Original Paper

Application of Efficient Data Cleaning Using Text Clustering for Semistructured Medical Reports to Large-Scale Stool Examination Reports: Methodology Study

Hyunki Woo[1*], BS; Kyunga Kim[1,2*], PhD; KyeongMin Cha[1], MS; Jin-Young Lee[3], MD, PhD; Hansong Mun[3], MD, PhD; Soo Jin Cho[3], MD; Ji In Chung[3], MD, PhD; Jeung Hui Pyo[3], MD; Kun-Chul Lee[4], PhD; Mira Kang[1,3], MD, PhD

[1]Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul, Republic of Korea
[2]Statistics and Data Center, Research Institute for Future Medicine, Samsung Medical Center, Seoul, Republic of Korea
[3]Center for Health Promotion, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[4]Jason TG, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Mira Kang, MD, PhD
Center for Health Promotion
Samsung Medical Center
Sungkyunkwan University School of Medicine
81 Irwon-ro, Gangnam-gu
Seoul, 06351
Republic of Korea
Phone: 82 2 3410 3882
Fax: 82 2 3410 0054
Email: mira90.kang@samsung.com

Abstract

Background: Since medical research based on big data has become more common, the community's interest and effort to analyze a large amount of semistructured or unstructured text data, such as examination reports, have rapidly increased. However, these large-scale text data are often not readily applicable to analysis owing to typographical errors, inconsistencies, or data entry problems. Therefore, an efficient data cleaning process is required to ensure the veracity of such data.

Objective: In this paper, we proposed an efficient data cleaning process for large-scale medical text data, which employs text clustering methods and value-converting technique, and evaluated its performance with medical examination text data.

Methods: The proposed data cleaning process consists of text clustering and value-merging. In the text clustering step, we suggested the use of key collision and nearest neighbor methods in a complementary manner. Words (called values) in the same cluster would be expected as a correct value and its wrong representations. In the value-converting step, wrong values for each identified cluster would be converted into their correct value. We applied these data cleaning process to 574,266 stool examination reports produced for parasite analysis at Samsung Medical Center from 1995 to 2015. The performance of the proposed process was examined and compared with data cleaning processes based on a single clustering method. We used OpenRefine 2.7, an open source application that provides various text clustering methods and an efficient user interface for value-converting with common-value suggestion.

Results: A total of 1,167,104 words in stool examination reports were surveyed. In the data cleaning process, we discovered 30 correct words and 45 patterns of typographical errors and duplicates. We observed high correction rates for words with typographical errors (98.61%) and typographical error patterns (97.78%). The resulting data accuracy was nearly 100% based on the number of total words.

Conclusions: Our data cleaning process based on the combinatorial use of key collision and nearest neighbor methods provides an efficient cleaning of large-scale text data and hence improves data accuracy.

(J Med Internet Res 2019;21(1):e10013) doi: 10.2196/10013

KEYWORDS

data cleaning; text clustering; key collision; nearest neighbor methods; OpenRefine

Introduction

In all of the industries, including the medical field, complex and diverse (structured, semistructured, unstructured) data have been growing dramatically for decades [1-3]. Although most health data have been digitalized, it is still not easy to handle medical records such as examination reports or physician's notes because they are historically based on paper records and generated data mainly in semistructured or unstructured forms. In addition, they may contain a variety of nonidentical duplicates, typographical errors, inconsistencies, and data entry problems [4-7].

High performance analysis requires clean and high-quality data to yield reliable results [8-11]. Therefore, efficient data cleaning takes precedence to improve the quality of data and obtain accurate analysis results [12]. However, researchers are commonly faced with many obstacles in transforming the data into a clean and high-quality dataset owing to diverse patterns of typographical errors and duplicates.

For text analysis of semistructured or unstructured data, we can use a paid program such as SAS Content Categorization (SAS Institute Inc) or IBM Watson Content Analytics (IBM) [13,14]. However, these programs are very expensive and are not readily available to individual researchers because they are mainly sold to companies or research groups. Also, these programs require extensive practice and experience.

