Metrics for Tag Cloud Evaluation

Úrsula Torres-Parejo[1(✉)] , Jesús R. Campaña[2] , Maria-Amparo Vila[2] ,
and Miguel Delgado[2]

[1] Department of Statistics and Operational Research,
University of Cádiz, Cádiz, Spain
ursula.torres@uca.es
[2] Department of Computer Science and Artificial Intelligence,
University of Granada, Granada, Spain

Abstract. Since their appearance Tag Clouds are widely used tools in Internet. The main purposes of these textual visualizations are information retrieval, content representation and browsing of text. Despite their widespread use and the large number of research that has been carried out on them, the main metrics available in the literature evaluate the quality of the tag cloud based only on the query results. There are no adequate metrics when the tag cloud is extracted from text and used to represent information content. In this work, three new metrics are proposed for the evaluation of tag clouds when their main function is to represent information content: coverage, overlap and disparity, as well as a fourth metric: the balance, in which we propose a way to calculate it by using OWA operators.

Keywords: Tag cloud · Content representation · Metrics · Coverage
Overlap · Balance · OWA operators

1 Introduction

In [6–8] we widely studied tag clouds and proposed a standard method to generate them. These works were focused on the extraction of valuable information stored in textual databases. The main aim was to represent this information to users with no previous knowledge of a textual database content.

With this objective in mind, we established a complete methodology for text processing. This methodology includes the tasks of syntactic and semantic preprocessing, generation of an intermediate form, postprocessing and visualization through a tag cloud. The novelty of the proposal was the preservation of text semantics, due to the fact that related terms could remain together in the visualization, that is, the tag cloud generated was a multi-term tag cloud.

To evaluate the tag cloud obtained through this methodology when used for text retrieval, we used the precision, recall and $F1$ Score metrics [2]. But, to evaluate it as a tool of content representation, we did not find adequate metrics in the literature, thus modifying some of those proposed in [9] to assess the tag cloud obtained from the query results.

In this work, we establish a formal definition for the modified metrics, "coverage" and "overlap", and propose a new one, the "disparity". The last one fixes the inconveniences found in other existing metrics such as balance or entropy. We expose these inconveniences in the next section. In addition, we give a formal definition for the balance, establishing a new way of calculating it through OWA operators.

This paper is organized as follows: Sect. 2 presents a brief summary of the existing metrics to evaluate the tag cloud, exposing their inconveniences for evaluation of content representation in tag clouds. Section 3 proposes new metrics and illustrates the proposal with an example. Finally, Sect. 4 gives some conclusions and future work.

2   Existing Metrics

There are not a lot of metrics in literature to evaluate the goodness of a tag cloud and much less when it works as a tool for content representation.

We found the first in [1]. The authors define the **entropy** of a tag cloud as follows:

Let $t \in T$ be a tag in a tag cloud T:

$$Entropy(T) = -\sum_{t \in T} p(t) log\{p(t)\}$$

where

$$p(t) = \frac{weight(t)}{\sum_{t \in T} weight(t)}$$

Entropy quantifies the weight disparity between tags. If it is low, the tag cloud is significant or effective. If, on the contrary, it is high, the weights of the tags are uniform, which visually is not very informative. A tag cloud will be effective if it consists of significant tags.

The inconvenience of this metric is that it is unbounded, so it is difficult to know when its value can be considered high or low.

A set of metrics to capture the structural properties of the tag cloud generated from the query results is defined in [3,9]. These metrics are:

1. **Coverage.** Gives the fraction of the query set C_q covered by the tag cloud S. This metric takes values between 0 and 1. If it is close to 0, the tag cloud covers few objects in the query set, but if it is close to 1, it covers many objects.
2. **Overlap.** Different tags in S may be associated with the same objects in C_q. With this metric the extension of such redundancies is evaluated. This metric also takes values in the interval [0, 1]. If it is close to 0, there is little overlap. If, on the contrary, the value of this metric is close to 1, the overlap is high and the tags are not very different from each other.
3. **Cohesiveness.** Measures the closeness of the objects in each query set associated to each tag in the tag cloud, according to the relationships between these objects.

