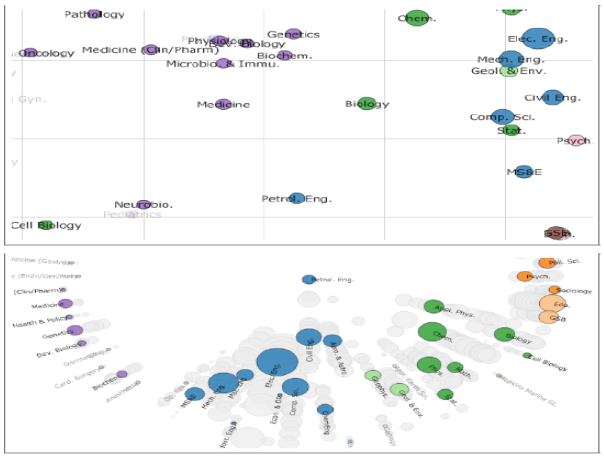

Figure 1: The case of Petroleum Engineering. The top visualization shows a 2D projection of pair wise topical distances between academic departments. In 2005, Petroleum Engineering appears similar to Neurobiology, Medicine, and Biology. Was there collaboration among those departments? The bottom visualization shows the undistorted distances from Petroleum Engineering to other departments by radial distance. The connection to biology disappears: it was an artifact of dimensionally reduction. The visual encoding of spatial distance in the first view is interpretable, but on its own is not trustworthy [1]. measure as the basis of these similarity scores.

There are quite many information being poured on almost every subject and topics. On internet it has been a great jargon of messy webs data being published day in and out. It is not only essential but needy at the same time to arrange each of them. Scholars in the field of IT are working hard to smoothen out this problem of present or for that matter later days. Most people rely on search Technologies (Search-engine we commonly call) to find required information. This model has proved highly effective tool for locating required data or information. Yet considering the nature of multi-face subjects and documents, the problem of data mining at a particular subject remain challenging and mind bogging.

Our goal is to find short description of papers published by the students for further analyzing. As the manually reading the document collection is infeasible due to both the size of the corpus and the expertise required. The size of text corpora often increase of what a person can read and process. While statistical topic models have the potential to aid large-scale exploration but also has a scarcity of real world analyses involving topic models. When the models are deployed, they involve time-consuming verification and model refinement [2]. We present a visualization system for the term-topic distributions produced by topic models for analyses of coherent papers.

II. RELATED WORK

Many researchers at various levels are engaged to find appropriate solution to this problem. Several solutions to the problem of understanding large documents corpora

Manuscript received on January, 2013.

Alpa K. Oza, Information Technology, Parul Institute of Engineering and Technology, Gujarat Technological University, Ahmedabad, India.

include Exemplar-based Visualization [4] and FacetAtlas [5]. Document-level tools use projection-based techniques to visualize relationships between documents in a collection. Many of these visualizations [4] map a set of documents to a 2D display according to document similarity. These visualization help users understand the corpus as a whole, but do not enable exploration of individual documents. Early topic modeling research has focused on building new topic models and improving algorithms. Researchers have typically used browsers to evaluate model algorithms. While a new interactive visualization technique [5] that enables users to navigate and analyze large multifaceted text corpora with complex cross-document relationships. A method for visualizing topic model that creates a navigator of the documents, allowing users to explore the hidden structure that a topic model discovers. These browsing interfaces [3] reveal meaningful patterns in a collection, helping end-users explore and understand its contents in new ways.

We reviewed topic modeling, focusing on *Latent Dirichlet Algorithm* (LDA) [6] which is simplest probabilistic topic models. LDA is a popular approach of discovering latent topics in a text corpus by automatically learning distributions of words that tend to co-occur in the same documents. While LDA produces some sensible topics, a prominent issue is the presence of "junk topics" [7] comprised of incoherent term grouping. Model outputs often need to be verified by domain experts and modified [1] to ensure they correspond to meaningful concepts in the domain of analysis.

