Ensuring Veracity in Heterogeneous Data Mining

Douglas Fraser
UEA Registration: 100189521
CMP-7023B – First Assessed Exercise

February 16, 2017

1 Introduction

The problems of data mining Big Data were first summarized in three words: Volume, Velocity, and Variety. However, as data mining was applied to disparate sources simultaneously, e.g. the Web, a fourth V was identified: Veracity (Yin et al., 2008). Veracity can be roughly thought of as arising out of the intersection of Volume and Variety. More data, of different types from different sources, structured differently (or only somewhat), and variable in numerous ways. This heterogeneous data complicates the work associated with the data fusion process. Given that traditional data mining algorithms work under the assumption of independent homogeneous samples, how can the unification process ensure the veracity of the resulting data? This paper briefly examines the ways heterogeneity and veracity intersect, providing examples of data related conflicts and the latest research into ways to resolve them.

2 Heterogeneity Related Conflict

Heterogeneity is a multifaceted concept and dealing with it requires various resolution strategies. These strategies depend on the specifics of the data and the data mining project. Pluempitiwiriyawej and Hammer (2000) divide heterogeneity related conflicts into three broad classes: structural, domain, and data related. Additionally, Veracity is associated with several concepts which reflect a different aspect: truthfulness, reliability, and quality (or accuracy). These concepts and what's required to ensure them impact the resolution of heterogeneity related issues. This results in a variety of tasks needed in data pre-processing to sufficiently clean it up before analyzing it and creating models.

To elucidate these tasks, a simple example of a data mining project is used to illustrate how different conflicts arise and could be dealt with to ensure veracity. It is based on credit agencies and the data management issues they face, e.g. Experian is somewhat notorious for a low quality of data in their systems resulting in numerous lawsuits over the years

(OnlineAthens, 2014). Recent research into how to resolve the different conflicts for real world problems is also briefly mentioned

2.1 Structural Conflict

The most obvious conflict in cleaning and unifying heterogeneous data results from schematic differences between data sets. Examples would be a field in one schema missing from the other schema or fields with similar or equivalent names. Another common issue is a field in one schema that maps to several in the other schema. All are sources of inaccuracies within the ideal unified set of data to be used by data mining algorithms.

For a credit agency, an example of schema conflicts would be in the following situation. Assume data from two companies is being used in a data mining project. Both schemas have a "name" field, but one company stores the person's entire name within the field. Unfortunately, the other company stores only the first name in that field because "surname" is the field containing the last name. Another example would be a difference in the sets of values used to denote the range of a person salary - one company could have a finer granularity to the scale than the other and these scales do not logically mesh together. The obvious solution is to define a mapping between the two schemas along with functions for cleaning and regularizing the data.

Schema matching is a well researched problem and many techniques and algorithms have been devised. One would be the HDSM algorithm (Chen et al., 2012) which can handle not only 1:1 schema mapping, but n:1, 1:n, and m:n. But with the advent of Big Data and the complexity of heterogeneous data, fully automatic algorithms are needed in order to cope with the growing workload. What if a data mining process was used on the schemas to determine mapping functions and field compatibilities? Shelake and Bhojane (2013) describe a system that uses n-grams, natural language processing techniques and data value analysis to determine how to map schemas. An obvious extension to this research would be to utilize metadata about the schemas as part of the process. This could possibly be combined with the techniques described in Nuamah et al. (2016) to make logical inferences based on the metadata and the schema data mining to further improve performance.

2.2 Domain Conflict

Domain conflicts result from semantic related differences between data sets. To wit, there is a discrepancy in how the data in associated fields for multiple schemas is interpreted. This can result from the units used for numerical data being different, e.g. millimeters versus inches. Thus, combining the values is not possible without first transforming one. Alternatively, the numerical precision could differ, resulting in a loss of accuracy if data is unified without considering the ramifications. Finally, the last possible type of domain conflict comes about from formatting issues. The data values are compatible, but there are different conventions used in their representation, resulting in accuracy and quality problems if they are not considered.

These three issues are easily described again using the credit agency example. Salary data for a person could be stored as a whole number by one company, but as a real number by the second. In the case of international companies, salary data would logically be stored based on the local currency of each branch. A data mining project that uses data from multiple countries would need to perform exchange rate calculations before creating an unified view. Finally, there are different standards per region that specify what characters are used for decimal points, e.g. America uses periods for decimal points, but European countries use commas. The reverse applies for denoting groups of 1000, thus care must be taken in merging data.

A key technique in resolving semantic conflicts, and thus improving data accuracy and quality, is through the use of ontologies. An ontology clarifies the set of concepts underlying the domain, their properties, and their linkage. With this information, decisions on how to resolve semantic conflicts can be made. An early example of research into ontologies and automated semantic conflict resolution for heterogeneous data is in Ram and Park (2004). Another use of ontologies is that they can define constraints for data to assist in detecting outliers and false values (Ristoski and Paulheim, 2016). As the Semantic Web grows, ontologies and metadata will be essential for successful automation of data mining and the KDD process.

