Fuzzy Metadata Strategies for Enhanced Data Integration

Hiba Khalid[1,2], Esteban Zimanyi[1] and Robert Wrembel[2]
[1]*Department of Code, ULB, Brussels, Belgium*
[2]*Department of Informatics, PUT, Poznan, Poland*

Keywords: Fuzzy Sets, Fuzzy Union, Crisp Sets, Metadata, Data Integration, Knowledge Base.

Abstract: The problem of data integration is one of the most debated issues in the general field of data management. Data Integration is typically accompanied by a concept of conflict management. The problem's root arises from different data sources and the probability of how each data source corresponds to another. Metadata is also another important yet, highly overlooked concept in these research areas. In this paper we propose the idea of leveraging metadata as a binding source in the process of integration. The research technique relies on exploiting textual metadata from different sources by using Fuzzy logic as a coherence measure. A framework methodology has been devised for understanding the power of textual metadata. The framework operates on multiple data sources typically a data source set can contain 'n' number of datasets. In case of considering two data sources the sources can be titled as primary and secondary. The primary secondary source is the accepting data source and thus contains more enriched metadata. The secondary sources are the requesting sources for integration and are also guided by textual data summaries, keywords, analysis reports etc. The Fuzzy MD framework operates on finding similarities between primary and secondary metadata sources using fuzzy matching and string exploration. The model then provides the probable answer for each set's association with the primary accepting source. The framework relies on origin of words and relative associativity rather than the common approach of manual metadata enrichment. This not only resolves the argument of manual metadata enrichment, it also provides a hidden solution for generating metadata from scratch as a part of the integration and analysis process.

1 INTRODUCTION

Data integration is a concept that revolves around the idea of combining two or more data sources based on the criteria of relevance and the probability of a possible match. The idea is somewhat simple however, the reality of data integration is far more difficult and time consuming. Enormous amount of data is generated every single day. This data could be in structured or unstructured representation. There is no actual physical binding between the "N" number of data sources. To integrate related sources or to even generate a simple representation of linkage a lot of man and machine power is required. This increases the cost of data integration and processing projects. Thus, to understand the data sources more extensively and to decrease the machine cost involved in analysing and matching databases a more comprehensive solution-based resolution must be designed. Some of the most common data integration techniques are:

- Data consolidation
- Hybrid Data Integration
- Extract Transform Load (ETL)
- Enterprise Information Integration Technologies (EII)
- Data Federation
- Data Propogation
- Enterprise Application Integration (EAI)

The aforementioned technologies do manage everyday business prospects of data integration and resolution. The spectrum and scope of available technologies does not efficiently correspond to changing and variable data forms, data structures and most importantly big data. For instance, in the case of semantic web. The increasing amount of semantic data cannot be linked together using traditional approaches. Also, the amount of data and irregularity requires methodologies that can both adapt and change according to specific application, scientific of business requirements.

Khalid, H., Zimanyi, E. and Wrembel, R.
Fuzzy Metadata Strategies for Enhanced Data Integration.
DOI: 10.5220/0006905200830090
In *Proceedings of the 7th International Conference on Data Science, Technology and Applications (DATA 2018)*, pages 83-90
ISBN: 978-989-758-318-6

Metadata is one aspect that is common to all forms of data sources. Traditional, relational, graph, columnar etc. All data sources do have a common representation and idea regarding metadata. The most crude yet accurate definition of metadata is "data about data". Another representation of metadata can be regarded as information that elaborates the content inside a data source or datasets. Metadata has its many representations and forms. Some of the most common metadata types include descriptive, structural and administrative. Each type serves its own purpose and complexity. Descriptive metadata for instance deals with information that identifies an entity or objects or documents e.g. titles, keywords used, abstract of documents, summaries and authors or creators of a project for instance. The metadata is a powerful set of data that can be enhanced, enriched and exploited to develop reasonable insights into the data itself. Metadata provides an opportunity to resolve the data integration problem to a certain extent if properly cultured and deployed over the data sources. The metadata can be utilized to draw data based context summaries before the actual integration of sources. The only hurdle is the availability of good quality metadata. A good structured metadata can provide more useful insights into the data and can also look into the problems of conflict resolution based on the learning curve of metadata repositories.

