A Survey on Data Extraction and Data Duplication Detection

Yashika A. Shah
Amravati, India
e-mail: yashah0694@gmail.com

Snehal S. Zade
Amravati, India
e-mail: snehalzade12@gmail.com

Smita M. Raut
Amravati, India
e-mail: swatiraut93@gmail.com

Shraddha P. Shirbhate
Amravati, India
e-mail:
shraddhashirbhate30@gmail.com

Vijeta U. Khadse
Amravati, India
e-mailvijetakhadse20@gmail.com

Anup P. Date
Amravati, India
e-mail:dateap@gmail.com

***Abstract*—** Text mining, also known as Intelligent Text Analysis is an important research area. It is very difficult to focus on the most appropriate information due to the high dimensionality of data. Feature Extraction is one of the important techniques in data reduction to discover the most important features. Processing massive amount of data stored in a unstructured form is a challenging task. Several pre-processing methods and algorithms are needed to extract useful features from huge amount of data. Dealing with collection of text documents, it is also very important to filter out duplicate data. Once duplicates are deleted, it is recommended to replace the removed duplicates. This Paper review the literature on duplicate detection and data fusion (remov e and replace duplicates).The survey provides existing text mining techniques to extract relevant features, detect duplicates and to replace the duplicate data to get fine grained knowledge to the user.

Keywords- *Text feature extraction, text mining, query search, text classification.*

*****

I. INTRODUCTION

. Society is increasingly becoming more digitized and as a result organizations are producing and storing vast amount of data. Managing and gaining insights from the produced data is a challenge and key to competitive advantage. Web based social applications like people connecting websites results in huge amount of unstructured text data. These huge data contains a lot of useful information. People hardly bother about the correctness of grammar while forming a sentence that may lead to lexical syntactical and semantic ambiguities. The ability of finding patterns from unstructured form of text data is a difficult task.

Data mining aims to discover previously unknown interrelations among apparently unrelated attributes of data sets by applying methods from several areas including machine learning, database systems, and statistics. Many researches have emphasized on different branches of data mining such as opinion mining, web mining, text mining. Text mining is one of the most important strategy involved in the phenomenon of knowledge discovery. It is a technique of selecting previously unknown, hidden, understandable, interesting knowledge or patterns which are not structured. The prime objective of text mining is to diminish the effort made by the users to obtain appropriate information from the collection of text sources [1].

Thus, our focus is on methods that extract useful patterns from texts in order to categorize or structure text collections. Generally, around 80 percent of company's information is saved in text documents. Hence text mining has a higher economic value than data mining. Current research in the area of text mining tackles problems of text representation, classification, clustering, information extraction or the search for and modelling of hidden patterns. Selection of characteristics, influence of domain knowledge and domain- specific procedures play an important role. The text documents contain large scale terms, patterns and duplicate lists. Queries submitted by the user on web search are usually listed on top retrieved documents. Finding the best query facet and how to effectively use large scale patterns remains a hard problem in text mining.

However, the traditional feature selection methods are not effective for selecting text features for solving relevance issue. These issues suggests that we need an efficient and effective methods to mine fine grained knowledge from the huge amount of text documents and helps the user to get information quickly about a user query without browsing tens of pages. The paper provides a review of an innovative techniques for extracting and classifying terms and patterns. A user query is usually presented in list styles and repeated many times among top retrieved documents. To aggregate frequent lists within the top search results, various navigational techniques have been presented to mine query facets.

Text Mining Models:

Text mining tasks consists of three steps: text preprocessing, text mining operations, text post processing. Text preprocessing includes data selection, text categorization and feature extraction. Text mining operations are the core part of text mining that includes association rule discovery, text clustering and pattern discovery as shown in Figure1. Post processing tasks modifies the data after text mining operations are completed such as selecting, evaluating and visualization of knowledge. It consists of two components text filtering and knowledge cleansing. Many approaches [2] have been concerned of obtaining structured datasets called intermediate forms, on which techniques of data mining [3] are executed. Text filtering translates collection of text documents into selected intermediate form(IF) which means Knowledge cleansing or discovering patterns. It can be structured or semistructured. Text mining methods like clustering, cla- ssification and feature extraction falls within document based IF. Pattern discovery and relationship of the obje- ct, associative discovery, visualization fall within object based documents.

