# Comparing Context Free Cleaning for Different Users Applied to Tagged Data using Similarity Metrics for Enhanced Tag Cloud

**Rinku Chavda[1], Dr. Sohil Pandya[2]**
[1,2] Assistant Professor, MCA Department,
[1,2] Sardar Vallabhbhai Patel , Institute of Technology (SVIT), Vasad, India
[1]rinku.chavda.12@gmail.com, [2]sohilpandya@gmail.com

**Abstract**

Today is the era of a scientific invention of resources on the internet; for these universities can go ahead for gaining competitive advantage only by trained data analysis. The present article highlights the subject of free data cleaning for advanced tag cloud by evaluating values of user-defined “Tags”, through the different string similarity metrics, where “Tags” are assigned by users which are given to the referenced resource. Authors of the present article suggested an honest formula to analyze the matching of value to appropriate values of Tags. Abandoned string similarity metrics were used, to explore the differences of two strings and observe the results. Experimental results highlight the approach that can effectively clean the data without referenced data.

**Keywords:** Context free data cleaning, Information Retrieval, String similarity metrics, Tag Cloud

## 1. INTRODUCTION

Utilization of Internet is rising in recent ages so it is a primary duty of a computer scientist to expend more and easier services on the Internet to search for specific resources. While searching for a document inside an organization document repository is a slovenly and lengthened job. Due to their unstructured nature, haphazard storage and different naming rules it becomes untidy and lengthened procedure, but required to be regained instantly.

Gujarat Technological University (GTU) is a big organization, authors have taken here as to clear the current concept as an example. University has numerous notifications, circulars; notes which are published regularly on the frequent basis. From these uploaded documents, they are tagged by users for the future use as per their need.

In the present paper, a researcher has put an honest attempt to implement an improved tag cloud designed which will help for information retrieval using data cleaning process, using various string similarity metrics for different users. For the data cleaning implementation portion, tags with their frequencies which are given by users are taken for this model as an input. During tagging of resource, users utilize their own word as tag designed for their usefulness for upcoming reference.

For one resource, sometimes multiple users use the different or same word as tags to tag a particular resource. While using the same tag the several of possibility occurs some examples are a correct spell of tag, incorrect spell of tag, similar kind of tag or shortening of a tag are used probably.

Therefore, to solve these circumstances, authors have utilized similarity metrics where the tags are balanced to find resemblance among tags, perform the substitution of tags and produce tag cloud for information retrieval from cleaned tag list which turns into a fine-tuned list of tags following the application of data cleaning process.

## 2. Data Cleaning For Enhanced Tag Cloud

Data cleaning is the process of noticing and shifting corrupt or inaccurate tags from a record set and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the tag list and then replacing with the correct data. In the next stage of cleaning, a reference dataset must to be consistent inside the numerous users though tagging any resources. The variations observed or detached may have been formerly sourced by user entry errors or by different data dictionary definitions of similar entities for tagging.

A tag cloud (word cloud or weighted list in visual design) is a visual demonstration for text data, more specifically used to depict keyword metadata (tags) on websites, or to imagine free form text [17].

Conclusively in very few words of this article is can be listed as, grounded on the frequency of tags, in downhill order, they are matched with other tags, and based on the match value, it is definite that whether they should be substituted or not?
In next section the procedure demonstrate that examines suitability of tags to become member of reference dataset and/or replace the tags by other matching tags which are often used among different users.

The proposed method has two key mechanisms: clustering and nearest string. It has an significant parameter acceptableDist, which is the minimum acceptable distance required during comparing and altering (ranges from 0.0 to 1.0, where 0.0 is non-similar string and 1.0 is similar string).To measure the distance, scientifically used Damerau Levenshtein string similarity metrics:

The Damerau-Levenshtein distance as below is a distance (string metric) between two strings, i.e., finite sequence of symbols, given by counting the minimum number of operations needed to transform one string into the other, where an operation is defined as an insertion, deletion, or substitution of a single character, or a transposition of two adjacent characters.

$$d_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if} \min(i,j) = 0, \\ \min \begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 \\ d_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \\ d_{a,b}(i-2,j-2)+1 \end{cases} & \text{if } i,j > 1 \text{ and } a_i = b_{j-1} \text{ and } a_{i-1} = b_j \\ \min \begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 \\ d_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

…….. [20]

Where, each recursive call matches one of the cases covered by the Damerau–Levenshtein distance:

- da,b(i-1,j)+1corresponds to a deletion (from a to b).
- da,b(i,j-1)+1} corresponds to an insertion (from a to b).
- da,b(i-1,j-1)+1(ai ≠ bj)corresponds to a match or mismatch, depending on whether the respective symbols are the same.
- da,b (i-2,j-2)+1 corresponds to a transposition between two successive symbols.