The metadata has many forms and complexities associated with it. For the scope of this research the metadata type has been restricted to "textual" content. Thus, all metadata analysed, acquired and created is text based. No formal numerical representations have been considered for the initial research phase. The second most important concern regarding metadata management is the use of techniques, formulas or algorithms that can devise a learning and classification path for the metadata under consideration. Eurostat metadata is a very detailed and rich in contextual information regarding various European statistics. This metadata provides the opportunity to analyse and locate similar information between different sources for integration, processing, analysis etc.

Text processing is the primary research aspect for analysing textual metadata. The classification of keywords, meaning of different headers, titles, sources, date of publishing etc. all represent valuable and useful information. The text can be analysed using various algorithms and can be designated under different categories using predictive and classifying algorithms.

In this paper we propose a fuzzy based approach in developing metadata integration strategies along with Naïve Bayes representation for text analysis and classification. The new process differs in two major aspects as compared to previous approaches. First, the unstructured metadata documents can also be exploited using the fuzzy logic representation i.e. various classes have been designed for unstructured information to fall into. It defines a more open context for metadata to associate with. For example, a simple table summary of 20 lines without any title or any other description can be classified based upon "open fuzzy classes". The text in the description is analysed and a probabilistic match based on keywords extracted is sent through the framework pipeline for "possible matches" and "possible classes". Secondly, besides the set of keywords a set of synonyms is derived from dictionary and kept in a repository. This repository can then be accessed to find similar meaning words in other summaries of data sources. The repository also maintains a key-link or historical information. This is regarded as the learning curve for future work. For example. Each iteration of classification access the words, possible keywords, possible matches and repository synonyms. Once a match is successfully classified the original word "OW", the associated keyword "OKW" and the derived or accessed synonym "DSA" is kept as a link in the historical table. This table will be used as a feedback generator in the next phase of the research i.e. concerned with the learning curve of metadata repositories.

In summary the contributions of the research paper includes the following:

- We propose and formulate a new methodology for utilizing metadata for more accurate or enhanced data integration.
- We have deployed Fuzzy logic and Naïve Bayes as ground algorithms for proof of concept and algorithm efficiency monitoring.
- We conducted our experiments by utilizing the standardized and relevant Eurostat data and metadata. This provides the ground work for analytical analysis and validated experimentation

The remaining of the paper is organized as follows. The problem definition is formulated in section 2.1. In section 2.2 an overview of the designed Fuzzy metadata framework is provided with the details of input, processes and outputs obtained. The study variables and framework components are also discussed in this section. Section 2.3 discusses the methodology proposed from start till the end of framework and problem execution. Section 3 discusses the concepts are literature studied and utilized in the construction of this research. It reviews the existing work conducted in this field. Section 4

concludes the research presented and provides the next direction for this research as future work.

2 PROPOSED METHODOLOGY

In this section the proposed methodology is discussed with a Eurostat dataset metadata values and scenario. The dataset used for analysis is the Hi-tech and science from Eurostat. The metadata utilized corresponds to the same dataset with interlinking metadata to other connected datasets.

2.1 Problem Statement

The metadata management for data integration is a necessary and proficient process. However, the industry and application management does not confide to developing, producing or enriching existing metadata. The deployment and use of metadata can not only improve the data analytics industry, it can also improve the underlying problematic and costly process such as data binding, data integration, associativity, relevance structure, relational aspects and conflict resolution.

The metadata associated with different types of data can be treated in similar ways. The metadata represents information about data that can be exploited to obtain insights such as the degree of information available, the time and the actual content of the data available. Thus, by improving the metadata under consideration the participation of external entities can be improved and enhances. It is a topical data interaction. Only the top information layer communicates before an actual association can be constructed between two or more distinct data sources.

The ideology behind the purposeful deployment of metadata as a binding entity is to ensure a minimum data read and exploration run. Since, the primary problem addressed is the increasing data and the cost required to find similar, relevant and duplicate data. The metadata if properly cultured and generated can decrease the cost as well as improve the integration as well as conflict resolution process. The process involves many data pre-processing and data association techniques. It is to be kept in mind that the data in this context is “metadata”. The metadata in this context is treated as data to provide relevant insights into the functionality, relations and relevance to other metadata components.