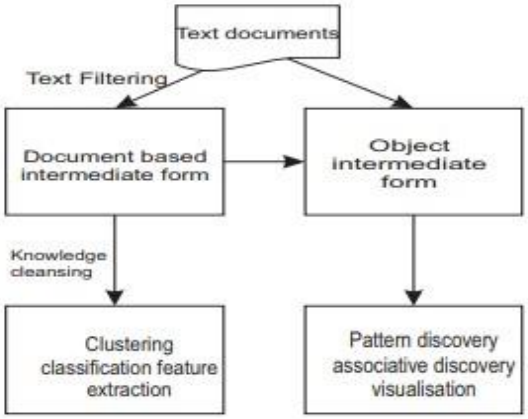

Figure 1: Text mining framework

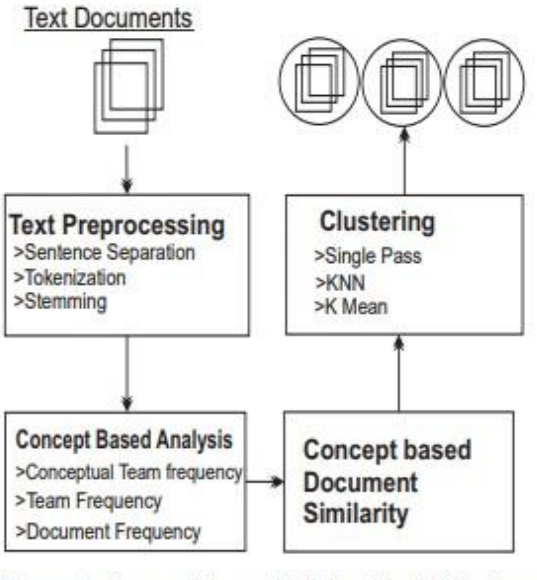

Figure 2: Concept based Mining Model System

When documents contains terms with same frequency. Two terms can be meaningful while the other term may be irrelevant. Inorder to discover the semantic of text, the mining model is introduced. Figure 2 represents a new mining model based on concepts. The model is proposed to analyse terms in a sentence from documents. The model contains group of concept analysis, they are sentence based concept analysis, document based concept analysis and corpus based similarity measure [4]. Similarity measure concept based analysis calculates the similarity between documents. The model effectively and efficiently finds matching concepts between documents, according to the meaning of their sentence.

II. FEATURE EXTRACTION

a) Feature Mining for Text Mining:

Li et al.,[5] designed a new technique to discover patterns i.e., positive and negative in text document. Both relevant and irrelevant document contains useful features. Inorder to remove the noise, negative documents in the training set is used to improve the effectiveness of Pattern Taxonomy Model PTM. Two algorithms HLF mining and N revision was introduced. In HLF mining, it first finds positive

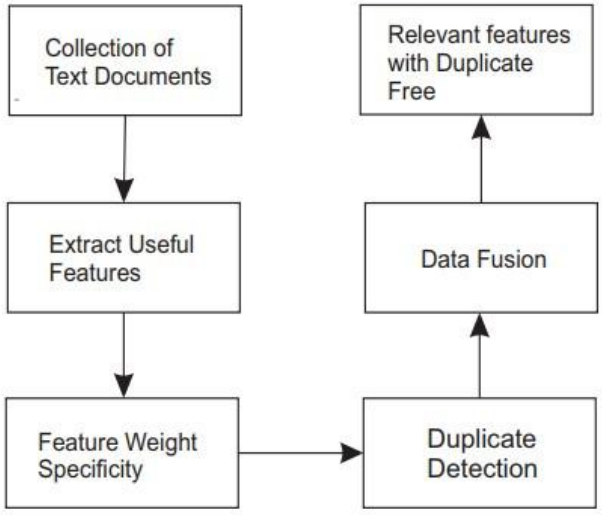

Fig. 3 HLF Mining

features, discovers negative features and then composes the set of term. The offenders are selected by ranking the negative documents. The weights are initialized to the discovered terms of negative patterns. NRevision algorithms explains the terms weight based on their specificity and distribution in both positive and negative patterns Zhong et al., [6] has presented an effective pattern discovery technique which includes the process of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. The proposed model outperforms other pure data mining-based methods, the concept based models and term-based state-of the- art models, such as BM25 and SVM. Li et al., [7] proposed two algorithms namely Fclustering and Wfeature to discover both positive and negative patterns in the text documents. The algorithm Fclustering classifies