The below algorithm is refered from reference [16] which is very scientific and useful for proposed research study for different users:

1. Convert all the Alphanumeric values to Number format e.g. I,one,First,1st, 1 ST to1
2. Keep list of Domain Specific entries of tags e.g. Degree Engineering, Deg. Engi., Bachelor of Engineering to B.E.
3. Retrieve list of tags (listing) with its frequency in descending order.
4. Repeat while (listing has tags to compare)

a. Read tag to compare from listing

b. Retrieve list of tags (listj) with its frequency in descending order where freq(tagj ) ≤ freq(tagi) and tagi∉ listj.

c. Repeat while (listj has tags to compare)

i. Convert tagi and tagj to lowercase

ii. ii. Compare tagi with tagj

iii. If the compare value is greater or equal 0.9 thresholds value, then perform replacement of tags else keep that two tags as a separate tags

## 3. Experimental Results & Discussion

The Delicious.com is one of the esteemed websites for social bookmarking over the Internet; it is also called web-based tagging system. The proposed method is applied to experimental data taken from the account of delicious.com of two distinct users.

This website allows you to add resource as a bookmark as well as attach some extra information related to the resource like Title of URL, different keywords based on the users' point of view and remark also [16]. Based on the data can get a list of tags with its recurrences and the tag cloud is as below:

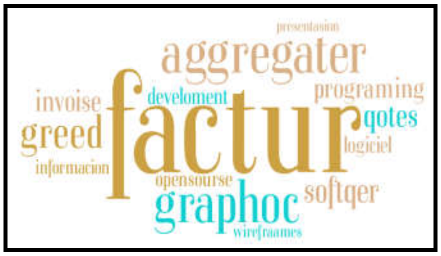

**Figure 1. Tag Cloud before Applying Data Cleaning for User-1**

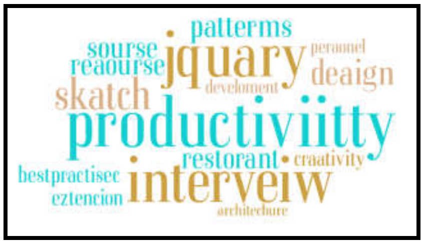

**Figure 2. Tag Cloud before Applying Data Cleaning for User-2**

After applying Damerau-Levenshtein string similarity metric on two dataset of different users, several results like how many records replaced (total, correctly, incorrectly, not replaced in context of spelling mistake), were found and are discussed in this section. Damerau-Levenshtein algorithm with 0.8 similarity metrics value, we found replacement rules as shown in Table 1 based on two users tagging data. The table is showing count for replacement of tags with correct tags where the tags are misspelled.

After applying Damerau-Levenshtein algorithm, Table: 1 shows values of replaced records, correctly replaced, incorrectly replaced, not replaced where the acceptableDis is greater or equal to 0.8. There were about 406 and 386 records of user 1 and user 2, respectively out of them total 241 and 259 records were marked as replacement. From those replacements, 85.71% 22.82% and 6.22% were identified as correct replacement, incorrect replacement, and not replaced tags; respectively for user 1 and 80.52%, 22.82% and 6.22% were identified as correct replacement, incorrect replacement, and not replaced tags, respectively for user 2. Based on replacement, the generated tag clouds of user 1 and user 2 are as below:

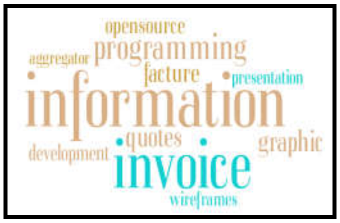

Figure 3. Tag Cloud after Applying Data Cleaning for User-1

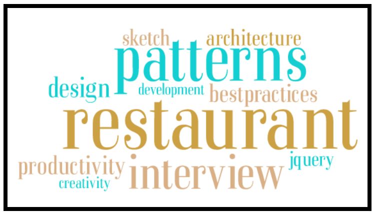

Figure 4. Tag Cloud after Applying Data Cleaning for User-2

For User 1, 241 records out of which 68 records contain incorrect tags which are entered by users. Using Damerau-Levenshtein algorithm, 60 records out of 68 records of incorrect tags are replaced with correct tags (Table: 1). Hence, from correctly replaced list of tags, 6.22 % tags remains unchanged.

