FUZZY MATCHING: AN ALTERNATIVE TECHNIQUE FOR MERGING EXTRACTED WEB DATA

L. Q. Zian[1], N.Z. Zulkarnain [1] and Y.J. Kumar [1]

[1]Fakulti Teknologi Maklumat dan Komunikasi,
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian
Tunggal, Melaka, Malaysia.

Corresponding Author's Email: [1]alexlee7399@gmail.com

Article History: Received 28 March 2024; Revised 8 April 2024;
Accepted 11 May 2024

ABSTRACT: Web scrapping has been a popular method for collecting data from websites. This is because data on the internet is updated frequently thus making it a good source for getting accurate information. However, the non-homogeneous nature of each website may cause the data from the different internet web sources to have different data making the quality of the data inconsistent. Previous study has proposed the use of record linkage method to merge data from multiple websites. The record linkage method proposed by previous study used deterministic technique to match data which match the string of matching variable to merge data. However, deterministic technique requires the matching variable to be an exact match to be able to match. This study explores the use of fuzzy matching technique as an alternative technique. A comparison in this study found out that fuzzy matching has a slightly better performance in merging web data. However, the main drawback of fuzzy matching is that it is hard to determine the threshold to trigger a match. Therefore, the future work should focus on exploring an optimal method on determining the threshold for fuzzy matching to making the process more streamlined.

KEYWORDS: *Web scrapping, Data Merging, Deterministic Matching, Fuzzy Matching*

1.0 INTRODUCTION

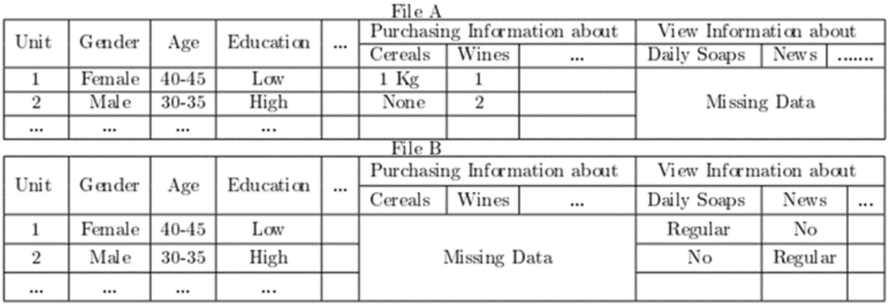
The internet is a huge repository for information. As we progress through the era of Internet of Things, the information on the internet has grown enormously. For most users, the internet nowadays is displayed in a user-friendly interface which is easy to understand. The potential of the internet is boundless due to the useful raw information in it. However, obtaining information can be tedious for a normal human due to the sheer size of the internet a person needs to navigate. Web scrapping is a technique that used to scrap web data from the internet which is introduce in 1994. It can scrap any form of data such as image, text, documents, video, or sound. Web scraper can extract web data with relatively less time and low cost compared to manual extraction.

Web scraping is widely popular in the field of market research and business intelligence. The ability of web scraper to collect large volume of data is very useful for company to collect information that are beneficial for their business. The most prominent application of web scraping is the usage of web scraper in collecting online market price for market research and other different application that require mass data collecting [1] [2]. Despite the advantages of web scraping, studies has stated that this process are challenging to manage due to the non-homogeneous nature of each website[3]. These inconsistencies in data can cause the quality of data to degrade [4]. These implications can be solved by using data sources from different web sources as supplement using data fusion which also known as deterministic matching technique. However, deterministic matching technique requires the matching variable of to be exact same for matching to occur. This can be a detriment for merging data from web sources as web data are often not organized. Despite normalizing the data can resolve these problems, web data have different data representation or data structure in different website. This can make normalization task a difficult process [5]. Therefore this will explore an alternative technique to deterministic matching knowns as fuzzy matching to be used in web scraping.