the terms into three categories general, positive, negative automatically without using parameters manually. After classifying the terms using Fclustering, Wfeature is executed to calculate the weights of the term. Wfeature is effective because the selected terms size is less than the average size of the documents. The proposed model is evaluated on RCV, Trec topics and Reuters 21578 dataset, the model performs much better than the term based method and pattern based method. Xu et al., [8] experimented on microblog dimensionality reduction- A deep learning approach. The approach aims at extracting useful information from large amount of textual data produced by microblogging services. The approach involves mapping of natural language texts into proper numerical representations which is a challenging issue. Two types of approaches namely modifying training data and modifying training objective of deep networks are presented to use microblog specific information. Meta-information contained in tweets like embedded hyperlinks is not explored. Nguyen et al., [9] worked on review selection using Micro-reviews. The approach consists of two steps namely matching review sentences with micro reviews and selecting a few reviews which cover many reviews. A heuristic algorithm performs computionally fast and provides informative reviews.

b) Feature Extraction for Classification: Khadhim et al., [10] [11] developed two weighting methods TF -IDF and TF- IDF (Term Frequency/Inverse Document Frequency) global to reduce dimensionality of datasets because it is very difficult to process the original features i.e, thousands of features.

PCA and Random Projection RP: Principal Component Analysis (PCA) is a simple technique used to explore and visualize the data easily. PCA extracts useful information from complicated data sets using non parametric method. It determines a lower dimension space by statistical method. Based on eigen value decomposition of the covariance matrix transformation matrix of PCA is calculated and thereby computation cost is more and it is also not suitable for very high dimensional data. The strength of PCA is that there are no parameters to fine tune and also no co-efficient is required to adjust. Fradkin et al., [12] [13] reported a number of experiments by evaluating random projecton in supervised learning. Different datasets were tested to compare random projection and PCA using several machine learning methods. The results show PCA outperforms RP in efficiency for supervised learning. The results also shows that RP's are well suited to use with nearest neighbour and with SVM classifier and are less satisfactory with decision trees.

III. DUPLICATE DETECTION AND DATA FUSION

Duplicate detection is the methodology of identification of multiple semantic representation of the existing and similar real world entities. The present day detection methods need to execute larger datasets in the least amount of time and hence to maintain the overall quality of datasets is tougher. Papenbrock et al., [17] proposed a strategic approach namely the progressive duplicate detection methods which finds the duplicates efficiently and reduces the overall processing time by reporting most of the results than the existing classical approaches. Bano et al., [18] executed innovative windows algorithm that adapts window for duplicates and also which are not duplicates and unnecessary comparisons is avoided. The duplicate records are a vital problem and a concern in knowledge management [19]. To Extract duplicate data items an entity resolution mechanism is employed for the procedure of cleanup. The overall evaluation reveals that the clustering algorithms perform extraordinarily well with accuracy and f- measure being high. Whang et al., [20] investigates the enhancement of focusing on several matching records. Three types of hints that are compatible with different ER algorithms:(i) an ordered list of records, (ii) a sorted list of record pairs, (iii) a hierarchy of record partitions. The underlying disadvantage of the process is that it is useful only for database contents.