The selected decision processing technique for metadata classification and association is Fuzzy logic. Fuzzy logic is a representative method of finding relevance of various items belonging to a certain category. The hard and fast rule of fuzzy logic states that any object, item or entity can show its relevance or contribution while lying with in two extremes 0 and 1. It is more efficient and useful than having crisp boundaries and thresholds. Because fuzzy logic provides the items that need to be categorized with more options to fall into. The traditional or classic categorization is based on the concept of truth and false i.e. either an item/product/element etc belongs in the set or does not belong in the set. Fuzzy logic provides an alternative and more appropriate solution by advising the degree to which an item might belong to a category or more than one category. Fuzzy logic operates on the functionality of “Partial Truth” i.e. a defined degree to which a certain element is associated with the fact or truth etc.

Fuzzy logic comprises of two very simple concepts Fuzzy sets and Fuzzy rules. Fuzzy sets like any other set are made up of collection of items. The distinct feature of fuzzy sets is that each item in this set belongs or represents a degree of association or linkage. For example if a set of players is created then the Fuzzy set will contain all players, age height etc. A crisp boundary is a hard or harsh line that behaves like a threshold meter. If a query like “who is a football player and who is a basketball player from Set A” is initiated. The crisp boundary will take the average threshold and classify players above a certain height with Basketball players because of a known fact. However, in Fuzzy sets a relational aspect can be developed. We can easily conclude that all players are tall to a certain height and from this information we can predict more accurate answers by not following an average calculation.

It also considers the exceptions and outlier cases such as a basketball player who might be a few inches shorter than average height. This is regarded as the “Fuzzy Knowledge”. This knowledge gathered from fuzzy sets can be effectively combined for enhanced predictions, classifications etc by using “Fuzzy Rules” to perform decision making processes based on the fuzzy set knowledge. In the context of our research the textual metadata composites (metadata composites: is a conceptual set of related metadata keywords combined into triples and quadruples depending upon their association such as synonyms can be combined into a triple meta-composite) and metadata sets can be created that can conform to a category based on the degree of relevance.

For example, From Eurostat dataset column headers can be extracted or profiled for metadata. These headers can provide insight into the kind of data that might reside in the rows. For example the

date column would be of type date. The countries column provides two insights i.e. the information is contextual and is possibly among the European countries. Now Fuzzy Categories can define the degree to which these can conform to a category. Suppose the Superlative Category is land area. Now, the date column header conforms least to land area. However, each country has different land areas and thus this does have a conforming degree say 0.6. Thus, metadata composites and fuzzy sets can predict and corelate information.

Fuzzy systems are designed on fuzzy logic that comprise of primarily four distinct parts:

- Fuzzifier: crisp data input set is converted into fuzzy set by using fuzzy linguistic variable and membership functions.
- Fuzzy Rules& Inference Engine: an inference on the fuzzy set is generated based on Rule base.
- Defuzzifier: The fuzzy output set is directed into a valued crisp output set by using the membership functions in defuzzification process.

The aforementioned linguistic variables in fuzzification step are terms or variables that contain words or sentences. Linguistic variables are both input and output variables containing sentences or words. For example from Eurostat metadata of Sciences and technology database is an extensive and detailed textual metadata representation. It also provides contextual and physical links to other related metadata. This is visible by simple analysis of the metadata file. For a system to recognize the headers, labels, columns and physical links is important. Thus, a fuzzy system allows the framework to draw conclusions based on degree of participation and relevance to fact. In our research context the very first query handles is based on employer search text query: "the availability of skilled data scientist in EU". This is a simple search query. The ideology is to take the context of the search query, examine the Eurostat metadata, generate relevance and find connected metadata that can answer this particular query.

The second phase is to take the same query and explored metadata and locate an external source besides Eurostat like UK Gov statistics and create a binding property based on metadata relevance matching. This all is acquired using textual pre-processing, keyword generation, fuzzy logic and dictionary fact validation in this research. The textual information extracted from the Eurostat metadata file is evaluated and pre-processed for identifying common labels, main column names, headers, sub sections, hyperlinks etc. each category is maintained with its available text. A scrum length of four consecutive words is used for constructing connection between words to identify how the word was placed or can be placed in a sentence.

Membership functions are the primary functions in any FLS system.

These membership functions are deployed in both fuzzification and defuzzification process. The aim of these functions is to map non fuzzy inputs to the corresponding fuzzy linguistic terms and same applies for defuzzification where the fuzzy linguistic terms are mapped to match values i.e. output values or terms. The important concept is the occurrence of a similar value over different sets. Fuzzy sets allow the values to appear in more than one correspondence thus increasing the relativity in the context of metadata connectivity. For instance, the metadata keywords are grouped and classified into classes. For set A the keywords generated could represent the following A:{economic, economic value, economic indicator, pay scale}. Each of these words has a relevant value in the membership function based on their degree of relevance to the label "Economic Indicators". Thus, the membership function display different forms based on the value of relevance such as triangle, trapezoid and gaussian shapes etc.