2.0 RELATED WORKS

Data fusion also plays an important role in data science and can be used to combine data from different data sources. A data fusion study conducted in 2004 [6] propose the use of statistical matching technique which match both dataset with their common variable as shown in Figure 1. This approach can be compared to k-nearest neighbour with donor as the training set and recipient as a test set. Some element from the recipient set of k best matching donor will be selected and the matching distance will be calculated through standard measure such as euclidean distance. Rässler also mentioned that some variable between donor and recipient must be 100% match so it will become a critical variable. Critical variable is important in this statistical matching as it can prevent some misclassification. For example, the variable of men will not bet matched women to prevent women characteristic such as 'pregnant' to be predicted. This technique has been proven useful in a non-parametric statistical dataset [7]. This study uses Micro Statistical Matching (MiSM) to perform data integration between primary and secondary data of agriculture data.

Apart from combining data, data fusion process can also be used handle missing data. These can achieved by merging information from multiple data source. In a study merge the data is merged from multiple sources into a single database[8]. The data fusion is conducted via imputation with implicit model such as nearest neighbour. The main idea of the technique is to give the variable of the receiver from the whole vector of donor as show in the imputation scheme. The imputation scheme works optimally on numerical data by using various method such as nearest neighbour to handle the missing data. However, the imputation method cannot work if the missing data is a categorical data. Lewaa, et al. mentioned that Statistical Matching (SM) technique is effective at completing data file from different source that does not contain the same unit as shown in Figure 2 [9]. Lewaa, et al. also stated that SM can be an interesting method for market researchers to combine data as this can save a lot of time while providing richer data source for decision making.



File A

Unit	Gender	Age	Education	...	Purchasing Information about			View Information about		
					Cereals	Wines	...	Daily Soaps	News	....
1	Female	40-45	Low		1 Kg	1		Missing Data		
2	Male	30-35	High		None	2				
...	...	...	...							

File B

Unit	Gender	Age	Education	...	Purchasing Information about			View Information about		
					Cereals	Wines	...	Daily Soaps	News	...
1	Female	40-45	Low		Missing Data			Regular	No	
2	Male	30-35	High					No	Regular	
...	...	...	...							

Figure 1: Statistical matching in combining data from different dataset [10]

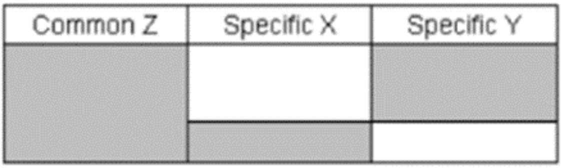


Figure 2: Matching Variables [8]

To combine the data extracted from web sources, previous studies had study the approaches of using web scraping with record linkage to merge data from multiple sources. Record linkage is a statistical technique use in the domain of database to match record using the common variable such as name, data and etc [10]. Record linkage match record by using deterministic matching which is done by comparing two strings for a match as shown in Figure 4. Its simplicity made it widely used in data related study for merging data into an integrated dataset.

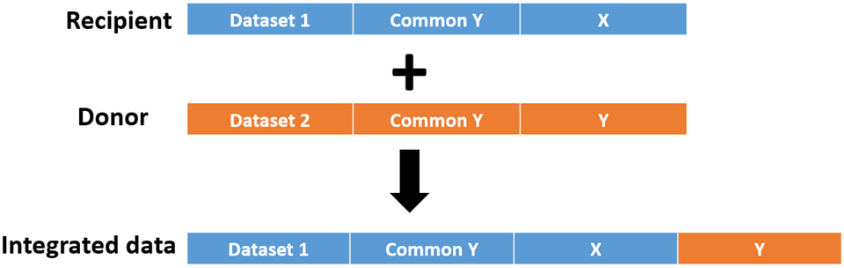
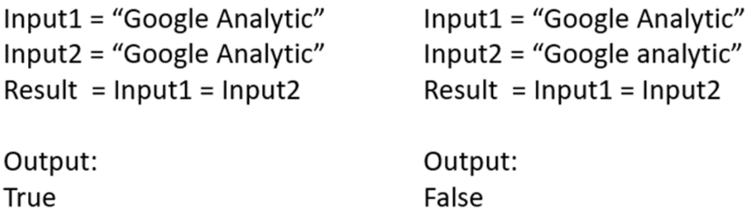


Figure 3: Record linkage method for merging data



```
Input1 = "Google Analytic"
Input2 = "Google Analytic"
Result  = Input1 = Input2

Output:
True
```

```
Input1 = "Google Analytic"
Input2 = "Google analytic"
Result  = Input1 = Input2

Output:
False
```