Latent topics are often presented to analysts as a list of probable terms [8] which imposes on the analysts the potentially arduous task of inferring meaningful concepts from the list and verifying that these topics are responsive to their goals. In contrast to existing tools for summarizing LDA model output, Termite aims to support the domain-specific task of building and refining topic models [2].

Methods chiefly followed/explored are like Termite or Visualization Technique for assessing Textual Topic Models by Jason Chuang, Manning and J. Heer[2]. The trio scholars have presented a very unique and effective solution to the problem of data mining in their paper published at University level. It is a term topic distribution produced by topic model as the name suggests. They have elaborated Termite system design and made a model of word similarity solution by using G^2 statistics.

III. THE DESIGN OF A UNIVERSITY RESEARCH PAPER MODULE

Our interest in model-driven visualization stems were tasked with investigating the impact of interdisciplinary collaboration at Gujarat Technological University. Our method applied the idea that we could identify influences and convergent lines of research across disciplines by detecting shared language use within university-wide publications. Manually reading the document collection is infeasible. A visual analysis tool for exploring university's years of Ph.D these from number of departments.

When using topic models to analyze a text collection, it is critical that the discovered latent topics be relevant to the domain task. Prior work suggests that the quality of a topic is often determined by the coherence of its constituent words [7]. Input to our model are words, from which we get output of departmental similarity, based on the text of each department's thesis. The perfection of the generative process for documents is achieved by considering LDA priors on data distribution over topics and on topic distribution over words. Researchers L. Alsumait, D. Barbara, J. Gentle and C. Domeniconi at George Mason University have spade enough work on LDA generative models. They have made analysis to distinguish junk topics from legitimate ones. The term junk means topics of insignificance. Various criteria based on methodologies such as uniform word distribution, vacuous, background distribution etc are experimented in details.

There are initially two models constructed, each representing an approach to textual similarity in the literature. The first metric is based on **term similarity**, for ranking and filtering terms. By surfacing more discriminative terms that measures faster assessment and comparison of topics i.e. measuring the overlap of words. The second **is topic similarity**, for sorting terms to reveal clustering patterns i.e. in which we measure similarity in a lower dimensional space of inferred topics [1] [2].

Our goal is to support effective evaluation of term distributions associated with LDA topics. The tool is designed to assess the quality of individual topics and all topics as a whole. We encode term probabilities as circles, with areas proportional to the number of dissertations filed in a given year. Distance between circles encodes one of the similarity measures [1]. Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis [1] visualizes through Landscape view, Department view and Thesis Views such as shown in Fig 1 with Landscape view.

While the **term-topic matrix** [2] shows term distributions for all latent topics. Here matrices support comparison across both topics and terms. The matrix view means rows corresponded to terms and columns to topics. Showing all words in term-topic matrix is neither desirable nor feasible due to large vocabularies with thousands of words. So we use this term-topic matrix visualization in our system as it can filter the display to show the most probable or salient terms. We can choose 10 terms out of 250 terms. Displaying over 250 terms does require a significant amount of scrolling which greatly reduces effectiveness of visualization.

While term-topic model represents term frequency which is topics usage in whole corpus and also the documents related to it. We also used topic similarity which measures for co-occurrence and collocation likelihood between all pairs of words. Collocation means the probability that a sequence of words occurs more often in a corpus, for example, "parallel coordinates" is right phrase while "coordinates parallel" is just not the right one [2] as shown in fig 2.

Earlier model just visualized measure of frequent terms on the topical similarity basis. The visualization reveals two addition views, the word frequency view and document view where word frequency view shows the topic's word usage relative to full corpus and the document view shows the representative documents belonging to the topic as shown in Fig. 3 [2].