Data cleaning using Excel's "remove duplicates" function has been done before, but it is mostly unpractical to clean the data using Excel tools. Some of the nonidentical duplicates still remain because they are not recognized as duplicates when special characters or punctuations appear [5,6,15,16]. Duplicate detection tools such as the Febrl system, TAILOR, and BigMatch were also used in cleaning data. However, Febrl has usability limitations such as slowness, unclear error messages, and complicated installations [17-20]. The listed programs are rather complex to the average users who do not have experience with programming and language functions.

Many researchers who interpret and clean the local datasets are domain experts and are not familiar with the programming language [21]. Thus, researchers need user friendly cleaning tools. OpenRefine can identify all types of strings and remove duplicates without the difficulties of programming and is a free, open source tool. OpenRefine contains the following 2 clustering methods: key collision methods and nearest neighbor methods. We proposed a data cleaning process using both text clustering methods in OpenRefine to improve accuracy of semistructured data.

Methods

We performed data cleaning of 574,266 stool examination reports conducted at Samsung Medical Center from 1995 to 2015. Data for this study were extracted from DARWIN-C, the clinical data warehouse of Samsung Medical Center. According to the data cleaning process proposed in Figure 1, we conducted data cleaning by clustering and merging parasite names and investigated its performance.

As described in Figure 1, the proposed data cleaning process consists of the following 4 steps: preprocess, text facet, systematic cleaning, and manual cleaning. In the preprocess, only names related to parasites (ie, helminth or protozoa) in raw text data were extracted using the regular expression functions of STATA MP 14.2 version [22]. The extracted words were then uploaded on OpenRefine 2.7. In the text facet step, the number of occurrences was browsed for each word.

The systematic cleaning step consists of text clustering and value-merging. Two clustering methods (ie, key collision and nearest neighbor) are used in a complementary manner to identify word clusters, each of which is expected to contain a correct word and its wrong representations with diverse forms of typographical errors (called "wrong values"). Key collision methods work by creating an alternate representation of a key that contains only the most significant or meaningful parts of a string and by clustering different strings together based on the same key. Because key collision methods are fast and simple in a variety of contexts, they have been often used for text clustering. We sequentially used 4 key collision methods including fingerprint, N-gram fingerprint, Metaphone3, and Cologne phonetic in OpenRefine. Nearest neighbor methods (also known as kNN) are widely used for clustering as well. These methods are slower but more accurate because they calculate the distance between each value. We sequentially used two nearest neighbor methods, the Levenshtein distance method and Prediction by Partial Matching method in OpenRefine. We combined both methods to enhance the accuracy [23].

For each identified cluster, the wrong values are converted to their correct word by value-merging. Because OpenRefine provides a convenient user interface that lists the correct word and its wrong values in each cluster in descending order of occurrence frequency, researchers can easily recognize the correct word and conduct the value-merging task. For "*Clonorchis sinensis*" in stool examination report data, a variety of wrong expressions were noticed in the same cluster, such as clonorchis sinesis, clnorchis sinensis, clonorchis cinensis, clonrchis sinensis, and clornorchis sinensis (Figure 2). By looking at the word list, we were able to efficiently choose "*Clonorchis sinensis*" as the correct word and make a quick decision to convert all the others to "*Clonorchis sinensis*" In the final step, the remaining words that did not belong to any cluster were investigated and manually cleaned when necessary.

Figure 1. Flow chart of our data cleaning process. PPM: prediction by partial matching.

