VAPD - A Visionary System for Uncertainty Aware Decision Making in Crime Analysis

Florian Stoffel, Dominik Sacha, Geoffrey Ellis and Daniel A. Keim, *Member, IEEE*

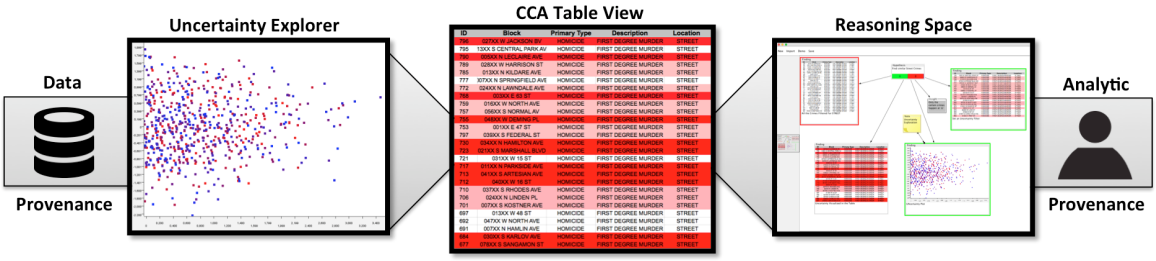

Fig. 1. VAPD-Overview: The *CCA Table View* to solve the analysis task, the *Uncertainty Explorer* for understanding different uncertainties and their impacts, and the *Reasoning Space* that supports uncertainty aware sensemaking. The basis for integrating analysis, user and uncertainty information are data and analytic provenance techniques.

Abstract— In this paper we describe a visionary system, VAPD, which supports crime analysts in uncertainty aware decision making in use of comparative case analysis. In this scenario, it is crucial for crime analysts to get an accurate estimate of uncertainties included in their data as well as those caused through data transformations and mappings, thus supporting analysts in calibrating their trust in the pieces of evidence gained through data analytics. VAPD consists of one data processing and three visualisation components that adopt a set of guidelines for handling uncertainties. The system focuses on conveying an accurate estimate of these uncertainties on processes and uncertainties that occur within its natural language processing components. Text data analysis is ambiguous and error prone, but is nevertheless an important part of the data analysis. Through its innovative handling of uncertainties, VAPD enables transparent and reliable decisions based on uncertainty-aware visual analytics.

Index Terms—Uncertainty, Provenance, Trust-Building, Crime Analysis.

1 Introduction

In many different application domains, data analysts and decision makers face complex data, complex application problems, and an increasingly complex data analysis process. Based on the outcomes of the data analysis, decisions are made that affect not only automated processes, but also humans.

An application domain where the need for uncertainty aware decision making is immediately clear is criminal data analysis. Data analysts make sense out of huge amounts of criminal reports and intelligence data, which exceed the manual information comprehension capabilities of the involved humans. As a consequence, automatic methods to find patterns, identify similar crimes, or process text data from crime reports are deployed. With the increased usage of automatic data analysis methods and the rise of technical complexity, the technical competence of humans is exceeded too. This forces users to trust the output of automatic data analysis methods without being able to comprehend the data used as the foundation for decisions. A common application of crime data analysis is the *CCA* (*Comparative Case Analysis*), which is the process of "determining whether multiple crimes have been committed by the same offender" [1].

In this work, we depict how state of the art guidelines for uncertainties by Sacha et al. [4] can be implemented in a visionary prototype for criminal data analysis, VAPD.

• *Data Analysis and Visualization Group, University of Konstanz. E-mail: forename.lastname@uni-konstanz.de*

Manuscript received 31 Mar. 2015; accepted 1 Aug. 2015; date of publication xx Aug. 2015; date of current version 25 Oct. 2015.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

2 Uncertainty, Awareness, and Trust in Visual Analytics

Sacha et al. [4] describe the role of uncertainty, awareness and trust in visual analytics. Uncertainty propagates from its origin (the data source) through further transformations within a VA system to the humans trust building process. The latter is affected by the user's awareness of these uncertainties. Configurations between system uncertainties and the humans awareness of these uncertainties are classified. The chance of errors is highest when the analyst is unaware of uncertainties or mistakenly believes that there are no uncertainties in a system. In order to support trust calibration between human and machine, the authors provide guidelines and examples for handling uncertainties:

G1: *Quantify Uncertainty in Each Component*
G2: *Propagate and Aggregate Uncertainties*
G3: *Visualise Uncertainty Information*
G4: *Enable Interactive Uncertainty Exploration*
G5: *Make the System Functions Accessible*
G6: *Support the Analyst in Uncertainty Aware Sensemaking*
G7: *Analyse Human Behaviour in order to Derive Hints on Problems*
G8: *Enable Analysts to Track and Review their Analysis*

3 VAPD - The System

In the following sections, we outline the main components of VAPD, introduce possible sources of uncertainty within the data processing pipeline, and highlight application of the guidelines by Sacha et al. [4]. In addition, we introduce different kinds of provenance used appropriately to capture and leverage uncertainty for decision making.

3.1 Data Analysis and Uncertainty

To process crime data, we illustrate VAPD as a three component data analysis pipeline (middle of Figure 2).

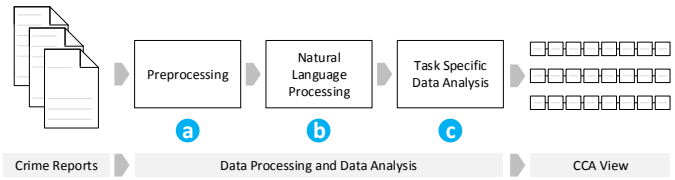

Fig. 2. Illustration of the VAPD data analysis process. Starting with the crime reports on the left, a three stage data processing pipeline is used to generate the CCA table view.