Fuzzy rule base is a knowledge base or a simple generic rule base that controls and monitors the output variable. The most simplest form of fuzzy rule is a general representation of IF-Then conditional statements with a conclusive variable statement. However, for the context of this research a more advanced rule base has been generated. The details of this rule base are beyond the scope of this research paper. However, the rules developed for understanding metadata are generated alongside standard fuzzy rules. The example of simple fuzzy rules are provided in Table1.

Table 1: Example of Fuzzy Rule Base.

S.no	Fuzzy Rules
1	IF(Label is 'Economic') AND (Header is "Statistical") THEN operation is Economic
2	IF(label is 'confidential') AND (Header is 'Paygrade') THEN operation is Salary
2	IF (Header is 'Forecast') AND (label is 'statistical') AND (Keyword is 'revised') THEN function is revised-data-stats

Fuzzy set operations are responsible for the fuzzy rule evaluations and projections. These are also responsible for combine fuzzy rules evaluations and performance. The first step is the evaluation of each rule. After the evaluation completion a compilation process is initiated. The rules can be combined in different groups thus allowing the metadata classes or upper categories to group under one. Some of the

most basic operations include MIN, MAX, UNION, COMPLEMENT, OR, AND.

Defuzzification is the last step in the FLS system. After a complete inference has been performed on textual metadata a value is returned. This is regarded as the fuzzy value. To attain a crisp output value the defuzzification process is utilized. In this step membership functions again play a primary role. Defuzzification has some basic standardised algorithms for calculating a deriving a crisp value. However, the details of each algorithm are beyond the scope and subject of this paper. The names of some most commonly deployed algorithms include centre of gravity, centre of gravity for singletons, left most maximum, right most maximum and so on.

A generic representation of the fuzzy algorithm is provided below:

```
   begin
       Definition {Linguistic
Variables, Terms}
   Initialize (membership Functions();)
   Construct Rule-Base ();
   Initialize (Rule-Base[][]);
   Fuzzification();
   //Use membership functions to
convert crisp data values to fuzzy
values.
   Evaluate (Rule-Base[][]);
   Combine[];
   //Combines results of each rule
   //Inference Module
   Defuzzification();
   //convert the attained output data
to non-fuzzy values.
   End.
```

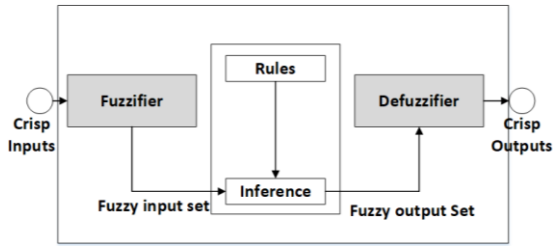

Figure 1: The overall representation of the functioning of a fuzzy system.

Figure 2 represents the designed framework for operational textual metadata utility with fuzzy logic. Figure 3 provides insight into the working and pre-processing of metadata before input sets are sent to Mamdani fuzzy controller or Naïve Bayes. The framework concludes currently in four possible outputs. Each output is distinct and serves a purpose for integration architecture and contribution. As explained earlier the fuzzy outputs are in the form of relevance or more precisely in the form of conformance degree. The ranges are between 0 to 1 and can have many values such as degree conformance of 01., 0.11.0.12 etc. The first of the four outputs is the graph based connected metadata representation. This is attained by analysing the primary source of metadata i.e. in our case the Eurostat file. The header of 'Related Metadata' is explored that contains links or connections to other metadata. This is structured in the form of graph for further connections and explorations. The second output is based on classification. This generated flags for information inside metadata that corresponds to domain categories such as in sciences and technology the fields of biology, electrical, robotics computer science all ca co-exist. The third output of framework is the main fuzzy conformance degree for each meta-class and metadata category. The metadata from primary and secondary sources is explored and a relation of degree is conformed. If the primary and secondary source have more in common the fuzzy out or crisp output set has higher degrees such as 0.8, 0.7, 0.6, 0.5 etc.

