In-depth Interactive Visual Exploration for Bridging Unstructured and Structured Document Content

Axel J. Soto* Ryan Kiros† Vlado Keselj* Evangelos Milios*

Abstract

Semi-structured data refers to the combination of unstructured and structured data. Unstructured data is free text in natural language, while structured data is typically stored in tables and following a data schema. Recent statistics shows that 80% of the data generated in the last two years is unstructured. However, one interesting observation is that free text usually comes along with some structured data, or meta-data describing or adding more information about the text. In this paper we present ViTA-SSD, a Visual Text Analytics Tool for Semi-Structured Data. This tool aims at extracting interesting patterns in semi-structured data through the joint consideration of the free text and the meta-data. This represents a challenging task because an effective approach needs the combined effort of text mining algorithms and human experts who can drive the exploration process in a meaningful way. A related challenge is the appropriate visualization and understanding of the patterns found. In order to address these challenges, our visual analytics tool takes advantage of a novel dimensionality reduction and a fast user-supervised clustering method. We showcase our tool here as well as we reflect on some lessons learned from the development and evaluation of our tool.

1 Introduction

Even though research in machine learning and text mining has contributed to the development of methods for automatic document analysis, finding actionable insight in text data usually requires a human user that queries and interacts with the data in order to find answers to her questions. Yet the field has so far not fully developed human-in-the-loop approaches to enable the user to perform visual data exploration and visual analytics, so that the user may arrive at meaningful and relevant results [3].

We will use the term *semi-structured data* to refer to the data that have an unstructured component (free text) and a set of structured fields associated with each document (meta-data). Despite the fact that semi-structured data can be found in many scenarios, little work has been done on analyzing information from semi-structured data. Kandel et al. [2] report interviews with 35 data analysts, where they indicate the processing of semi-structured data as one of the major analysts' challenges.

For instance, an analyst aided by a text mining tool may find several clusters of documents describing problems that are similar or related. However, the analyst would also like to relate this information to the meta-data present in the documents. In the context of aviation incident reports[1], some possible questions an analyst may ask include: are certain incident descriptions more likely to be produced in certain locations than others? To answer this type of questions, or to even discover unexpected relationships in our data, we need computational support to bridge the analysis of the structured and unstructured components of the data.

In this work, we present a Visual Text Analytics Tool for Semi-Structured Documents (ViTA-SSD). The system has two main user-oriented goals. First, to enable a user to discover patterns hidden in the text and meta-data by means of a visual and interactive exploration of the data. This is done by combining the results of a user-driven document clustering with the associated document meta-data. Second, to make the discovered patterns understandable and verifiable by the users. To address these goals we have developed a novel dimensionality reduction method, which is one of the core elements of the system to represent the unstructured data internally. We have also redesigned a clustering approach for large document corpora that allows user supervision. We will give a brief overview of the main components of this system here, and we will demo the system during the workshop. We conclude this paper formulating general design principles that aim at being applicable to any other visual text analytics systems.

2 Statistical and Data Mining Techniques

2.1 Dimensionality reduction methods on text generate a new (and typically smaller) set of latent

*Dalhousie University, Canada – soto@cs.dal.ca

†University of Toronto, Canada

[1] We use the ASRS data for examples this paper http://asrs.arc.nasa.gov/

features that can capture richer semantic relationships. This compact representation serves two other purposes. First, it provides a two-dimensional representation of the corpus that users can interact with. Second, since the reduced space is a good approximation of the raw representation of the corpus, this compact representation allows measuring similarity among documents considerably faster, and hence clustering algorithms can be faster too.