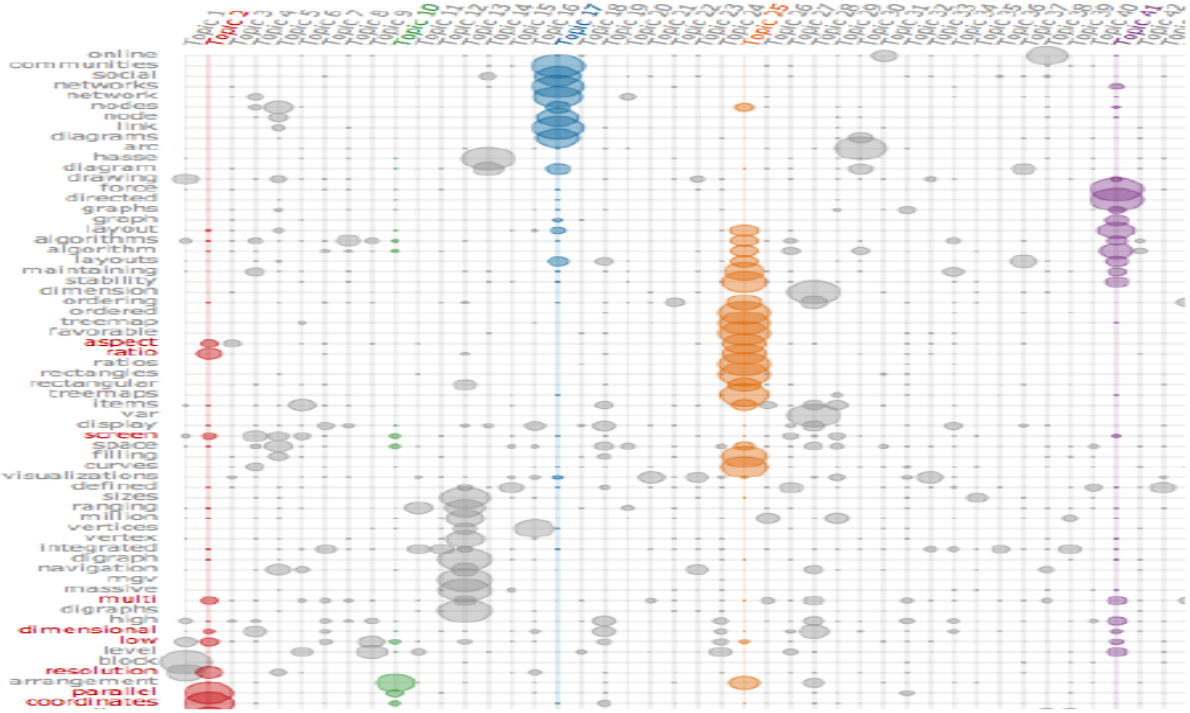

Fig 2: Topic Similarity, shows co-occurrence of terms which is revealed by clustering patterns.