2.3 Data Conflicts

This third category relates to issues with the values of data fields that have to be merged. If data for a field from one source is distinctly different from the data from other sources, the conflict needs to be resolved in a logical way. For homogeneous numerical data, statistical techniques can determine if certain values are merely acceptable outliers or if the data should be considered suspect. This is a reflection of the accuracy aspect of veracity. For categorical data, data corruption may result in misspellings of labels which are easily handled through a simple mapping process to resolve accuracy issues. But what if multiple data objects need to be reconciled? This is different than independent homogeneous data points. In Big Data applications involving many sources that may conflict, evaluating the truthfulness or reliability of them is important towards creating a single representation (Sun et al., 2012).

One of the conceptual advances in dealing with data mining heterogenous data is to regard it as an information network (Han and Gao, 2009). Viewing data objects as connected with links provides a rationale for using graph type algorithms to mine the link associated information. This is the basis for the truth discovery type algorithms that resolve heterogeneous data related conflicts by traversing the graph and calculating reliability estimates, Liu et al. (2016) and Li et al. (2016) being two recent examples. Once object reliability has been estimated, then data values can be further processed to arrive at some type of consensus depending on the needs of the application, e.g. discarding data sources deemed to be too unreliable. Another possible application for network based algorithms is to aid in data imputation – the most likely values for empty data fields of a specific object could be inferred by examining the fields of the entire set of similar and related objects linked to it.

To illustrate the concept of source reliability and veracity for the credit agency example, imagine another credit agency has been bought by Experian and so database records need to be merged. There are multiple records for a man named John Doe, some associated with one Social Security Number (SSN) and some with another. Due to data entry errors, the SSN is incorrect for some of Experian's records. If records are blindly merged based only on a few key fields like name and SSN, with no thought to establishing reliability, one John Doe could end up being associated with the criminal history records of a second John Doe. But by considering the records of address changes of the two men and their connected master records, the veracity of both referring to the same man could be established. In this manner, problems with data cross-contamination could be prevented. Also possible is establishing the reliability or veracity of specific fields used as foreign keys, e.g. the SSN. Unfortunately, this example is not merely a theoretical one.

3 Conclusion

Establishing the veracity of information and resolving heterogeneous data conflicts are already important concerns in various Big Data related applications. They will become even more important once the Internet of Things concept becomes a reality. There is no chance of a global data schema standard for IoT devices, nor is the data reported guaranteed to be always accurate and reliable. Deliberate corruption of data will also be a concern, especially in a world where Internet related security is already a significant issue. Fortunately, the fundamental issues around heterogeneity and veracity are understood. Based on the research reviewed, two basic ideas that will drive further developments are better metadata and more abstraction. Better metadata, e.g. about data schemas, will allow for automated inference based reasoning to solve issues related to semantic and schematic heterogeneity. More abstraction, such as viewing heterogenous data as linked in an information network, will enable new insights to be developed. The result will be more automated and intelligent algorithms leading to more sophisticated processes and applications. These will be needed to deal with the zettabytes of information that the KDD community will face.

References

Chen, W., Guo, H., Zhang, F., Pu, X., and Liu, X. (2012). Mining Schema Matching Between Heterogeneous Databases. In *Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on*, pages 1128–1131. IEEE.

Han, J. and Gao, J. (2009). Research Challenges for Data Mining in Science and Engineering. *Next Generation of Data Mining*, pages 1–18.

Li, Y., Li, Q., Gao, J., Su, L., Zhao, B., Fan, W., and Han, J. (2016). Conflicts to Harmony: A Framework for Resolving Conflicts in Heterogeneous Data by Truth Discovery. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):1986–1999.

Liu, W., Liu, J., Duan, H., He, X., and Wei, B. (2016). Exploiting Source-Object Network to Resolve Object Conflicts in Linked Data. *arXiv preprint arXiv:1604.08407*.

Nuamah, K., Bundy, A., and Lucas, C. (2016). Functional Inferences over Heterogeneous Data. In *International Conference on Web Reasoning and Rule Systems*, pages 159–166. Springer.

OnlineAthens (2014). Lawsuit filed against credit reporting giant Experian. Retrieved from http://onlineathens.com/2014-06-16/lawsuit-filed-against-credit-reporting-giant-experian.

Pluempitiwiriyawej, C. and Hammer, J. (2000). A Classification Scheme for Semantic and Schematic Heterogeneities in xml Data Sources. *TR00-004, University of Florida, Gainesville, FL*.

Ram, S. and Park, J. (2004). Semantic Conflict Resolution Ontology (SCROL): An Ontology for Detecting and Resolving Data and Schema-Level Semantic Conflicts. *IEEE Transactions on Knowledge and Data engineering*, 16(2):189–202.

Ristoski, P. and Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on the World Wide Web*, 36:1 – 22.

Shelake, V. M. and Bhojane, V. S. (2013). A Novel Approach for Multi-Source Heterogeneous Database Integration. In *Machine Intelligence and Research Advancement (ICMIRA), 2013 International Conference on*, pages 184–190. IEEE.

Sun, Y., Han, J., Yan, X., and Yu, P. S. (2012). Mining Knowledge from Interconnected Data: A Heterogeneous Information Network Analysis Approach. *Proceedings of the VLDB Endowment*, 5(12):2022–2023.

Yin, X., Han, J., and Philip, S. Y. (2008). Truth Discovery with Multiple Conflicting Information Providers on the Web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808.