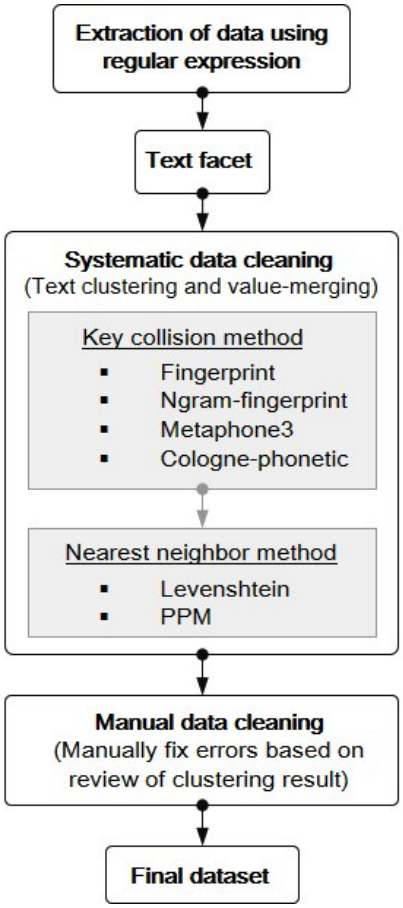
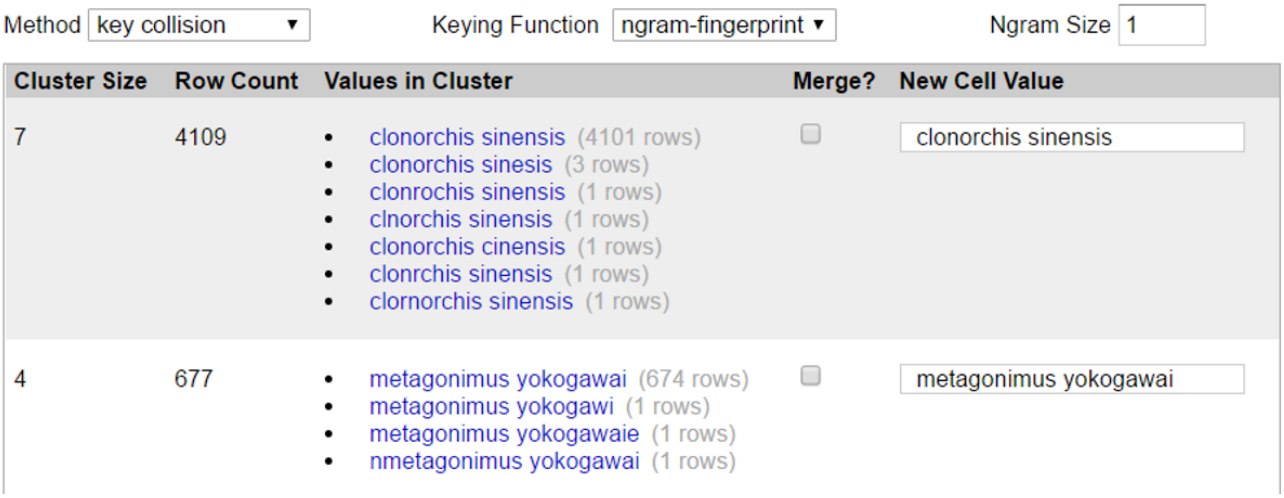


Figure 2. Representative screenshot of OpenRefine interface used for value-merging task.



Results

A total of 1,167,104 words in 574,266 stool examination reports were surveyed, and words not related to the names of helminth or protozoa were excluded from the study. We discovered 30 correct words and 45 patterns of typographical errors and duplicates (Multimedia Appendix 1). The key collision methods were able to cluster the patterns of typographical errors and duplicates with the correct word except for 6 patterns. The nearest neighbor methods were able to cluster the patterns of typographical errors and duplicates with the correct word except for 2 patterns (Table 1).

Table 1. List of typographical errors that could not be clustered with the correct word by each method.

Correct word	Typographical error	Key collision	Nearest neighbor
Negative	• Native	✗[a]	✗
	• Negaitve	✓[b]	✗
Endolimax	• Eolimax	✗	✓
	• Endolix	✗	✓
Entamoeba	• Etamoeba	✗	✓
Lamblia	• Lamdlia	✗	✓
	• G.lamblia	✗	✓

[a]✗: Typographical error is not clustered with correct word.

[b]✓: Typographical error is clustered with correct word.

Table 2. Correction rates by each method.

Method	Correction rate by the number of typographical error patterns[a], %	Correction rate by the number of typographical error words[b], %
Key collision	86.67	91.67
Nearest neighbor	95.56	97.22
Using both	97.78	98.61

[a]The number of corrected typographical error patterns divided by the total number of typographical error patterns multipled by 100 (%).

[b]The number of corrected typographical error words divided by the total number of typographical error words multipled by 100 (%).