4. **Relevance.** It is defined as the overlap between the set of the results obtained with the query (C_q) and the set of the objects retrieved with each tag (C_t). It is calculated as the fraction of results in C_t that are also in C_q.
5. **Popularity.** A tag in S is popular in C_q if it is associated with many objects in C_q.
6. **Independence.** Two tags in S are independent if the objects they recover are not similar to each other. The metric in this case is similar to the cohesiveness, but the latter is calculated for each pair of tag sets.
7. **Balance.** A tag cloud S is balanced if its tags represent a similar number of objects in C_q. The balance takes values in the range $[0, 1]$. A tag cloud is considered to be balanced if the value of this metric is close to 1. This metric is calculated as the fraction between the minimum and the maximum number of objects retrieved through a tag, so only two values are considered for its calculation.

From these metrics, only the coverage, overlap, and balance are suitable for its adaptation to evaluate the tag cloud generated with the purpose of representing text content. The other metrics are only useful for tag clouds coming from query sets or if they are calculated for one or two isolated tags.

Furthermore, the balance metric has the inconvenience that for its calculation, only two tags are considered, the tag which more objects represents and the one representing less. A tag cloud is said to be unbalanced if this metric is close to 0, but if we think for example in a tag cloud with all tags representing the same number of objects except for one tag which differs so much from the others (see tag cloud in Fig. 1 where *tag8* is much smaller than the others), we could say that this tag cloud is unbalanced when it is not. For this reason, this metric does not seem appropriate to us.

Fig. 1. Example of balanced tag cloud said to be unbalanced according to [9]

In [5] the metric **Selectivity** is also proposed, which measures the number of objects filtered in a tag cloud when a tag that has no relation with the former tag is selected. Other metrics such as **Simplicity** or **Detailedness** can be found in [4] to evaluate the cluster grouping into the tag clouds.

3   Metrics Proposed for the Evaluation of the Tag Cloud as a Tool of Content Representation in Textual Databases

In [7,8] we defined a methodology of text processing in databases that had as final step the visualization of the text through a tag cloud. This representation helps in content identification and in querying and browsing tasks, for textual data.

To evaluate the goodness of the information retrieved through the tags, we used the precision, recall and F-Score metrics [2], which are very standardized measures for these purposes. But, for evaluating the tag cloud as a content representation tool, we find few specific metrics that do not meet our requirements.

We take the metrics "coverage" and "overlap" from [9] and adapt them to evaluate the tag cloud as a tool of content representation. The metric "balance" has the inconvenience of taking only two values for its calculation, the minimum and maximum weights, being a metric very influenced by the extreme values, which could lead us to erroneous conclusions.

A way to calculate the balance avoiding these abnormal values in the weight is excluding first the outliers. Values that exceed 1.5 times the interquartile range are considered to be outliers. In this way the influence of values due to errors in the data would be avoided, but still only two values would be considered for the calculation of the balance.

Next, we define three metrics to evaluate the tag cloud when it is used as a tool for content representation in textual databases: the coverage, the overlap and the disparity. The third one is related with the balance and the entropy, but is bounded and considers all the tags for its calculation. In addition, a new way of calculating the balance through OWA operators is proposed, avoiding first the outliers as we exposed above.

3.1   Coverage, Overlap and Disparity

Let x_i be a tag of a cloud X and t_i be a tuple in a set of tuples T, then we call $T(x_i)$ to the set of tuples associated to a tag t_i.

We calculate the coverage of X as:

$$cov(X) = \frac{card(\cup_{x_i \in X} T(x_i))}{card(T)} \qquad (1)$$

This metric takes values in the interval [0, 1]. A value close to 1 indicates that the tag cloud represents most of the content of the database.

The overlap is calculated through the expression:

$$over(X) = avg_{i \neq j} \left(\frac{card(T(x_i) \cap T(x_j))}{min\{card(T(x_i)), card(T(x_j))\}} \right) \qquad (2)$$

This metric also takes values between 0 and 1. A value close to 0 means that tags in the tag cloud represent different objects in the database.