Duplicate records do not share a strategic key but they build duplicate matching making it a tedious task. Errors are induced because the results of transcription errors, incomplete information and lack of normal formats. Abraham et al., [21] [22] provides survey on different techniques used for detecting duplicates in both XML and relational data. It uses elimination rule to detect duplicates in database. Elmagarmid et al., [23] present intensive analysis of the literature on duplicate record for detection and covers various similarity metrics, which will detect some duplicate records in exceedingly available information. The strengths of the survey analysis in statistics and machine learning aims to develop a lot of refined matching techniques that deem probabilistic models. Deduplication is an important issue in the era of huge database [24]. Various indexing techniques have been developed to reduce the number of record pairs to be compared in the matching process. The total candidates generated by these techni- ques have high efficiency with scalability and have been evaluated using various data sets. The training data in the form of true matches and true non matches is often unavailable in various real-world applications. It is commonly up to domain and linkage experts for decision of the blocking keys. Papadakis et al., [25] presented a blocking methods for clean- clean ER over Highly Heterogeneous Information Spaces (HHIS) through an innovative framework which comprises of two orthogonal layers. The effective layer incorporates methods for construction of several blockings with small probability of hits; the efficiency layer comprises of a rich variety of techniques which restricts the required number of pairwise matches. Papadakis et al., [26] focuses to boost the overallblocking efficiency of the quadratic task on Entity Resolution among large, noisy, and heterogeneous information areas. The problem of merging many large databases is often encountered in KDD. It is usually referred to as the Merge/Purge problem and is difficult to resolve in scale and accuracy. The Record linkage [27] is a wellknown data integration strategy that uses sets for merging, matching and elimination of duplicate

records in large and heterogeneous databases. The suffix grouping methodology facilitates the causal ordering used by the indexes for merging blocks with least marginal extra cost resulting in high accuracy. An efficient grouping similar suffixes is carried out with incorporation of a sliding window technique. The method is helpful in various health records for understanding patient's details but is not very efficient as it concentrates only on blocking and not on windowing technique. Additionally the methodology with duplicates that are detected using the state of the art expandable paradigm is approximate [28]. It is quite helpful in creating clusters records.

Bronselaer et al., [29] focused on Information aggregation approach which combine information and rules available from independent sources into summarization. Information aggregation is investigated in the context of inferencing objects from several entity relations. The complex objects are composed of merge functions for atomic and subatomic objects in a way that the composite function inherits the properties of the merge functions.

Sorted Neighborhood Method (SNM) proposed by Draisbach et al., [30] partitions data set and comparison are performed on

the jurisdiction of each partition. Further, the advances in a window over the data is done by comparison of the records that appears within the range of same window. Duplicate Count Strategy (DCS) which is a variation of SNM is proposed by regulating the window size. DCS++ is proposed which is much better than the original SNM in terms of efficiency but the disadvantage is that the window size is fixed and is expensive for selection and operation. Some duplicates might be missed when large window are used. The tuples in the relational structure of the database give an overview of the similar real world entities such tuples are described as duplicates. Deleting these duplicates and in turn facilitating their replacement with several other tuples represents the joint informational structure of the duplicate tuples up to a maximum level. The incorporated delete and then replacement mode of operation is termed as fusion. The removal of the original duplicate tuples can deviate from the referential integrity. Bronselaer et al., [31] describes a technique to maintain the referential integrity. The fusion Propogation algorithm is based on first and second order fusion derivatives to resolve conflicts and clashes. Traditional referential integrity strategies like DELETE cascading, are highly sophisticated. Execution time and recursively calling the propagation algorithm increases when the length of chain linked relations increases. Bleiholder et al., proposes the SQL Fuse by inducing the schema and semantics. The existential approach is towards the architecture, query languages, and query execution. The final step of actually aggregating data from multiple heterogeneous sources into a consistent and homogeneous datasetand is often inconsiderable. Naumann et al., [32] observes that amount of noisy data are in abundance from several data sources. Without any suitable techniques for integrating and fusing noisy data with deviations, the quality of data associated with an integrated system remains extremely low. It is necessary for allowing tentative and declarative integration of noisy and scattered data by incorporating schema matching, duplicate detection and fusion. Subjected to SQL-like query against a series of tables instance, oriented schema matching covers the cognitive bridge of the varied tables by alignment of various corresponding attributes. Further, a duplicate detection technique is used for multiple representations of several matching entities. Finally, the paradigm of data fusion for resolving a conflict in turn merges around with each individualistic duplicate transforming it into a unique singular representation Bleiholder et al., [33] explains a conceptual understanding of classification of different operators over data fusion. Numerous techniques are based on standard and advanced operators of algebraic relations and SQL. The concept of Co-clustering is explained from several techniques for tapping the rich and associated meta tag information of various multimedia web documents that includes annotations, descriptions and associations. Varied Coclustering mechanisms are proposed for linked data that are obtained from multiple sources which do not matter the representational problem of precise texts but rather increase their performance up to the most minimally empirical measurement of the multi- modal features. The two channel Heterogeneous Fusion ART (HF-ART) yields several multiple channels divergently. The GHF-ART [34] is designed to effectively represent multimedia content that incorporates Meta data to handle precise and noisy texts. It is not trained directly using the text features but can be identified as a key tag by training it with the probabilistic distribution of the tag based occurrences. The approach also incorporates a highly and the most adaptive methodology for active and efficient fusion of multimodal.