The word “native” was the only pattern not clustered as “negative” out of all typographical errors by any clustering method because of the high inconsistency rate of the 2 words (2/6 characters, 33%). All typographical errors and duplicates except “native” were clustered correctly. We achieved a high correction rate of 98.61% by the number of typographical error words and 97.78% by the number of typographical error patterns when using both clustering methods (Table 2). After systematic data cleaning of 1,167,104 words, only 1 word with a typographical error remained and was revised manually. Thus, the accuracy of systematic data cleaning was nearly 100% based on the number of total words.

Discussion

Many researchers have made great efforts to study data analytics methodology, but there have been relatively few studies on data cleaning methodology for unexpected typographical errors [24,25]. It is rare to find a report that quantitatively analyzes the performance of data cleaning methods because they are often undocumented and used in nonofficial ways [24]. In this study, we suggested an efficient way of data cleaning for large-scale medical text data and investigated its cleaning performance. Although several methods of text analysis exist, it is not easy for general researchers to use these methods. Most methods are not readily available or have limitations in usability. Therefore, there is a need for more feasible and user friendly methods for cleaning large-scale text datasets.

We employed OpenRefine for data cleaning because of the following advantages. First, individual researchers can easily access and use OpenRefine because it is a free and open source tool. Second, OpenRefine provides researchers with an easy interface to clean the data without difficulties of programming. Third, one can easily fix rare typographical errors (which are not automatically corrected) manually and have the opportunity to modify false positive clustering [6,23].

However, we still need much effort to review each clustering result and decide whether to merge, especially in cases where the number of clustering is extremely large. In addition, formal technical support for OpenRefine is not available, and it is supported by user forums or communities. Despite these limitations, OpenRefine is a useful and effective support tool for labor-intensive and time-consuming data cleaning of semistructured data.

Our data cleaning process can be applied to other types of semistructured text data because we observed that the combinatorial use of key collision and nearest neighbor methods resulted in efficient and reliable data cleaning.

Acknowledgments

This study was supported by Samsung Medical Center grant (SMX1170601).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Patterns of parasite names in stool examination reports.

[PDF File (Adobe PDF File), 39KB-Multimedia Appendix 1]