Figure 4: Process of deterministic matching

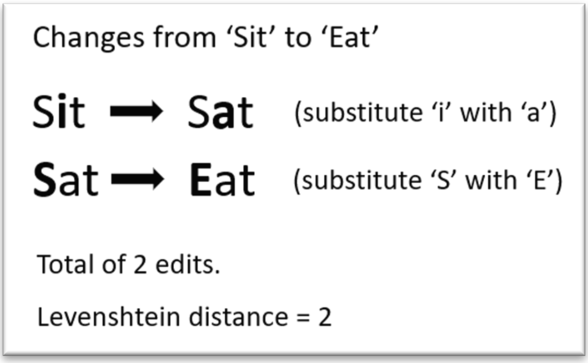
Record linkage is widely used in data related field like web scraping to match data from different data sources for merging data[10-15]. Record linkage was also used on creating synthetic data set to assist in data mining and data analysis. This approach was use by Namazi-Rad et al. in creating geo specific census and survey data using individual data such as wealth index and household data [12]. This approach is used create a reliable data for a smaller area as national statistical agencies cannot capture the data in a specific area while using a direct estimate bas on high level region. However there is one disadvantage of deterministic matching is that it does not account for spelling mistakes, missing words, spacing or letter cases. Therefore deterministic matching can be sensitive to typos or incomplete identifier as shown in Figure 5. This can be an implication on extracting data from internet as data extracted from the internet are not consistent[14].

As a conclusion, data merging can be done by using a matching variable in a data to match with other data using deterministic matching. However, deterministic matching require the matching variable to be an exact match for a matching to occur. Therefore, this study will explore an alternative technique to merge data.

2.1 Fuzzy Matching

Fuzzy matching is also known as fuzzy string matching is a data matching process that match data by using the approximate pattern of strings. The similarity of the string is measured in regard of the edit distance by using distance measuring metric such as Levenshtein distance as shown in equation 1 to measure the difference between two sequences of words. It uses the minimum number of edits use to change a word such as insertion, substitution, or deletion [16] as shown in Figure 5. This make fuzzy match can find match even if there is misspell in the words. In a practical application, a ratio function will be used to compute the similarity threshold. A predetermined threshold will be set depending on how similar a pair of strings will be matched[17].

$$lev_{a,b}\ (i,j) = \begin{cases} \max(i,j), if \min(i,j) = 0 \\ min \begin{cases} lev_{a,b}\ (i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1 \end{cases} ,otherwise \end{cases} \tag{1}$$



Changes from ‘Sit’ to ‘Eat’

Sit ➡ Sat (substitute ‘i’ with ‘a’)

Sat ➡ Eat (substitute ‘S’ with ‘E’)

Total of 2 edits.

Levenshtein distance = 2

Figure 5: Example of calculating edits between 2 strings

3.0 METHODOLOGY

This study will employ record linkage on merging data from two software review websites: 1) softwareadvice.com and 2) capterra.com. This study will use the alternative approach to impute the “Trialability” data from Capterra into the dataset from Software Advice. The dataset from Software Advice will be use as the recipient and “Software Name” will be use as the matching variable since the data need to be imputed to the correct software. This study will compare the performance of deterministic matching and fuzzy matching techniques in web scraping. This study will use the FuzzyWuzzy module (python module) to perform the fuzzy matching process by computing the Levenshtein distance similarity between a pair of strings to match them together. Figure 6 shows how fuzzy matching techniques is implemented in the web data extraction process.

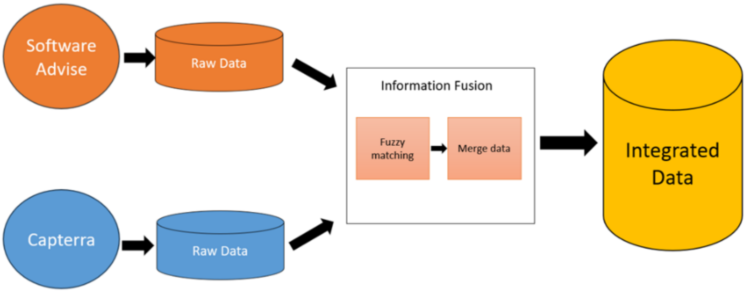


Figure 6: Data merging using fuzzy matching technique.

