ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

Inconsistency Resolution and Joint Information Extraction from Multiple Data Sources

I. Carol[1], Dr. S. Britto Ramesh Kumar[2]

Research Scholar, Department of Computer Science, St. Joseph's college (Autonomous), Trichy, Tamil Nadu, India[1]

Assistant Professor, Department of Computer Science, St. Joseph's college (Autonomous), Trichy, Tamil Nadu, India[2]

Abstract: Web content mining describes the discovery of useful information from the Web contents/data/documents/information. The presence of useful information in different repositories is undeniable. The aim of Information Retrieval (IR) is to retrieve the information by eliminating factors such as noise, redundant and irrelevant from the repository. Retrieval of information from multiple data sources increases the complexity. In the time of IR different type of problem arrived. Inconsistency is one such important difficulty faced by IR in multiple data sources. Different solutions have been proposed with varied levels of results. The aim of the paper is to provide a solution for the inconsistency problem using Semantics, Possibility and Para-Consistent Logic, Meta Data and User Relevance Feedback in web content mining.

Keywords: Web content mining, Information retrieval, Multiple data sources, Inconsistency

I . INTRODUCTION

Web Mining is the use of the data mining techniques to automatically discover and extract information from web documents/services. In this paper, we present an overview of challenging issues in web mining. We discuss mining with respect to multiple data sources referred here as we mining. In particular, our focus is on different problems in multiple data sources (MDS). And how to solve the inconsistency problem on MDS.

Web mining could be classified into three categories: Web content mining, Web structure mining, and Web usage mining

Web content mining tries to discover valuable information from Web contents (i.e. Web documents). Web content is mainly referred to textual objects. The web content data consist of unstructured data such as free texts, semi-structured data such as HTML documents, and structured data such as data in the tables or database generated HTML pages. Advantages of Web Content Mining are, It helps in structuring the information from the web, it has the capability to mine all types of documents, it can be extended to use graph theory concepts and extract structured, semi-structured information. It can be used to create visual tools that can be used for web information extraction.

Web structure mining involves in modelling Web sites in terms of linking structures. It is used to construct Web page communities or find relevant pages based on the similarity or relevance between two Web pages. The challenge for Web structure mining is to deal with the structure of the hyperlinks within the Web itself.

Web usage mining tries to reveal the underlying access patterns from Web transaction or user session data that recorded in Web log files. Capturing Web user access interest or pattern can not only provide help for better understanding user navigational behaviour, but also for efficiently improving Web site structure or design. As more organizations rely on the Internet and the World Wide Web to conduct business, the traditional strategies and techniques for market analysis need to be revisited in this context.

A. Information retrieval

IR is used to retrieve all relevant documents automatically while at the same time retrieving as few of the non-relevant as possible. Actually IR has the primary goals of indexing text and searching for useful documents in a collection and now a day's research in IR includes modeling, document classification and categorization, user interfaces, data visualization, filtering, etc.(data are stored in multiple places. If one query put in to the IR system, that IR system searches data from multiple places. That is called Multiple Data Sources (MDS).

II. REVIEW OF LITERATURE

The literature review has been made in detail on web content mining and MDS. The existing architectures, models and algorithms of various authors are also analyzed.

Huiwen Fu et al. [6] stated that the sheer scale of web and diverse authorships, caused use of different syntaxes to express similar or related information. They proposed a technique to extract information from multiple sites. Still two major problems related to the Web content mining arose (1) Web query information integration, to enable querying multiple Web databases (which are hidden in the deep Web) (2) Schema matching e.g. integrating Yahoo and Google's directories to match concepts in the hierarchies.

M.S. Shashidhara et al. proposed a method comprising of three phases of extraction namely selection, pre-processing and presentation of useful documents from streaming on-

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

line sources. It is evident that the extraction of knowledge was considerable to an extent, yet this method does not address MDS.

R. Kosala et al. [10] regarded that in the recent years, useful patterns were discovered from Multiple Data Sources (MDS) and recognized it as a critical and hot research area in data/web mining and machine learning.

V. Lesser et al. [2] opined that mining in MDS benefits the industries and businesses in easier and smarter use of information.