Numerous approaches have used variants of Principal Component Analysis (PCA), Latent Semantic Indexing, or Multi-Dimensional Scaling as dimensionality reduction methods for text. The main problem with PCA-like approaches and classical scaling methods is that they focus mainly on retaining large pairwise distances, which is undesirable for document corpora because distances between similar documents is not well preserved. The machine learning community has been active in the area of dimensionality reduction and feature learning, and many algorithms based on deep learning have been proposed [1]. Deep learning typically involves a greedy layer-by-layer training of bipartite Markov random fields known as Restricted Boltzmann Machines. This type of training could be used to learn feature hierarchies from data as a preprocessing step to learning tasks or for initialization of the weights of a deep neural network. These methods have been shown to lead to better embeddings than most popular non-linear dimensionality reduction methods [1]. While these methods have an expensive training phase, they can handle new incoming data efficiently.

ViTA-SSD incorporates a novel unsupervised dimensionality reduction method which we shall denote as *deep t-distributed correlative subspace mapping* (dt-CSM). Our algorithm uses a deep pre-trained network to learn a mapping by maximizing: $r(P, Q)$, where r is the Pearson correlation, and P and Q denote document pairwise similarities in the original and latent space, respectively. This objective function is inspired by [6]. In this way we can extend a linear dimensionality reduction method to learn a highly non-linear parametric mapping.

2.2 Fast user-supervised clustering. One important design choice in exploratory data analysis systems is given by the tradeoff between accuracy and speed of the text mining methods. Clustering is one crucial component of ViTA-SSD as it is used to support corpus exploration and sense-making [4]. Therefore, we have incorporated the clustering method proposed in [5], which was shown to have a significant speed-up with a marginal drop of the clustering accuracy compared to the traditional k-means.

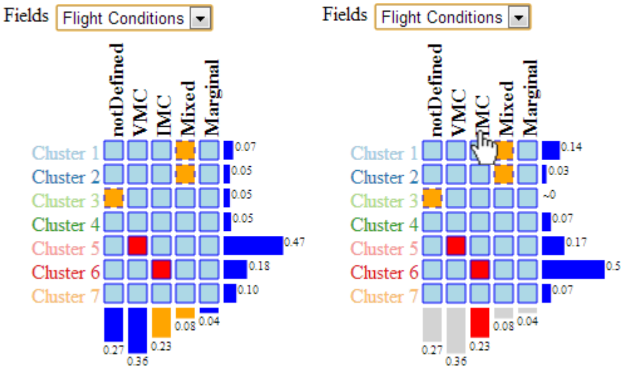
Yet another important feature of visual analytics systems is to provide some sort of supervision to better reflect users' intentions. Similar to [7] and others therein, our approach makes use of feature supervision for influencing cluster formation. In these approaches, relevant terms are provided to the algorithm, so that they can have an impact when calculating distances among documents. Our approach differs from previous works, in the sense that the feature supervision strategy can be applied on a different feature space, e.g. after a non-linear dimensionality reduction step. In this way, we have adapted the fast clustering algorithm presented in [5] to modify distances to centroids based on the occurrence of the suggested terms without the need to compute distances in the original feature space.

3 ViTA-SSD Interface

ViTA-SSD aims at supporting the exploration of semi-structured data and helping to identify relevant patterns in the data. There are many different ways the data can be explored, such as searching, filtering or clustering. Clustering provides a way of aggregating related documents and get a sense about what they have in common by interacting with their main keywords and representative documents. Clusters are also a critical component for the correlation of related free-text descriptions with particular meta-data values. The interface of ViTA-SSD is divided in six coordinated panels. Due to space limitations we will just describe the two most important visualizations: the *document exploration* and the *meta-data analysis* panel.