For User 2, 259 records out of which 68 records contain incorrect tags which are entered by users. Using Damerau-Levenshtein algorithm, 62 records out of 76 records of incorrect tags are replaced with correct tags (Table: 2). Hence, from correctly replaced list of tags, 5.41 % tags remains unchanged.

Table 1. Incorrect Tags Altered For User-1

| Tag | Freq. | Replaced With | Tag Freq. |
|---|---|---|---|
| Aggregater | 8 | Aggregator | 1 |
| factur | 9 | Facture | 2 |
| graphoc | 8 | Graphic | 8 |
| softqer | 6 | - | - |
| develoment | 5 | Development | 8 |
| Informacion | 5 | information | 28 |
| Invoice | 4 | Invoice | 10 |
| Qotes | 2 | Quotes | 2 |
| Wirefraames | 2 | wireframes | 2 |
| Logiciel | 2 | - | - |
| Programing | 6 | programming | 9 |
| Opensourse | 2 | opensource | 2 |
| Presentasion | 2 | presentation | 2 |
| Greed | 7 | - | - |

**Table 2. Incorrect Tags Altered For User-2**

| Tag | Freq. | Replaced With | Tag Freq. |
|---|---|---|---|
| Deaign | 5 | Design | 9 |
| Jquary | 8 | Jquery | 6 |
| Development | 2 | development | 3 |
| Bestpractisec | 5 | best practices | 10 |
| Eztencion | 4 | - | - |
| Architechure | 2 | architecture | 8 |
| Patterms | 7 | Patterns | 15 |
| Productiviitty | 9 | productivity | 10 |
| Reaourse | 5 | - | - |
| Sourse | 3 | - | - |
| Creativity | 5 | creativity | 1 |
| Interview | 8 | Interview | 13 |
| Skatch | 6 | Sketch | 5 |
| Restorant | 5 | restaurant | 20 |
| Peraonel | 2 | - | - |

Following results, percentage of correctly replaced tags (CR %), percentage of incorrectly replaced tags (IR %) and percentages of not replaced tags (NR %) with respect to incorrect tag were derived. Out of various acceptableDist for 0.8 there were 85.71% and 80.52% values were replaced correctly for user 1 and user 2, respectively; 21.57% and 23.94 values replaced incorrectly for user 1 and user 2, respectively, 5.88% and 5.41% values are not replaced for user 1 and user 2 respectively, with respect to total replacements (as depicted in Table 3 and Fig. 5, 6 & 7).

**Table 3. Replacement (Correct, Incorrect, and Not Replaced)**

| Threshold 0.8<= | Total | Correct | Incorrect | Not Replaced |
|---|---|---|---|---|
| User 1 | 241 | 85.71 | 22.82 | 6.22 |
| User 2 | 259 | 80.52 | 23.94 | 5.41 |