3.1 Determining the similarity threshold for fuzzy matching

Determining the best threshold for fuzzy matching is depended by case on cases basis. In this study, the matches we consider true positives in fuzzy matching are matches that has difference are letter cases and spacing. In this study, the threshold is initially set to 0.9. The output data of fuzzy matching will be compared to the output data of deterministic matching. In the 0.9 threshold configuration, the matching process merged additional 21 observations compared to deterministic matching. Due to the nature of how fuzzy matching matches data, not all the 21 observations are considered true positive. Therefore the false positive data will be removed manually based on the matching we considered as true positive. After removing all the false positives, there are only left 5 observations that are consider true positives shown in the Table. Thus, adjustment is made to the threshold configuration (+1 threshold in each test) until the desired result are met. Table 1 and Figure 7 shows that the threshold for the desire matching for the data use in this study is between 0.95 and 0.97. Therefore 0.95 will be use as the threshold for the fuzzy matching process.

Table 1: Additional data merged using fuzzy matching technique compared to deterministic matching technique.

Similarity Threshold	Additional Data merged compared to deterministic matching	True positive additional data
0.90	+21	+5
0.91	+6	+5
0.92	+6	+5

0.93	+6	+5
0.94	+6	+5
0.95	+5	+5
0.96	+5	+5
0.97	+5	+5
0.98	+4	+5
0.99	+4	+5
1.00	+0	0

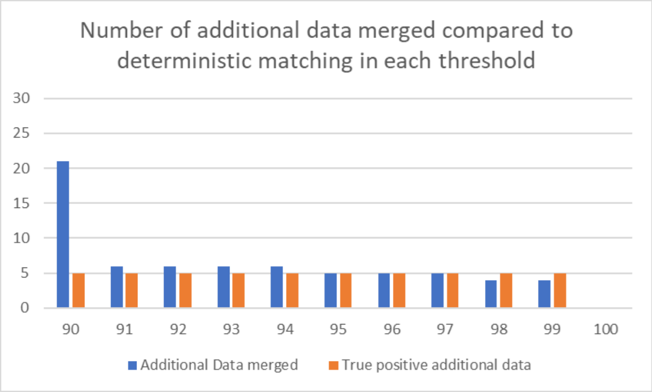


Figure 7: Graph of additional data merged using fuzzy matching technique compared to deterministic matching technique.

3.2 Evaluating matching validity and comparing data merging performance

Previous study has proved that performance of merging data is hard to evaluate [18]. A study in 2020 stated that the main aim of data merging is to create a synthetic dataset for statistical inference. Thus, it is crucial for the donor data to become the subset of the recipient data [19]. This study will use overlap index in equation 2 as measure the overlap between two data sets by dividing the size of intersection by the smaller size of the two data sets:

$$overlap(X,Y) = \frac{|X \cap Y|}{\min(|X|,|Y|)} \tag{2}$$

To compare the performance of deterministic matching and fuzzy matching, the output data information of each technique such as the number of integrated observations into the primary data and distribution will be used for evaluation.

4.0 RESULT

After the completion of statistical matching, the preservation of variables are measured using Overlap index. The result of the measurement shows that the overlap index equal to 1 (Overlap index = 1). The result of overlap index is expected in this study as the trialability variable from Capterra are merged with the data from Software Advice. This merging process cause the data from Capterra becomes a subset in Software Advice. Vijaymeena, M. K. and Kavitha, K. stated that when a set becomes a subset of another set, the overlap index is equal to 1 which can be applied to the case of this study [20].

A comparison is conducted between the output data of deterministic matching and fuzzy matching in Table 5. The result is obtained from merging the data of "Trialability" feature from Capterra dataset (auxiliary data) into Software advice data (primary data). There are 341 observations in the Capterra data, with 271 "No" and 70 "Yes". Result in Table 6 also shows that using deterministic matching techniques, 121 observations from Capterra are merged into the primary data. On the other hand, fuzzy matching merged 126 observations into the primary data.

As a summary of the comparison, fuzzy matching merges slightly more data compared to deterministic matching. This is due to fuzzy matching match data by computing the similarity distance of 2 strings. Therefore, fuzzy matching allows the match of observations by considering spacing and letter cases as in the case of this study as shown in Table 2 which shows the additional 5 observations that is matched using fuzzy matching. This makes fuzzy matching more robust in matching data extracted from the internet compared to deterministic matching.