References

1. Zhang Y, Qiu M, Tsai C, Hassan MM, Alamri A. Health-CPS: Healthcare Cyber-Physical System Assisted by Cloud and Big Data. IEEE Systems Journal 2017 Mar;11(1):88-95. [doi: 10.1109/Jsyst.2015.2460747]
2. Das T, Kumar P. Big data analytics: A framework for unstructured data analysis. Int J Eng Sci Technol 2013;20135(1):A.
3. Tsai C, Lai C, Chiang M, Yang LT. Data Mining for Internet of Things: A Survey. IEEE Commun. Surv. Tutorials 2014;16(1):77-97. [doi: 10.1109/Surv.2013.103013.00206]
4. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014;2:3 [FREE Full text] [doi: 10.1186/2047-2501-2-3] [Medline: 25825667]
5. Groves A. Beyond Excel: how to start cleaning data with OpenRefine. Multimedia Information and Technology 2016;201642(2):18-22.
6. Ham K. OpenRefine (version 2.5). http://openrefine.org. Free, open-source tool for cleaning and transforming data. J Med Libr Assoc 2013 Jul;101(3):233-234 [FREE Full text] [doi: 10.3163/1536-5050.101.3.020]
7. Gallant K, Lorang E, Ramirez A. Tools for the digital humanities: a librarian's guide. 2014. URL: https://mospace.umsystem.edu/xmlui/bitstream/handle/10355/44544/ToolsForTheDigitalHumanities.pdf?sequence=1 [accessed 2018-11-14] [WebCite Cache ID 73uqoKFfO]
8. Chu X, Ilyas IF. Qualitative data cleaning. Proc. VLDB Endow 2016 Sep 01;9(13):1605-1608. [doi: 10.14778/3007263.3007320]
9. Wang L, Jones R. Big Data Analytics for Disparate Data. American Journal of Intelligent Systems 2017;20177(2):39-46. [doi: 10.5923/j.ajis.20170702.01]
10. Zhang S, Zhang C, Yang Q. Data preparation for data mining. Appl Artif Intell 2003 May-Jun;17(5-6) 2003:375-381. [doi: 10.1080/08839510390219264]
11. Anagnostopoulos I, Zeadally S, Exposito E. Handling big data: research challenges and future directions. J Supercomput 2016 Feb 25;72(4):1494-1516. [doi: 10.1007/s11227-016-1677-z] [Medline: 26811110]
12. Rahm E, Do H. Data cleaning: Problems and current approaches. IEEE Data Eng Bull 2000;200023(4):3-13.
13. Chakraborty G, Pagolu M, Garla S. Text mininganalysis: practical methods, examples,case studies using SAS. Cary, NC: SAS Institute; 2014. URL: http://support.sas.com/publishing/pubcat/chaps/65646.pdf [accessed 2018-11-14] [WebCite Cache ID 73uraoski]
14. Zhu W, Foyle B, Gagné D, Gupta V, Magdalen J, Mundi A. IBM Watson Content Analytics: Discovering Actionable Insight from Your Content. New York, USA: IBM Redbooks; 2014.
15. Katsanevakis S, Gatto F, Zenetos A, Cardoso A. Management of Biological Invasions. 2013. How many marine aliens in Europe URL: https://pdfs.semanticscholar.org/4dbe/0bc865391bd3a6100e112e7046675341ba18.pdf [accessed 2018-11-14] [WebCite Cache ID 73uswuwIP]
16. Dallas DC, Guerrero A, Khaldi N, Borghese R, Bhandari A, Underwood MA, et al. A peptidomic analysis of human milk digestion in the infant stomach reveals protein-specific degradation patterns. J Nutr 2014 Jun;144(6):815-820 [FREE Full text] [doi: 10.3945/jn.113.185793] [Medline: 24699806]
17. Hassanien A, Azar A, Snasel V, Kacprzyk J, Abawajy J. Big Data in Complex Systems: Challenges and Opportunities Berlin, Germany. New York City, United states of America: Springer Publishing Company; 2015.
18. Selvi P, Priyaa D. A Perspective Analysis on Removal of Duplicate Records using Data Mining Techniques: A Survey. International Journal of Engineering Technology Science and Research 1;20163(12):36-41.
19. Higazy A, El TT, Yousef A, Sarhan A. Web-based Arabic/English duplicate record detection with nested blocking technique. 2013 Presented at: Computer Engineering & Systems (ICCES), 8th International Conference on IEEE; 2013; Cairo, Egypt p. 313-318.
20. Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate Record Detection: A Survey. IEEE Trans. Knowl. Data Eng 2007 Jan;19(1):1-16. [doi: 10.1109/Tkde.2007.250581]
21. Larsson P. courses.cs.washington.edu. 2013. URL: https://courses.cs.washington.edu/courses/cse544/13sp/final-projects/p12-plarsson.pdf [accessed 2018-11-14] [WebCite Cache ID 73utiScvt]
22. Medeiros R. Using regular expressions for data management in Stata. West Coast Stata Users' Group Meetings, Stata Users Group 2007:-.
23. Clustering In Depth. 2016 URL: https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth [accessed 2018-11-14] [WebCite Cache ID 73utyFlHG]

24. Maletic J, Marcus A. Data Cleansing: Beyond Integrity Analysis. IQ 2000:2000-2209.
25. Chu X, Ilyas I, Krishnan S, Wang J. Data cleaning: Overview and emerging challenges. 2016 Presented at: Proceedings of the International Conference on Management of Data. ACM; 2016; San Francisco, CA, USA p. 2201-2206.

Edited by G Eysenbach; submitted 06.02.18; peer-reviewed by W Raghupathi, S Barteit, F Shen, F Wang; comments to author 30.07.18; revised version received 23.09.18; accepted 12.10.18; published 08.01.19

Please cite as:
Woo H, Kim K, Cha K, Lee JY, Mun H, Cho SJ, Chung JI, Pyo JH, Lee KC, Kang M
Application of Efficient Data Cleaning Using Text Clustering for Semistructured Medical Reports to Large-Scale Stool Examination Reports: Methodology Study
J Med Internet Res 2019;21(1):e10013
URL: https://www.jmir.org/2019/1/e10013/
doi: 10.2196/10013
PMID: 30622098