H. Zhao et al. [3] expounded in their research that information repositories are composed of heterogeneous and autonomous information sources. Nevertheless, the users expected that the entire collection as a single source would disclose a single, unambiguous answer for queries

However, they identified 'information conflicts' as an important issue in information integration from the various information sources.

Xingquan Zhu et al.[8] propounded a method for retrieving useful patterns and implement them. Although they retrieved, it became an expensive and perhaps even impossible to implement due to the presence of following difficulties in MDS:

Heterogeneity, Incompleteness, Inconsistency, Large Size.

Adhikari et al. [7] proposed a Pipelined feedback technique of mining multiple databases to improve significantly the accuracy of mining. They segregated the large databases by dividing them into sub-databases to enhance the accuracy.

However, the presence of local patterns in data sources produced inconsistent and conflicting inferences thereby preventing the identification of global interesting patterns. This paves a large gap between the reality (the raw data sources being inconsistent) and the objectives of mining on MDS.

S. Zhang et al.[18] is discovering knowledge from MDS centralized their research around Mono-Database Mining to discover patterns that are globally significant among participatory data sources. Causally, they were effective in reducing the search cost.

It is the case that for the following limitations are found in Mono-Database Mining:

- Complexity of resulting Data, Curse of Higher Dimensions
- Loss of some useful Patterns, Integration Problems
- Inconsistency among the Data Sources

Muhammed Fahad et al. [1] contributed towards automatically identifying semantic inconsistencies through a mechanism that detected semantic heterogeneities.

All the same, Application Dependence and need for Multiple Scans for each application arose as impedance in resolving inconsistencies.

S.C. Zhang et al. [9] proffered a technique namely, Local Pattern Analysis (LPA) to resolve inconsistency, where they clustered the data sources and mined for knowledge from individual relevant data sources. Finally, they integrated the knowledge from the data sources.

This technique is application-independent and is based on global pattern discovery that makes it efficient strategy for multiple data source discovery. Nonetheless, the frequency of mining various DB in LPA strategy is accounted as a limitation.

III. METHODOLOGY

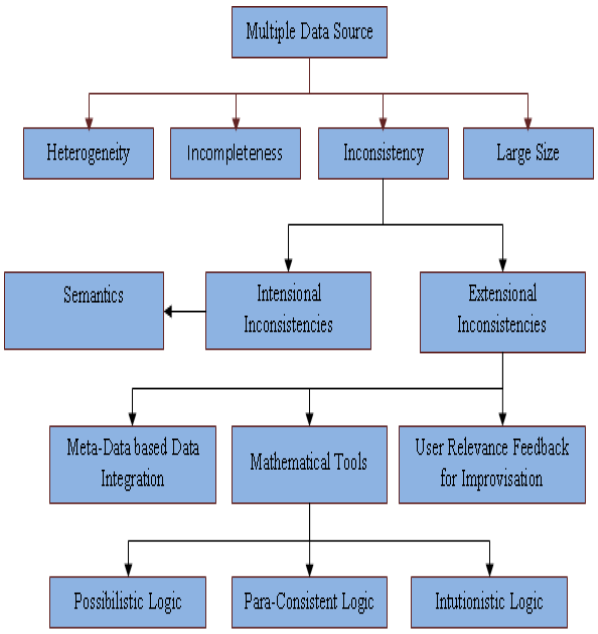

Fig 1. Inconsistency for MDS - Methodology

The Inconsistency Problem

All of the above approaches are based on the assumption that the data to be mined are already of high quality. The systems then only focus on identifying useful patterns at the attribute-value level. Unfortunately, in the real applications data sources are often inconsistent and conflicting.

A. Intensional Inconsistencies

The sources are in different data models. They have different schemas within the same data model, and their data is represented in different natural languages or different measurement systems. Such conflicts have often been termed semantic inconsistencies. Almost any mining system will address this issue when integrating the data in the pre-processing step.

B. Extensional Inconsistencies

There are factual discrepancies among the sources in data values that describe the same objects. Such conflicts are also referred to as data inconsistencies. No data mining system addresses this problem extensively, because Extensional inconsistencies can only be observed after Intensional inconsistencies have been resolved.