3 RELATED WORK

The research study and literature review associated with the research under discussion is broad spectrum. It does not confine to a specific data or domain classification. The idea of utilizing metadata in business intelligence and semantic web has been increasing over the past few years. The difficulty in this process is the availability of research and professional tools to both create and manage metadata. Business intelligence is a very broad field of reference. With the recent advancements in semantic web and big data the field of BI has become even more intriguing and powerful. The needs of business intelligence are extensive and include many challenges such as the computational complexity, analysis & reporting, output and performance accuracy, predictive analytic etc. These business applications are designed to provide solutions for above mentioned concerns. The primary concern of industrialists, professionals and researchers is the increase in data size and volume. It revolves around the data complexity such as retrieval of information, pre-processing, post processing, analysis etc.

Consequently, the semantic web is also increasing with increasing amounts of data connected to each other and available for use such as linked open data. The only problem is the web of metadata and metadata repositories are being set on second priority.

Very minimum applications and professional tools have been designed to work on metadata infrastructure. Google and other big tech giants use primary metadata for all data processing and their metadata architectures are improving every day. This is however not the case with every other industry and research institutes. In this respect semantic web and semantic search presents a very appreciative problem design.

The semantic (Laborie et al., 2015) ideology revolves around the concept of enriched and cultured metadata. The concept is to provide fully enriched metadata as a support layer in integration architectures. This study also provides and proposes the idea of using metadata connectivity and association in the form of relational graphs. The demand of sematic databases and semantic applications are also increasing thus, providing the perfect opportunity to embellish the metadata architectures and provide targeted solutions to illustrated and encountered problems. As mentioned earlier the tech giants like Google (Mnih et al., 2013) are producing research and professional products that incorporate the use of metadata and metadata repositories.

The research also provides insights into the deployment and use of new and enriched metadata types as compared to traditional schema and textual descriptions. Like fuzzy logic in our proposed framework research there are many other algorithms and machine learning solutions that have been devised recently to incorporate the effective learning process for metadata such as reinforcement learning modules. (Heess et al., 2012); (Watkins, 1989) indulge in using reinforcement learning as an experience generator. Similar techniques of learning can be deployed for the metadata learning modules. Correct predictions, associations can transform into learning experiences and metadata profiling can proficiently improve over time. The Atari (Mnih et al., 2013) was considered and is still to date one of the biggest scientific accomplishments and it was attained through deep learning and reinforcement learning.

The researchers are posing research questions on pre-processed metadata for encouraging reinforcement strategies and rule based systems. The metadata studies and association with data (Sutton and Barto, 1998); (Watkins, 1989) indicate that one form of metadata can conform to more than one domain of data classification. Thus, a replication metadata technique can be easily deployed to empower the datasets with communication architectures. The learning can be enhanced using experiences this is regarded as reward (Bellemare et al., 2015) based learning. The metadata can be fed into a learning systems and can learn from integration results such as rewards for correct classification, integration, identification of conflict etc. Similarly, consequences can eb delivered to serve as memory for metadata learning lakes.

The meta-data needs to be increased, enriched and put to better use if more aware applications are required. Also meta-data has the power to initiate learning from different machine learning algorithms in order to attain predictive accuracy. The final output is the metadata that could not be understood at the moment. This metadata could represent important information and thus a learning curve has to be designed for this.

4 CONCLUSIONS

In this paper we proposed a new concept in the field of metadata for data integration. Fuzzy technique allows the metadata distribution to fall under closely related possibilities. It improves the context and association of metadata with each other. This is promising for resolving the problem of data integration over traditional databases, semantic web databases and big data. The idea to pre-match is a necessity with the increasing amounts of information over the web and in general. The research notions that metadata when properly cultured can produce highly optimized and efficient results. The biggest advantage is the use of semantic orientation and representation. The connection between two data objects can be understood more efficiently if powerful metadata backs up the system. Also, metadata alone cannot script relations thus various algorithm such as Naïve Bayes for textual metadata render far better results than any other. However, Fuzzy logic not only improves the textual metadata predictions, associations, classifications etc. It also inhibits the feature of mapping conformance degrees i.e. how one part of metadata might or eventually correlates with another metadata set of values in the similar dataset.

The next phase of the study is to profile information using the existing data for generation of metadata. This technique is referred to as “Data Profiling”. Data profiling allows the machine learning algorithms to analyse datasets and derive descriptive and analytical statistics. The future work is concentrated in developing the fuzzy logic framework and adding a module of “Incomplete Metadata” for parts of data sources that do not have sufficient metadata attached.