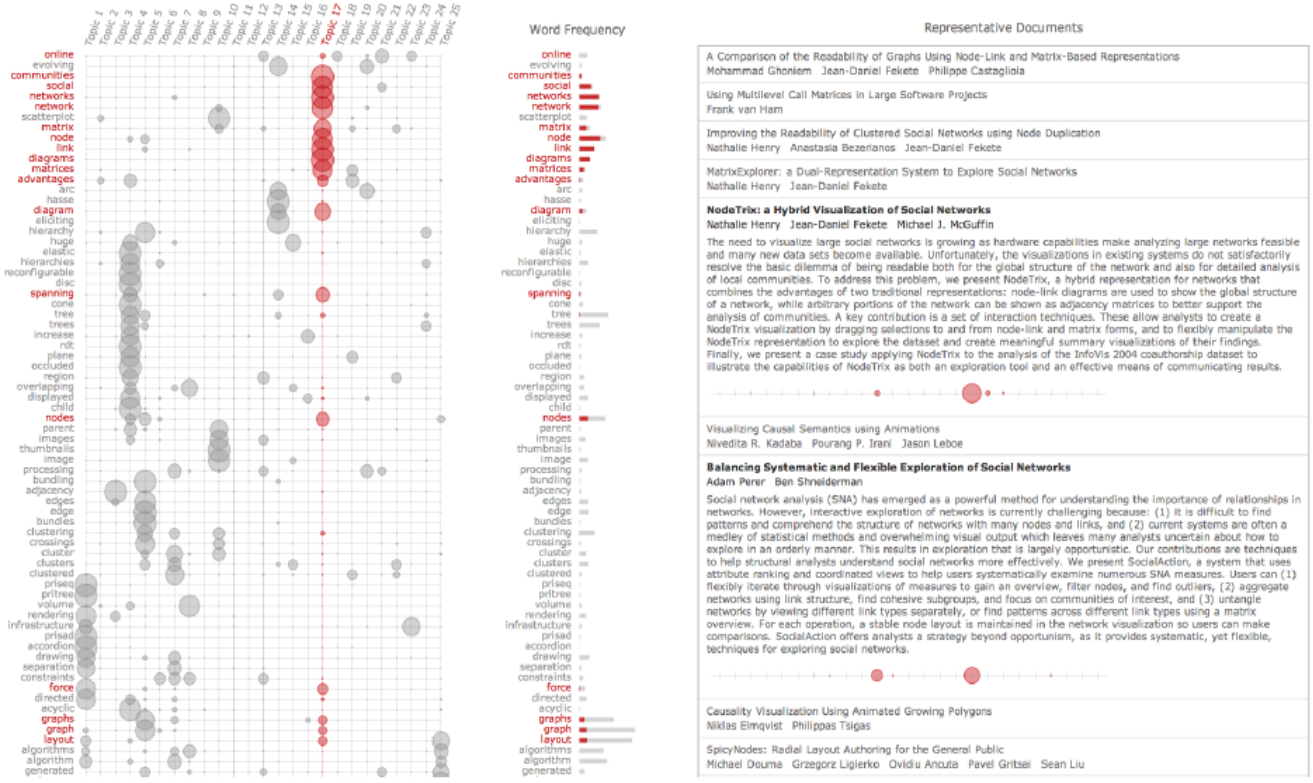

Figure 3: When a topic is selected in the term-topic matrix (left), the system visualizes the word frequency distribution relative to full corpus (middle) and shows the most representative documents (right).

IV. CONCLUSION

In this paper, we presented University research paper module for term-topic matrix. We also compared this modeling technique with other designing model-driven visualization for text analysis. Through this system students and university members will find easy to analyze different topics, its related similarities and also have a fast look on the preface or the abstract of the research paper. Along with the research paper's published date, name of Conference/Journal were it was published and other details of the researcher related to it are represented. The interactive model refinement can significantly improve the utility and reduce the cost of applying topic models to make sense of large text corpora. With the rapid growth on information being displayed it is essential to arrange in same for easy and smooth accessible findings of a particular subject. The term topic models are having more significance in a way to keep away the junk or irrelevant subjects on the net.

REFERENCES

[1] J. Chuang, C. D. Manning, and J. Heer, "Interpertation and Trust: Designing model-driven visualizations for text analysis", In CHI, 2012.

[2] J. Chuang, C. D. Manning, and J. Heer, "Termite: Visualization Techniques for assessing textual topic models", In ACM, 2012.

[3] Allison J. B. Chaney and D. M. Blei. "Visualizing Topic Models", In AAAI, 2012.

[4] Y. Chen, L. Wang, M. Dong and J. Hua. "Exemplar-based Visualization of Large Document Corpus", 2009. . IEEE Transactions on Visualization and Computer Graphics 15(6): 1161-1168.

[5] N. Cao, J. Sun, Y-R. Lin, D. Gotz, S. Liu and H. Ou. "FacetAtlas: Multifaceted Visualization for Rich Text Corpora", 2010. IEEE Transactions on Visualization and Computer Graphics 16(6): 1172-1181.

[6] D. M. Blei, A. Y. Ng and M. I. Jordan. "Latent Dirichlet Allocation", J Machine Learning Research, 3:993-1022, 2003.

[7] L. Alsumait, D. Barbara, J. Gentle and C. Domeniconi. "Topic Significant ranking of LDA generative models", In ECML, 2009.

[8] J. Chang, J. Bod-Graber, C. Wang, S. Gerrish and D. M. Blei. "Reading tes leaves: How Humans interpret topic models", In NIPS, pages 288-296, 2009.