Possibilistic logic extends classical logic by considering classical formulae associated with certainty levels and allows us to compute global inconsistency levels for such

knowledge bases. Interestingly, the decision form of the deduction problem in standard possibilistic logic is not harder than deduction in classical propositional logic.
Para-consistent logic aims to handle inconsistent pieces of information by isolating them.A Para-consistent logic is a logical system that attempts to deal with contradictions in a discriminating way. [19][20].
Alternatively, Para-consistent logic is the subfield of logic that is concerned with studying and developing Para-consistent (or "inconsistency-tolerant") systems of logic.

Intuitionistic logic, or Constructive logic, is a symbolic logic system that differs from classical logic in its definition of what it means for a statement to be true.
In classical logic, all well-formed statements are assumed to be either true or false, even if we do not have a proof of either.
In constructive logic, a statement is "true" only if there is a constructive proof that it is true and "false" only if there is a constructive proof that it is false. Operations in constructive logic preserve justification, rather than truth.
Our term for such meta-data is information features which includes
Timestamp: The time when the information in the source was validated. Latest information could be provided more importance.
Cost: The time it would take to transmit the information over the network, or the money to be paid for the information, or both. Cost-Effectiveness could be achieved.
Accuracy: Probabilistic information that denotes the accuracy of the information.
Availability: The probability that at a random moment the information source is available. This provides reliability.
Clearance: The security clearance level needed to access the information. Provides privacy protection.
Inconsistency resolution is a process in which users must be given a prominent voice.
Depending on their individual preferences (which are subject to individual situations), users must be allowed to decide how inconsistencies should be resolved.
Implemented based on two options are Thresholds of Acceptance (to eliminate data that is too old, too costly or lacking in authority, etc.) and Feature Weighting Approach (based on data quality and utility to the user). Through the use of the above mentioned techniques including Semantics, Possibility and Para-Consistent Logic, Meta Data and User Relevance Feedback, we can eliminate the inconsistencies in the Information Retrieval process involving Multiple Data Sources (MDS) that are unstructured, which is typical in several Web Mining scenarios. This makes Inconsistency Resolution and Joint Information Extraction a possibility.

IV. CONCLUSION

Through the use of the above mentioned techniques including Semantics, Possibility and Para-Consistent Logic, Meta Data and User Relevance Feedback, we can eliminate the inconsistencies in the Information Retrieval process involving Multiple Data Sources (MDS) that are unstructured, which is typical in several Web Mining scenarios. This makes Inconsistency Resolution and Joint Information extraction.