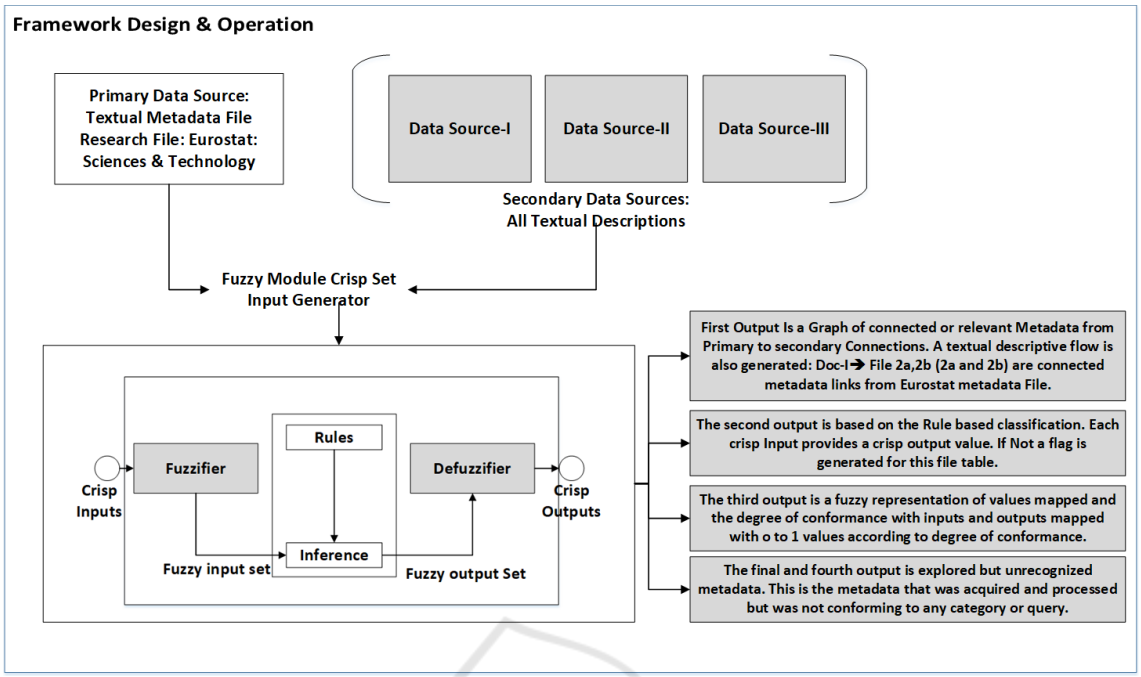

Figure 2: The Framework Design, Operation and perceived system outputs are described in this picture. It represents the functionality of the designed and attributed system for analysing textual metadata.