Table 2: Comparison of number of data merged between deterministic matching technique and fuzzy matching technique

	Capterra		**Data merging**			
			Deterministic matching		**Fuzzy matching**	
	Count	**Distribution**	**Count**	**Distribution**	**Count**	**Distribution**
No	271	0.794721408	84	0.694214876	88	0.701492537
Yes	70	0.205278592	37	0.305785124	38	0.298507463
Total	341	1	121	1	126	1

5.0 DISCUSSION

This study explores 2 different techniques of merging data from non-homogenous websites in web scraping which are deterministic matching and fuzzy matching. The procedure of the proposed approach for merging data are merging the data and comparing the performance of merging data of two different techniques.

The record linkage used in the merging data from two non-homogenous websites shows that record linkage can be used to merge data from different websites. The overlap index computed after the data merging procedure shows that the data from donor has been successfully imputed into the data set of recipients thus becoming a subset of recipient data set. This study explore 2 different techniques of matching data which are deterministic matching and fuzzy matching. Deterministic matching is a simple to use and proven technique used in several previous studies[10-15]. The similar technique was also use by Jamali-Phiri et al to address the data deficiencies in assistive technology [19]. The study uses data from 2017 survey on Living conditions among persons with disabilities in Malawi and the 2015-16 Malawi Demographic and Health survey for statistical matching. The result of the overlap index from the study was 0.909 which was close to 1. This indicates that deterministic matching technique can be used reliably to create synthetic data from merging two data sets. However, there also studies stated that the main disadvantage of deterministic matching is that the data need to exact same match to be matched thus making it unable to match data that has some dissimilarity such as letter case and spacing which are common in the web scrapped data.

Therefore, this study explores the techniques of fuzzy matching in matching web scrapped data. The nature of how fuzzy matching match data enables it to match data that is partially similar. This makes fuzzy matching suitable to use in matching data that has no unique identifier such as a data ID in a database. This practicability has been proven by several studies which use fuzzy matching in profile matching applications to match raw data that has no unique identifier. This study also shows that fuzzy matching is suitable to be use in the domain of web scrapping compared to deterministic matching. The string in the data we use as match identifier have inconsistencies such as letter cases and spacing. The ability of fuzzy matching to match partial similarity enable it to match data in this case.

However, the main disadvantage of fuzzy matching compared to deterministic matching in web scrapping is the computing complexity. Compared to fuzzy matching, deterministic matching is easier to use and understand. From the process of using both techniques in merging data, the deterministic matching techniques can be simply applied into the data merging process while fuzzy matching require us to determine the suitable threshold for the desired result.

6.0 CONCLUSION

As a summary, this study explores an alternative data matching technique known as fuzzy matching and A comparison of it performance of merging web data with deterministic matching technique is also made in this study. The comparison shows that fuzzy matching technique can merge slightly more data compared to deterministic matching which make a better technique to use in merging the web data used in this study. However, fuzzy matching requires user to determine the threshold of similarity which can be a challenge as the similarity threshold is dependent on the usage of data in different scenarios.

6.1 Future work

As the limitation stated in the conclusion, effort should be put into future work to investigate the optimal way of determining similarity threshold. This can ensure fuzzy matching is easier to be implemented in different scenario of merging web data.

ACKNOWLEDGMENTS

This research was supported by Universiti Teknikal Malaysia Melaka (UTeM) grant number KHAS-KKP/2021/FTMK/C00003.

REFERENCES

[1] J. Hillen, "Web scraping for food price research," *British Food Journal,* vol. 121, no. 12, pp. 3350-3361, 2019, doi: 10.1108/BFJ-02-2019-0081.

[2] C. G. Konny, B. K. Williams, and D. M. Friedman, "Big data in the us consumer price index: Experiences and plans," *Big Data for 21st Century Economic Statistics,* 2019.

[3] C. Osbat *et al.*, "What micro price data teach us about the inflation process: web-scraping in PRISMA," 2022.

[4] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings,* vol. 3, no. 1, pp. 91-99, 2022/06/01/ 2022, doi: https://doi.org/10.1016/j.gltp.2022.04.020.