REFERENCES

[1] Muhammad Fahad, NejibMoalla and AbdelazizBouras, "Detection and resolution of semantic inconsistency and redundancy in an automatic ontology merging system", J IntellInfSystSpringerScience+Business Media, LLC (2012) 39, pp 535–557.
[2] V. Lesser, B. Horling, F. Klassner, A. Raja, T. Wagner and S. Zhang, "BIG: an agent for resource-bounded information gathering and decision making", Artificial Intelligence vol. 118 (1–2), 2000, pp 197–244.
[3] H. Zhao and S. Ram, "Entity matching across heterogeneous data sources: an approach based on constrained cascade generalization", Data & Knowledge Engineering, Vol 66, Isssue 3, September 2008), pp 368–381.
[4] David Sánchez, Montserrat Batet, David Isern and Aida Valls, "Ontology-based semantic similarity: A new feature-based approach", Expert Systems with Applications (ScienceDirect) 39, 2012, pp 7718–7728.
[5] B. Turhan and A. Bener, "Analysis of naive Bayes' assumptions on software fault data: an empirical study", Data & Knowledge Engineering vol. 68 (2), 2009, pp 278–290.
[6] Huiwen Fu, DingrongYuan and Xiaomeng Huang, "Mining indirect association rules in multi-database ", System Science Engineering Design and Manufacturing Information (ICSEM).2012,3rd International Conference (Volume:2) .
[7] AnimeshAdhikari, PralhadRamachandrarao, Bhanu Prasad, and JhimliAdhikari, "Mining Multiple Large Data Sources", The International Arab Journal of Information Technology, Vol. 7, No. 3, July 2010.
[8] XingquanZhu, Bin Li a, Xindong Wu, Dan He and Chengqi Zhang, "CLAP: Collaborative pattern mining for distributed information systems", Decision Support Systems (ScienceDirect)52, 2011, pp 40–51.
[9] S.C. Zhang and M.J. Zaki, "Mining multiple data sources: local pattern analysis", Data Mining and Knowledge Discovery 12, 2006, pp 121–125.
[10] R. Kosala, H. Blockeel, "Web Mining Research: A Survey", KDD00, 2000, pp. 1–15.
[11] DietmarJannacha,KostyantynShchekotykhinb, Gerhard Friedrichb, "Automated ontology instantiation from tabular web sources"TheAllRight system, Web Semantics: Science, Services and Agents on the World Wide Web 7, 2009, pp 136–153
[12] Ivan Habernala and MiloslavKonopík b, "SWSNL: Semantic Web Search Using Natural Language", Expert Systems with Applications 40, 2013, pp 3649–3664.
[13] S. Benferhat, D. Dubois and S. KaciH. Prade, "Possibilistic merging and distance-based fusion of propositional information", International Journal of Annals of Mathematics and Artificial Intelligence 34 (1–3), 2002, pp 217–252.
[14] S. Benferhat, D. Dubois and H. Prade, "From semantic to syntactic approaches to information combination in possibilistic logic, in: B. Bouchon-Meunier (Ed.)", Aggregation and Fusion of Imperfect Information, Studies in Fuzziness and Soft Computing, PhysicaVerlag, 1997, pp. 141–151.
[15] D. Dubois, J. Lang and H. Prade, "Possibilistic logic", Handbook of Logic in Artificial Intelligence and Logic Programming, vol. 2, 1994, pp. 439–513.
[16] AmihaiMotro and Philipp Anokhin, "Fusionplex: Resolution of Data Inconsistencies in the Integration of Heterogeneous information Sources INFS-797/IT-803: Information Integration and Interoperation, Vol. 7, No. 2, June 2006.
[17] A. Hunter, "Evaluating the significance of inconsistencies", Proceedings of the International Joint Conference on AI(IJCAI'03), 2003, pp. 468–473.ed Logic, Reidel, 1977, pp. 8–37.
[18] S.C. Zhang, C.Q. Zhang and X.D. Wu, Knowledge Discovery in Multiple Databases, Springer-Verlag, June 2004, pp 234.
[19] M. Kifer, E.L. Lozinskii, "A logic for reasoning with inconsistency", Journal of Automated Reasoning 9 (2), 1992, 179–215.
[20] Zhang, Z., He, B., Chang, K. C.-C. Understanding Web Query Interfaces: Best-Effort Parsing with Hidden Syntax.SIGMOD-04, 2004.

[21] Tak-Lam Wonga, Wai Lam b, "An unsupervised method for joint information extraction and feature mining across different Web sites", Data & Knowledge Engineering 68, 2009, pp 107–125.
[22] Derek Sleemana, Laura Mossa, Andy Aikena, Martin Hughesb, John Kinsellab and Malcolm Simb, "Detecting and resolving inconsistencies between domain experts' different perspectives on (classification) tasks", Artificial Intelligence in Medicine 55, 2012, pp 71–86.
[23] S.Mahesha, Dr. M S Shashidhara and Dr. M. Giri, "An Efficient Web Content Extraction Using Mining Techniques", International Journal of Computer Science and Management Research, 2012.

BIOGRAPHIES

I. Carol is pursuing Doctor of Philosophy in Department of Computer Science, St. Joseph's College, (Autonomous), Tiruchirappalli, Tamil Nadu, India. He received his M.Phil degree from St. Joseph's College, Tiruchirappalli. He received his MCA degree from St. Joseph's College, Tiruchirappalli. He has published many research articles in the International conferences and journals. His area of interest is Data mining, Web mining.

Dr. S. Britto Ramesh Kumar is working as Assistant Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has published many research articles in the National/International conferences and journals. His research interests include Data Mining, Web Mining, and Mobile Networks.