The first part of the pipeline is a preprocessing module, capable of dealing with normalisation of input text data, strategies to cope with missing values, and data/application dependent data characteristics (Figure 2 a). Removal of parts of the input data is common, e.g. to remove non-printable characters or doubled words and this introduces additional data uncertainties. As does modifications of the input data to, for example, cope with missing numerical values. Sophisticated text preprocessing, e.g. automated spell checking, is also difficult because state of the art spell checkers have dictionaries created from a selection of documents from different domains, and hence do not match the application case terminology.

The second component, natural language processing (NLP), is operating on the text data beyond statistics generation (Figure 2 b). State of the art methods use models that describe the word contexts or typical occurrences together with labels, e.g. part of speech tags. To reach reasonable accuracy in the application of NLP methods, it is crucial that the input data corresponds to the training data to the largest possible extent. Methodically, there are also probabilistic methods used (e.g. topic modeling) that produce different results although they are initialized and applied with the same parameters. The implementation of NLP including uncertainty quantification and aggregation is still an open research topic. Grimmer and Stewart [3] highlight consequences and mistrust from domain experts because of this issue.

Lastly, VAPD uses an application specific data analysis component operating on the output of the preprocessor and natural language processing (Figure 2 c). The data analysis is based on the best knowledge and mental models of data analysts, but depends on the output of the preceding data analysis – and therefore also incorporates their uncertainties. Depending on the analysis tasks, the guidelines G1 and G2 are implemented.

To be able to measure the uncertainty coming from the data analysis, each involved component needs to be able to quantify uncertainty caused by its operations. In the text data processing, as well as the high level processing pipeline view, VAPD keeps track of and aggregates these uncertainties using the concept of data provenance. This makes sure that the origin of uncertainty, as well as the changes from analysis component to analysis component can be viewed and explored separately from each other. With respect to the work of Sacha et al. [4], data provenance for uncertainty is the method of choice to implement G2, *propagate and aggregate uncertainties*.

3.2 Visualisation and Visual Analytics

In this section we describe VAPD's user interface that consists of several visualisations.

3.2.1 CCA Table View

The CCA Table View enables the user to solve the actual analysis task of identifying and comparing similar crimes. The standard table representation is enriched with an aggregated uncertainty measure for each crime record that is provided by the analysis pipeline (G3). A first approach is to map the uncertainty information to the visual variable of colour (red, as shown in Figure 1). This uncertainty information is updated when the analyst changes the aggregation weights in the uncertainty explorer in order to understand the impact of different uncertainties (G4). Furthermore, it is possible to order or filter the crimes according to the crimes uncertainty measure. Finally, an analyst is able to explicitly mark good or bad results according to his expert knowledge (G5) for adjusting the similarity model automatically.

3.2.2 Uncertainty Explorer

This component is inspired by Correa et al.'s framework for uncertainty aware visual analytics [2] and provides several uncertainty plots (e.g. projections) to investigate input sensitivity and impacts of uncertainty sources (G4). In addition, the analyst can interactively apply weightings or filters to each uncertainty dimension that will trigger an uncertainty information update within the whole system. Importantly, the uncertainty explorer offers an analysis pipeline visualisation indicating which (e.g., NLP-) components introduce uncertainties.

3.2.3 Reasoning Space

This component supports uncertainty aware sensemaking (G6) by integrating hypothesis and several pieces of evidence that have been gained by using the system. Uncertainty awareness will be raised by adding cognitive cues (uncertainty information, trust measures) to the concepts within the reasoning space. Also, an analyst may explicitly rate hypotheses, findings or insights in order to externalise his trust status. Furthermore, this component will automatically calculate and relate uncertainty and trust measures (e.g., by counting the amount of supporting or contradicting evidence) as an unbiased counterpart to the human (G7). In addition, if items or dimensions of interest have been detected based on analysing analytic provenance trails, the system may automatically come up with alternative model configurations that may prove or disprove the analysts expectations. In the case of NLP-models only the result with highest probability is chosen and visualised. The system could also reveal the alternative model results that have gained high probability. Finally, the reasoning space tracks and visualises the entire analysis process and integrates exploration and verification phases. This enables the analyst to review and rethink his knowledge generation process (G8).

4 Discussion & Challenges

The VAPD system presented in this paper is a combination of several existing ideas, examples and prototypes that tackle the problem of handling uncertainties for uncertainty aware decision making in visual analytics. In our opinion the benefit comes with the integration of these approaches as illustrated for the *CCA* case. In the future, it will be important to discuss technical details and challenges for realising the VAPD vision. In particular, the description of the quantification and aggregation of uncertainties in the processing pipeline and how this can concretely be transferred to the visualisations, is curtailed due to limited space within this two page paper. However, VAPD offers a concrete example with which to discuss this topic with experts in the domain and improve the current state of the art.

Acknowledgments

This work was partly supported by the EU project Visual Analytics for Sense-making in Criminal Intelligence Analysis (VALCRI) under grant number FP7-SEC-2013-608142 and the German Research Foundation (DFG) project "Feature-based Visualization and Analysis of Natural Language Documents" (VisADoc).

References

[1] C. Bennell, J. Woodhams, and R. Mugford. Linkage analysis for crime. In G. Bruinsma and D. Weisburd, editors, *Encyclopedia of Criminology and Criminal Justice*, pages 2947–2953. Springer New York, 2014. (document)

[2] C. Correa, Y.-H. Chan, and K.-L. Ma. A framework for uncertainty-aware visual analytics. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 51–58. IEEE, 2009. 3.2.2

[3] J. Grimmer and B. M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21:267–297, 2013. 3.1

[4] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of the Visual Analytics Science and Technology)*, 23(01), January 2016. (document), 2, 3, 3.1