IV. CONCLUSION

This Paper reviews different techniques and framework to extract relevant features from huge amount of unstructured text documents. To guarantee the quality of extracted relevant features in a collection of text documents is a great challenge. Many text mining techniques have been proposed till date. However how effectively the discovered features are interesting and useful to the user is an open issue.

REFERENCES

[1]. R. Agrawal and M. Batra, "A Detailed Study on Text Mining Techniques," International Journal of Soft Computing and Engineering (IJSCE) ISSN, vol. 2, no. 6, pp. 2231–2307, 2013.

[2]. V. H. Bhat, P. G. Rao, R. Abhilash, P. D. Shenoy, K. R. Venugopal, and L. Patnaik, "A Data Mining Approach for Data Generation and Analysis for Digital Forensic Application," International Journal of Engineering and Technology, vol. 2, no. 3, pp. 313– 319, 2010.

[3]. Y. Zhang, M. Chen, and L. Liu, "A Review on Text Mining," In Proceedings of 6th IEEE International

Conference on Software Engineering and Service Science (ICSESS), pp. 681– 685, 2015.

[4]. S. Shehata, F. Karray, and M. S. Kamel, "An Efficient Concept-based Mining Model for Enhancing Text Clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1360–1371, 2010.

[5]. Y. Li, A. Algarni, and N. Zhong, "Mining Positive andNegative Patterns for Relevance Feature Discovery," In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 753–762, 2010.

[6]. N. Zhong, Y. Li, and S.-T. Wu, "Effective Pattern Discovery for Text Mining," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 1, pp. 30–44, 2012

[7]. Y. Li, A. Algarni, M. Albathan, Y. Shen, and M. A. Bijaksana, "Relevance Feature Discovery for Text Mining," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 6, pp. 1656–1669, 2015.

[8]. L. Xu, C. Jiang, Y. Ren, and H.-H. Chen, "Microblog Dimensionality Reduction—A Deep Learning Approach," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 7, pp. 1779–1789, 2016.

[9]. T.-S. Nguyen, H. W. Lauw, and P. Tsaparas, "Review Selection Using Micro-Reviews," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 4, pp. 1098– 1111, 2015.

[10]. A. I. Kadhim, Y. Cheah, N. H. Ahamed, L. A. Salman et al., "Feature Extraction for Co-occurrence-based Cosine Similarity Score of Text Documents," In Proceedings of IEEE Student Conference on Research and Development (SCOReD), pp. 1–4, 2014.

[11]. K. Srinivasa, A. Singh, A. Thomas, K. R. Venugopal, and L. Patnaik, "Generic Feature Extraction for Classification using Fuzzy C-means Clustering," In Proceedings of 3rd International Conference on Intelligent Sensing and Information Processing, pp. 33–38, 2005.

[12]. D. Fradkin and D. Madigan, "Experiments with Random Projections for Machine Learning," In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 517–522, 2003.

[13]. S. Joshi, D. Shenoy, P. rashmi, K. R. Venugopal, and L. Patnaik, "Classification of Alzheimer's Disease and Parkinson's Disease by using Machine Learning and Neural Network Methods," In Proceedings of Second International Conference on Machine Learning and Computing (ICMLC), pp. 218–222, 2010.

[14]. Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, "Mining Temporal Patterns in Time Interval-Based Data," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 12, pp. 3318– 3331, 2015.