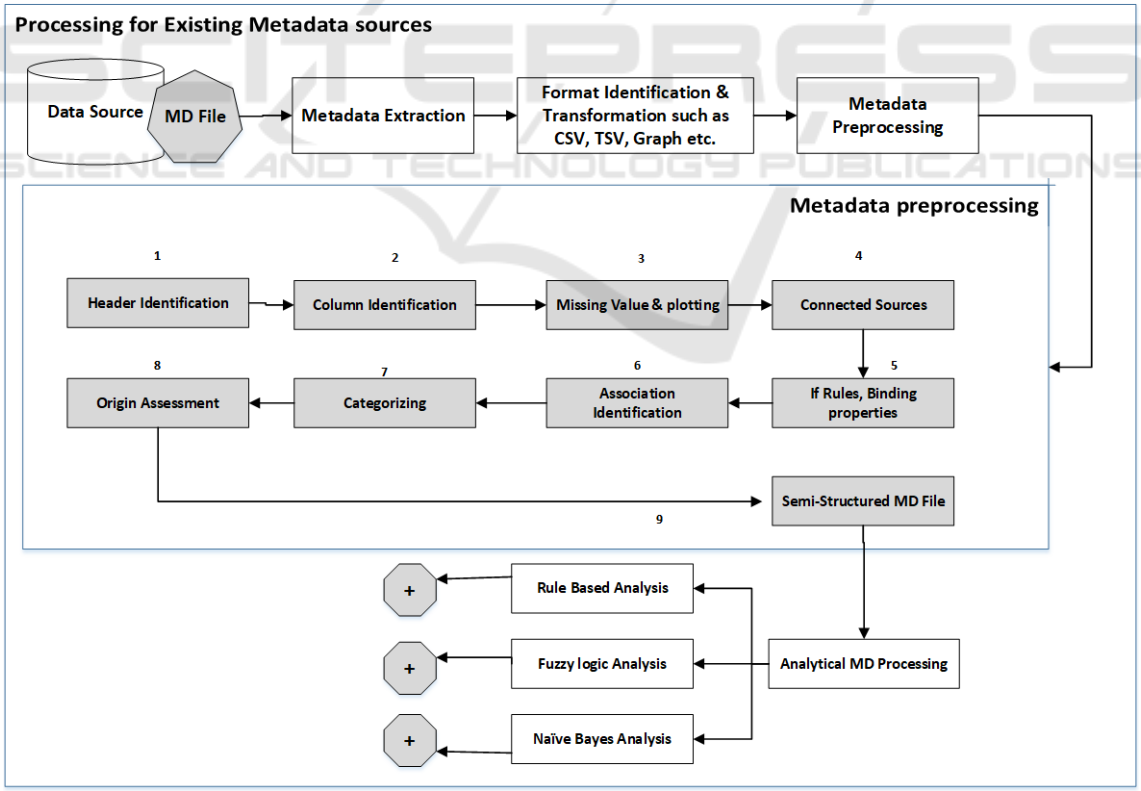

Figure 3: The illustration of metadata pre-processing and ingestion of metadata processed files into Naïve Bayes and Fuzzy Controllers.

ACKNOWLEDGEMENTS

This research has been funded by the European Commission through the Erasmus Mundus Joint Doctorate Information Technologies for Business Intelligence-Doctoral College (IT4BI-DC).

REFERENCES

Amorim, R.C., Castro, J.A., da Silva, J.R., Ribeiro, C., 2014. *In: Proc. of the 8th Re- search Conference on Metadata and Semantics Research, MTSR 2014*. Communications in Computer and Information Science, vol. 478, pp. 193

Auer, S., Demter, J., Martin, M., Lehmann, J.:., 2012. An extensible framework for high-performance dataset analytics. *In: Proc. of the 18th International Conference on Knowledge Engineering and Knowledge Management, EKAW 2012*. Lecture Notes in Computer Science, vol. 7603, pp. 353{362. Springer (2012)

Bellemare, M.G., Naddaf, Y., Veness, J., Bowling, M., 2015.The arcade learning environment: An evaluation platform for general agents (extended abstract). *In: Proc. of the 24th International Joint Conference on Arti_cial Intelligence, IJCAI 2015*. pp. 4148{4152. AAAI Press (2015)

Bellemare, M.G., Veness, J., Bowling, M., 2012. Sketch-based linear value function approximation. *In: Proc. of the 26th Annual Conference on Neural Information Processing Systems*. pp. 2222{2230 (2012).

Halevy, A.Y., Korn, F., Noy, N.F., Olston, C., Polyzotis, N., Roy, S., Whang S.E.., 2016. Goods: Organizing Google's datasets. *In: Proc. of the 2016 International Conference on Management of Data, SIGMOD Conference 2016*. pp. 795{806. ACM..

Heess, N., Silver, D., Teh, Y.W., 2012. Actor-critic reinforcement learning with energy based policies. *In: Proc. of the 10th European Workshop on Reinforcement Learning, EWRL 2012*. JMLR Proc., vol. 24.

L.C.B., 1995. Morgan Kaufmann, Residual algorithms: Reinforcement learning with function approximation. *In: Proc. of the 12th International Conference on Machine Learning*. pp.30

Laborie, S., Ravat, F., Song, J., Teste, O., 2015.*In: Actes du XXXIII_eme Congr_es IN-FORSID*. pp. 99.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.A., 2013.*Playing Atari with deep reinforcement learning. CoRR* abs/1312.5602.

Pandey, P., Pandey, D., Kumar, S., 2010. *Reinforcement learning by comparing immediate reward. CoRR* abs/1009.2566 (2010)

Sutton, R.S., Barto, A.G., 1998. *MIT Press*, Reinforcement learning - an introduction. *Adaptive computation and machine learning, MIT Press.*

Watkins, C., 1989.Learning from Delayed Rewards. Ph.D. thesis, Cambridge University.