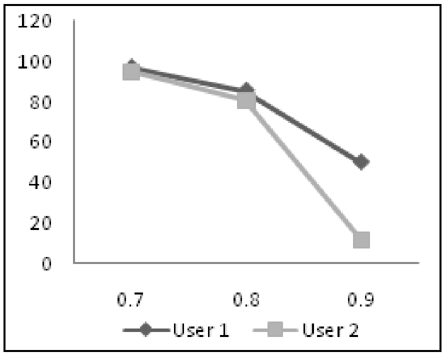

**Figure 5. Correctly Updated Tags for User-1 and User-2**

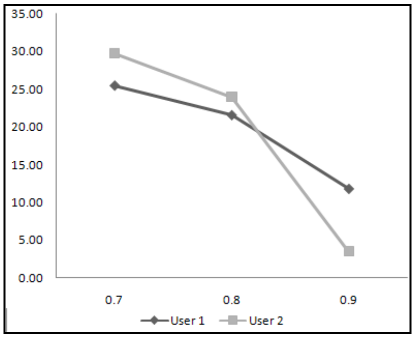

**Figure 6. In-correctly Updated Tags for User-1 and User-2**

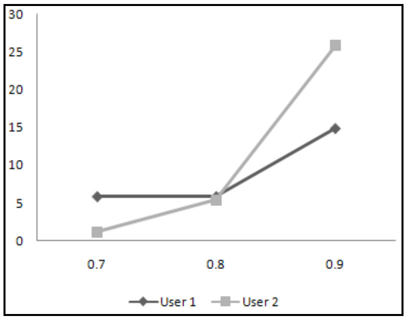

**Figure 7. Not Replaced Tags for User-1 and User-2**

## 4. Conclusion

In above experiments, one string similarity metric i.e. Damerau-Levenshtein for two users was evaluated. It is possible that other metrics or functions and/or various combinations of them for multiple users, as per the requirements, may give healthier results and this should be discovered in further experiments. The consequence of the experiments the accuracy of the algorithm and which motivate to use it for context free data cleaning of tags to generate improved tag cloud for improved information retrieval.

## References

[1] A.W. Rivadeneira, D.M. Gruen, M.J. Muller, and D.R. Millen, "Getting our head in the clouds: Toward evaluation studies of tagclouds", Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, (**2007**), pp. 995–998.

[2] Lukasz Ciszak "Application of Clustering and Association Methods in Data Cleaning", in Proc. of Int. Multiconference on Computer Science and Information Technology, Vol. 3, (2008), pp. 97-103.

[3] Hearst, Marti A., and Daniela Rosner. "Tag clouds: Data analysis tool or social signaller?" In Hawaii International Conference on System Sciences, Proceedings of the 41st Annual, pp. 160-160. IEEE, 2008.

[4] James Sinclair, and Michael Cardew-Hall, "The folksonomy tag cloud: when is it useful?" in Journal of Information Science, Vol 34 No 1, pp. 15-29, May 2007.

[5] Kim, Hak-Lae, John G. Breslin, Sung-Kwon Yang, and Hong-Gee Kim. "Social semantic cloud of tag: Semantic model for social tagging." In Agent and Multi-Agent Systems: Technologies and Applications, pp. 83-92. Springer Berlin Heidelberg, 2008.

[6] Knautz, Kathrin, Simone Soubusta, and Wolfgang G. Stock. "Tag clusters as information retrieval interfaces." In System Sciences (HICSS), 2010 43rd Hawaii International Conference on, pp. 1-10. IEEE, 2010.

[7] Kuo, Byron YL, Thomas Hentrich, Benjamin M. Good, and Mark D. Wilkinson. "Tag clouds for summarizing web search results." In Proceedings of the 16th international conference on World Wide Web, pp. 1203-1204. ACM, 2007.

[8] M.A. Hearst, and D. Rosner, “Tag clouds: Data analysis tool or social signaller?“, Proceedings of 41st Hawaii International Conference on System Sciences (HICSS 2008), Social Spaces minitrack, 2008.

[9] Satoshi Niwa, Takuo Doi, and Shinichi Honiden, “Web Page Recommender System based on Folksonomy Mining” in Proc. of the Third International Conference on Inforrmation Technology: New Generation (ITNG’06), pp. 388-393, April 2006.

[10] Seifert, Christin, Barbara Kump, Wolfgang Kienreich, Gisela Granitzer, and Michael Granitzer. "On the beauty and usability of tag clouds." In Information Visualisation, 2008. IV'08. 12th International Conference, pp. 17-25. IEEE, 2008.

[11] Sohil D Pandya, Dr. Paresh V Virparia “Clustering Approach in Context Free Data Cleaning”, in National Journal of System and Information Technology, Vol 2 No 1, pp. 83-90, June 2009.

[12] Sohil D. Pandya, Paresh V. Virparia and Rinku Chavda “Implementation of Folksonomy based Tag Cloud Model for Information Retrieval from Document Repository in an Indian University”, International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.5, No.1, February 2016

[13] W Cohen, P Ravikumar, S Fienberg “A Comparison of String Distance Metrics for Name-Matching Tasks” in Proc. of the IJCAI-2003

[14] T. Russell, “Cloudalicious: Folksonomy over time“, Proceedings of the 6th ACM/IEEEC-CS Joint Conference on Digital Libraries, 2006, p.364.

[15] Trant, Jennifer. "Studying social tagging and folksonomy: A review and framework." Journal of Digital Information 10, no. 1, 2009.

[16] Rinku Chavda, Sohil Pandya, “Improved Tag Cloud by Cleaning Tags using Context Free Data Cleaning”

[17] https://delicious.com/developers

[18] https://github.com/lucaong/jQCloud

[19] http://jquery.com

[20] http://www.w3schools.com/jquery

[21] http://en.wikipedia.org