[5] K. Sankpal, "A Review on Data Normalization Techniques," *International Journal of Engineering Research and,* vol. V9, 07/06 2020, doi: 10.17577/IJERTV9IS060915.

[6] S. Rässler, "Data Fusion: Identification Problems, Validity, and Multiple Imputation," *AUSTRIAN JOURNAL OF STATISTICS Volume,* vol. 33, pp. 153-171, 01/01 2004, doi: 10.17713/ajs.v33i1&2.436.

[7] R. D'Allerto and M. Raggi, "From collection to integration: Non-parametric Statistical Matching between primary and secondary farm data," *Statistical Journal of the IAOS,* vol. 37, pp. 1-11, 04/16 2021, doi: 10.3233/SJI-200644.

[8] G. Saporta, "Data fusion and data grafting," *Computational Statistics & Data Analysis,* vol. 38, no. 4, pp. 465-473, 2002/02/28/ 2002, doi: https://doi.org/10.1016/S0167-9473(01)00072-X.

[9] I. Lewaa, M. S. Hafez, and M. A. Ismail, "Data integration using statistical matching techniques: A review," *Statistical Journal of the IAOS,* vol. 37, pp. 1391-1410, 2021, doi: 10.3233/SJI-210835.

[10] M. Jamali-Phiri *et al.*, "Addressing data deficiencies in assistive technology by using statistical matching methodology: a case study from Malawi," *Disability and Rehabilitation: Assistive Technology,* vol. 18, no. 4, pp. 415-422, 2023/05/19 2023, doi: 10.1080/17483107.2020.1861118.

[11] F. D. d'Ovidio, P. Perchinunno, and L. Antonucci, "Data Integration Techniques for the Identification of Poverty Profiles," *Social Indicators Research,* vol. 156, no. 2, pp. 515-531, 2021/08/01 2021, doi: 10.1007/s11205-019-02255-0.

[12] M.-R. Namazi-Rad, R. Tanton, D. Steel, P. Mokhtarian, and S. Das, "An unconstrained statistical matching algorithm for combining individual and household level geo-specific census and survey data," *Computers, Environment and Urban Systems,* vol. 63, pp. 3-14, 2017/05/01/ 2017, doi: https://doi.org/10.1016/j.compenvurbsys.2016.11.003.

[13] S. R. and R. S., "Web Scraping Online Newspaper Death Notices for the Estimation of the Local Number of Deaths," in *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019)*, 2019, vol. 5: HEALTHINF, pp. 319-325, doi: 10.5220/0007382603190325.

[14] B. E. Shook-Sa, M. G. Hudgens, A. L. Kavee, and D. L. Rosen, "Estimating the Number of Persons with HIV in Jails Via Web Scraping and Record Linkage," *Journal of the Royal Statistical Society Series A: Statistics in Society,* vol. 185, no. Supplement_2, pp. S270-S287, 2022, doi: 10.1111/rssa.12909.

[15] T. Tuoto, D. Fusco, and L. Di Consiglio, "Exploring Solutions for Linking Big Data in Official Statistics," in *Studies in Theoretical and Applied Statistics*, Cham, C. Perna, M. Pratesi, and A. Ruiz-Gazen, Eds., 2018// 2018: Springer International Publishing, pp. 49-58.

[16] G. Navarro, "A Guided Tour to Approximate String Matching," *ACM Computing Surveys,* vol. 33, 04/06 2000, doi: 10.1145/375360.375365.

[17] H. Suman, H. Tamkiya, and A. Kushwah, *Candidate Background Verification Using Machine Learning and Fuzzy Matching*. 2020.

[18] T. De Waal, *Statistical matching: Experimental results and future research questions*. 2015.

[19] M. Jamali-Phiri *et al.*, "Addressing data deficiencies in assistive technology by using statistical matching methodology: a case study from Malawi," *Disability and Rehabilitation: Assistive Technology,* pp. 1-15, 2020, doi: 10.1080/17483107.2020.1861118.

[20] V. M K and K. K, "A Survey on Similarity Measures in Text Mining," *Machine Learning and Applications: An International Journal,* vol. 3, pp. 19-28, 03/30 2016, doi: 10.5121/mlaij.2016.3103.