Let us illustrate the clustering and its user-supervision process on the *document exploration.* The collection of documents at an initial stage is shown in Fig. 1(a). Let us assume that the analyst would like to have a cluster only devoted to bird strike incidents while keeping all other documents selected. We can see that cluster 4 (dark green) has the words *engine*, *emergency* and *landing* among the most representative ones. Note that *bird* is also listed, but with a smaller font. The result of searching for "bird" is shown in Fig. 1(b). We can see that most documents are a subset of the original cluster 4 with some others being close but in neighboring clusters. This illustrates the capacity of dt-CSM to locate related documents close to each other. The state of the system after "bird" is suggested as a relevant keyword (with influence factor at 10%) is shown in Fig. 1(c). We can note how cluster 4 has shrunk around the documents about bird strike incidents, and "bird" became the second largest word. The influence factor of the suggestion allows the user to control the extent that distances are adapted to cluster together documents containing the suggested term. Finally, Fig. 1(d) shows the clustering assignment after supervision with the influence factor at 80%. We can now see that the cluster keywords better reflect bird strike incidents. The cluster has become much more exclusive and distances in that neighborhood are distorted to benefit documents that contain "bird". Pushing this influence factor even further (90%-100%) makes all documents that contain the word "bird" (Fig. 1(b)) members of the same cluster.

The meta-data analysis panel is responsible for bridging and visualizing patterns between the unstructured and structured content (Fig. 2). This visualization is organized as a matrix layout, where rows represent clusters and columns represent meta-data values for a given meta-data field. Each cell is colored orange and red according to the existence of strong or very strong evidence of an unlikely correlation, defined as $p < 0.05$ or $p < 0.01$, where p quantifies the probability of expecting this correlation from the data distribution. The intuition behind a colored cell is described in Fig. 2(b). There are some other information and interactions available in this visualizations, which we do not comment on due to space reasons.



(a) (b)

Figure 2: Meta-data analysis panel for "Flight conditions" after filtering incident reports about turbulence encounters. (a)Very strong correlations can be identified e.g. between "Cluster 6" and "IMC" (Instrument Meteorological Conditions). (b) When the meta-data value "IMC" is hovered over, we can see how incidents that happen under this flight category are assigned to the different clusters. The large increase in documents assigned to "Cluster 6" (from 18% to 50%) suggests that an interesting correlation happens between documents assigned to "Cluster 6" and the "IMC" category.

4 Discussion

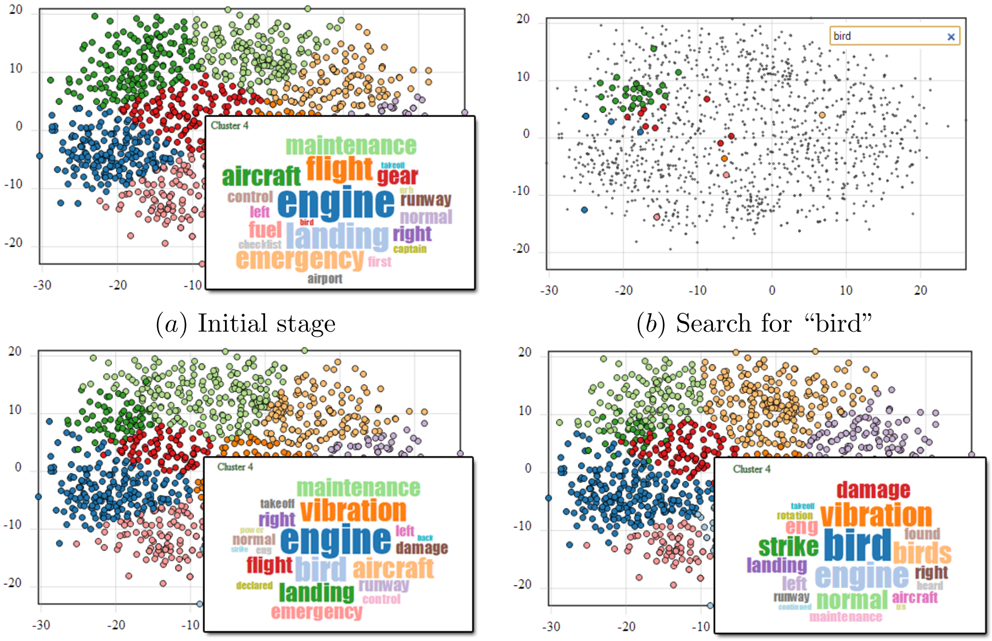
The development of our visual analytics tool allowed us to learn some lessons and to propose guidelines that are applicable to the design of any visual text analytics method:

Fast processing after interactions. Unfortunately many text mining and machine learning algorithms are compute intensive and the time needed to run may get prohibitive for large document collections or for real-time usage. In order to keep users engaged exploring in the document collection and interacting with the interface, it is important that the latency time between the interaction submission and its outcome can be as short as possible. This leads to a careful selection of algorithms that may be pre-computed off-line, and those that may be executable in real time. For the ones that need to run on-line, there is usually a clear trade-off between result precision and computation speed. In many scenarios a faster outcome with a slightly larger error may be more desirable than the slower but more accurate decision. One way to deal with this issue is to place this responsibility on the user by providing a way of manually controlling the computation time in tradeoff to its accuracy.

Simple method parameterization. Text mining methods usually require some sort of parameterization that makes the method more adaptable to different scenarios. However, we believe that domain experts tend to be reluctant to learn complex computing concepts if these are not strictly related to their work. The challenge here is to select methods that enable an intuitive parameterization for a lay person or design the interface in such a way that parameters are replaced by more user-oriented meta-parameters. Also, we found useful from our user study to provide a two-level option interface, where more advanced features are not visible but can be accessed on demand for increased flexibility.

Influencing similarity. Measuring similarity is key in text mining. An old and recurring question in text mining is what criteria should be followed to consider two documents to be similar to each other. There is no simple answer to this question as different users may regard different features to be relevant for similarity. Visual analytics can provide a way to deal with this issue as opposed to defining a fixed criterion or distance measure in advance. In this regard, we consider important to provide different channels for corpus exploration as well as different ways to influence the text mining algorithms.

One of the main characteristics of the modern phenomenon called *Big Data* is that documents are not only a large static repository, but also arrive as a never-ending stream of data. This imposes some limitations



(*a*) Initial stage (*b*) Search for "bird"

(*c*) *bird* suggested at 10% (*b*) *bird* suggested at 80%

Figure 1: Use case of the fast user-supervised clustering algorithm where the user is interested in obtaining one cluster about bird strike incidents.

on the methods we can use for dimensionality reduction on the document collections. Models based on deep architectures scale well with very large data sets because the model is trained in mini-batches of data, as opposed to other traditional dimensionality reduction methods that need to store all pairwise similarities, and hence suffer from a memory limitation. While models based on deep learning are slow to train, it is a one-time effort that can be performed off-line. Moreover, in the scenario of dynamic data, new incoming documents could be visualized or projected immediately by using the representation that was learnt based on previous documents.

ViTA-SSD successfully integrates novel machine learning methods with interactive visualizations to support exploration and pattern finding in semi-structured data. Numerical experiments have been carried out in order to assess and compare our dimensionality reduction method and fast user-supervised clustering algorithm. We have also run a user-study that allowed us to evaluate ViTA-SSD from a user-centered perspective.

Acknowledgements

This research has been supported by The Boeing Company and a grant from the Natural Sciences and Engineering Research Council of Canada.

References

[1] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[2] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer. Enterprise data analysis and visualization: An interview study. In *Visual Analytics Science and Technology, 2012. VAST 2012. IEEE Conference on*, 2012.

[3] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. *Mastering the Information Age-Solving Problems with Visual Analytics*. Florian Mansmann, 2010.

[4] Y. Liu, S. Barlowe, Y. Feng, J. Yang, and M. Jiang. Evaluating exploratory visualization systems: A user study on how clustering-based visualization systems support information seeking from large document collections. *Information Visualization*, 12(1):25–43, 2013.

[5] D. Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1177–1178. ACM, 2010.

[6] M. Strickert, A. J. Soto, and G. E. Vazquez. Adaptive matrix distances aiming at optimum regression subspaces. In *Proceedings of the European Symposium on Artificial Neural Networks*, pages 93–98. D-facto Publications, 2010.

[7] D. M. Witten and R. Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490), 2010.