[15]. A. Bartoli, A. De Lorenzo, E. Medvet, and F. Tarlao, "Inference of Regular Expressions for Text Extraction from Examples," IEEE Transactions on Knowledge © 2016 Global Journals Inc. (US) Global Journal of Computer Science and Technology Volume XVI Issue V Version I 17Year 2016 ( ) C and Data Engineering, vol. 28, no. 5, pp. 1217– 1230, 2016.

[16]. Q. Song, J. Ni, and G. Wang, "A Fast Clusteringbased Feature Subset Selection Algorithm for HighDimensional Data," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 1, pp. 1–14, 2013.

[17]. T. Papenbrock, A. Heise, and F. Naumann, "Progressive Duplicate Detection," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 5, pp. 1316–1329, 2015.

[18]. H. Bano and F. Azam, "Innovative Windows for Duplicate Detection," International Journal of Software Engineering and Its Applications, vol. 9, no. 1, pp. 95–104, 2015.

[19]. O. Hassanzadeh, F. Chiang, H. C. Lee, and R. J. Miller, "Framework for Evaluating Clustering Algorithms in Duplicate Detection," In Proceedings of the VLDB Endowment, vol. 2, no. 1, pp. 1282–1293, 2009.

[20]. S. E. Whang, D. Marmaros, and H. Garcia-Molina, "Pay- asyou- go Entity Resolution," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 5, pp. 1111– 1124, 2013.

[21]. A. A. Abraham and S. D. Kanmani, "A Survey on Various Methods used for Detecting Duplicates in Xml Data," International Journal of Engineering Research and Technology, vol. 3, no. 1, pp. 1–10, 2014.

[22]. J. J. Tamilselvi and C. B. Gifta, "Handling Duplicate Data in Data Warehouse for Data Mining," International Journal of Computer Applications (0975–8887), vol. 15, no. 4, pp. 1–9, 2011.

[23]. A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate Record Detection: A Survey," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, pp. 1–16, 2007.

[24]. P. Christen, "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 9, pp. 1537– 1555, 2012.

[25]. G. Papadakis, E. Ioannou, T. Palpanas, C. Nieder´ee, and W. Nejdl, "A Blocking Framework for Entity Resolution in Highly Heterogeneous Information Spaces," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 12, pp. 2665–2682, 2013.

[26]. G. Papadakis and W. Nejdl, "Efficient Entity Resolution Methods for Heterogeneous Information Spaces," In Proceedings of IEEE 27th International Conference on Data Engineering Workshops (ICDEW), pp. 304–307, 2011.

[27]. T. De Vries, H. Ke, S. Chawla, and P. Christen, "Robust Record Linkage Blocking using Suffix Arrays and Bloom Filters," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 5, no. 2, pp. 9–44, 2011.

[28]. O. Hassanzadeh and R. J. Miller, "Creating Probabilistic Databases from Duplicated Data," The VLDB Journal—The International Journal on Very Large Data Bases, vol. 18, no. 5, pp. 1141–1166, 2009.

[29]. A. Bronselaer and G. De Tr´e, "Aspects of object Merging," Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS), pp. 1–6, 2010.

[30]. U. Draisbach, F. Naumann, S. Szott, and O. Wonneberg, "Adaptive Windows for Duplicate Detection," In Proceedings of IEEE 28th International Conference on Data Engineering (ICDE), pp. 1073–1083, 2012.

[31]. A. Bronselaer, D. Van Britsom, and G. De Tre, "Propagation of Data Fusion," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 5, pp. 1330– 1342, 2015.

[32]. F. Naumann, A. Bilke, J. Bleiholder, and M. Weis, "Data Fusion in Three Steps: Resolving Schema, Tuple, and Value Inconsistencies." IEEE Data Xml Data," International Journal of Engineering Research and Technology, vol. 3, no. 1, pp. 1– 10, 2014. Engineering and Management, vol. 29, no. 2, pp. 21–31, 2006.

[33]. J. Bleiholder and F. Naumann, "Data Fusion," ACM Computing Surveys (CSUR), vol. 41, no. 1, pp. 1– 41, 2009.

[34]. L. Meng, A.-H. Tan, and D. Xu, "Semi-Supervised Heterogeneous Fusion for Multimedia Data CoClustering," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 9, pp. 2293–2306, 2014.