

CORRELATED TOPICS IN A SCALABLE MULTIDIMENSIONAL TEXT CUBE: ALGORITHMS AND AVIATION SAFETY CASE STUDY

BO ZHAO*, CINDY X. LIN*, ASHOK N. SRIVASTAVA**, NIKUNJ C. OZA**, AND JIAWEI HAN*

1. INTRODUCTION

Many organizations have large text repositories that contain information that is mission critical to the organization. NASA, for example, operates a safety reporting system known as the Aviation Safety Reporting System (ASRS) which collects voluntarily submitted aviation safety incident/situation reports from pilots, controllers, and others with the purpose to identify system-wide deficiencies and safety issues[2]. ASRS can receive as many as several thousand reports in a month and contains over 100,000 reports at this time. The analysts at ASRS analyze about 20% of the reports in detail and assign them to potentially several of 60 high-level anomaly categories and conduct other safety related studies with the reports. The reports also contain various 'fixed-field' pieces of information that identify the context and operating conditions of the flight such as weather and phase of flight. The reports are anonymous by law; thus, the author, her organization, and other identifying pieces of information are removed from the reports. The ASRS analysts also use automated tools to help them study the reports and compare them with others, but the vast majority of the work is done by skilled experts.

The problem that we focus on in this paper is the generation of a method to automatically uncover documents and topics that are correlated with a topic of interest and then analyze the resulting set of reports using a scalable multidimensional cube [3]. The multidimensional cube could consist of the fixed-fields already identified in a set of reports, but more interestingly, other topics that have been discovered in the text repository¹. These reports can offer a significant amount of insight into the main topic and its contributing factors. We use this as a running example throughout the paper to illustrate the performance and output of the system.

For example, an analyst may be interested in studying pilot fatigue due to FAA or experts in relevant areas indicating that fatigue is of interest because it is thought to be a contributing factor to aviation safety incidents. The authors of the aviation safety reports may not directly mention the word *fatigue* in their writeups. Instead, they may mention other phrases such as “*I was on the last leg of a 5 segment trip*” or *THIS WAS THE FINAL LEG OF A MULTI-LEG FLT AND I WAS MORE TIRED THAN I THOUGHT*. In these examples, the author does not directly state the word fatigue. In the last example, the author indicated that he/she was tired due to being on the final leg of a trip. Notice in this excerpt that the author uses abbreviations; ASRS documents are laden with abbreviations in the narrative sections.

In spite of these characteristics of the ASRS documents, we would like that the analyst can enter a set of relevant words, such as ‘fatigue’ and ‘tired’, and get back a set of related topics, fixed-field values, anomaly categories, and relevant documents. These results can then be further analyzed so that topics correlated to the main topic can be discovered. We now present the system that we developed to facilitate this analysis and give a case study describing an example of such analysis.

2. ALGORITHMS

2.1. Topic Relevance Analysis. In this section, we formally define the problem of topic relevance analysis as: given a keyword query $Q = \{q_1, q_2, \dots, q_{|Q|}\}$, we try to find topics θ that have high

*University of Illinois at Urbana-Champaign, bozhao3@uiuc.edu, xidelin2@uiuc.edu, hanj@cs.uiuc.edu

**NASA Ames Research Center, ashok.n.srivastava@nasa.gov, nikunj.c.oza@nasa.gov .

¹Note that these topics are different from the 60 high-level anomaly categories mentioned earlier.

relevance score $Rel(Q, \theta)$, which can be defined as the conditional probability of the topic given the query Q :

$$(1) \quad Rel(Q, \theta) = Pr(\theta|Q) \propto Pr(Q|\theta)Pr(\theta) = \left(\frac{1}{|D|} \sum_{d \in D} Pr(\theta|d) \right) \prod_{q_i \in Q} Pr(q_i|\theta)$$

2.2. Topic Correlation Analysis. In topic correlation analysis [1, 4], two topics are correlated if they appear in the same or similar contexts. The so-called ‘context’ can be interpreted as the documents that have higher probability of the topic. Therefore, we define the correlation of two topics α and β over a corpus $D = \{d_1, d_2, \dots, d_{|D|}\}$ as the cosine distance of their topic distribution vectors:

$$(2) \quad Col(\alpha, \beta) = Cosine(\vec{V}(\alpha), \vec{V}(\beta)) = \frac{\vec{V}(\alpha) \cdot \vec{V}(\beta)}{\|\vec{V}(\alpha)\| \|\vec{V}(\beta)\|},$$

where $\vec{V}(\alpha)$ is the topic distribution vector of α , i.e., $\vec{V}(\alpha) = (Pr(\alpha|d_1), Pr(\alpha|d_2), \dots, Pr(\alpha|d_{|D|}))$. Moreover, we want to measure the topic correlation in every cell of a text cube, in which case we will only consider documents in the cell in the topic distribution vector, i.e., for a cell c , $\vec{V}_c(\alpha) = (Pr(\alpha|d'_1), Pr(\alpha|d'_2), \dots, Pr(\alpha|d'_{|c(D)|}))$, where $c(D) = \{d'_1, d'_2, \dots, d'_{|c(D)|}\}$ is the document set in cell c .

2.3. Query Processing and Cube Computation. Given a target topic α and cell c , we want to find the topics most correlated to α in cell c . There are several strategies for executing such query. First, without any pre-computation, every document in the cell needs to be evaluated, which costs too much time. On the other hand, we could materialize some useful information in the text cube so that online queries can be handled efficiently. However, if we materialize all the cells in the cube, there will be too much storage overhead. To overcome the disadvantages of these two approaches, we find the topic correlation query can also be executed efficiently in a *partially materialized* cube, and propose a cube computation algorithm that can *optimize the storage* and guarantee *bounded query execution time*. Experiments justifies the performance of our method.

2.4. Case Study. We investigate the ‘fatigue’ issue in experiments. During topic modeling, we can successfully group ‘fatigue’ with some related words in the same topic, such as ‘long duty’ and ‘insufficient rest’. However, many other terms correlated to ‘fatigue’ are not discovered. By mining correlated topics, we are able to infer other more detailed, complex factors that relate to ‘fatigue’.

For example, we find one highly correlated topic has higher probabilities for the words ‘stress’, *etc.* In the representative report, the pilot first explained “ALTHOUGH I AM COMPLETELY FAMILIAR WITH THE AIRSPACE; I COMPLETELY FORGOT ABOUT THAT SEGMENT OF THE CLASS B”, then later actually mentioned it may be because of fatigue, and there was not enough rest between the flights. And the correlation is discover in cell ‘[Flight Phase]: cruise level’, which indicates pilots are most influenced by fatigue during that phase.

REFERENCES

- [1] D. M. Blei and J. D. Lafferty. Correlated topic models. In *NIPS*, 2005.
- [2] L. Connell. Aviation safety reporting system. Technical report, NASA Ames Research Center, 2010.
- [3] C. X. Lin, B. Ding, J. Han, F. Zhu, and B. Zhao. Text cube: Computing ir measures for multidimensional text database analysis. In *ICDM*, pages 905–910, 2008.
- [4] K. Salomatin, Y. Yang, and A. Lad. Multi-field correlated topic modeling. In *SDM*, pages 628–637, 2009.