And finally, we calculate the disparity with the next expression:

$$dis(X) = avg_{i \neq j} \left(\frac{|card(T(x_i)) - card(T(x_j))|}{max\{card(T(x_i)), card(T(x_j))\}} \right) \qquad (3)$$

As the two other metrics, the disparity takes values between 0 and 1. A value close to 1 indicates that the disparity between the weights of the tags is high, which is necessary in order to highlight the importance of the tags through the weight.

3.2   Balance Calculated Through OWA Operators

Ordered Weighted Averaging (OWA) Operators. An OWA operator of dimension n is a mapping [11]:

$$f : R^n \rightarrow R$$

that has an associated n vector W:

$$W = [w_1 \ \ w_2 \ \ w_3]^T$$

such that

$$1 \, . \, w_i \in [0, 1]$$
$$2 \, . \, \sum_i w_i = 1$$

Furthermore $f(a_1, \ldots, a_n) = \sum_j w_j b_j$ where b_j is the jth largest of the a_i.

An important aspect to consider in this operation is the re-ordering step, since a weight is associated with a particular ordered position of aggregate, that is, the $a_1, \ldots, a_n$ vector has to be previously ordered in a descending way.

In [10] Yager pointed out three important special cases of OWA aggregations:

1. F^*: In this case $W = W^* = [1 \ 0 \ \ldots \ 0]^T$. Then $F^*(a_1, \ldots, a_n) = Max_i(a_i)$,
2. F_*: In this case $W = W_* = [0 \ 0 \ \ldots \ 1]^T$. Then $F_*(a_1, \ldots, a_n) = Min_i(a_i)$,
3. F_{Ave}: In this case $W = W_{Ave} = [1/n \ \ldots \ 1/n]^T$. Then $F_{Ave}(a_1, \ldots, a_n) = \frac{1}{n} \sum_i a_i$.

Balance Calculated Through OWA Operators. If we consider the two first special cases of OWA operators pointed out in [10], we can express the balance as:

$$bal(S) = \frac{F_*(C)}{F^*(C)} \qquad (4)$$

where F_* and F^* are OWA operators of size n with associating weighting vectors $W_* = [0 \ 0 \ \ldots 1]^T$ and $W^* = [1 \ 0 \ \ldots 0]^T$, respectively and $C = \{card(T(x_1)), card(T(x_2)), \ldots, card(T(x_n))\}$ is the set of cardinals from the associated sets $T(x_i)$ to each tuple t_i.

Before applying this formula we have to exclude the outliers from C, identifying these as the values that exceed 1.5 times the interquartile range.

The balance takes values in the interval [0, 1]. A value close to 0 indicates that the tag cloud is considered to be unbalanced.

3.3   Example of Calculation of Coverage, Overlap, Disparity and Balance Through OWA Operators

Table 1 presents several movie titles selected from FilmAffinity[1] in the category of romance.

**Table 1.** Titles of movies selected from FilmAffinity

	Movie titles		Movie titles
1	A true love story never ends	11	True love
2	This is not your story	12	Love and air sex
3	True romance	13	Holiday affair
4	Shakespeare in love	14	A holiday to remember
5	Love and other cults	15	Love on the air
6	A holiday for love	16	My true love story
7	Love actually	17	Our story needs no filter
8	West side story	18	Love story
9	A taste of romance	19	Love is in the air
10	The history of love	20	Roman holiday

After cleaning the text in Table 1 and following the methodology explained in [8], two tag clouds have been generated considering different supports in terms of absolute frequency. We can see them in Fig. 2.

For both tag clouds, metrics proposed in Subsects. 3.1 and 3.2 have been calculated. We can see the values obtained in Table 2.

The details of the calculation of balance through OWA operators is in Table 3, where B is the ordered vector of items in C.

As we can see, Tag Cloud 1 performs better in coverage and disparity, but worst in overlap. The Tag Cloud 1 has smaller support than Tag Cloud 2, so a

Tag Cloud 1 (Support = 2)   Tag Cloud 2 (Support = 3)
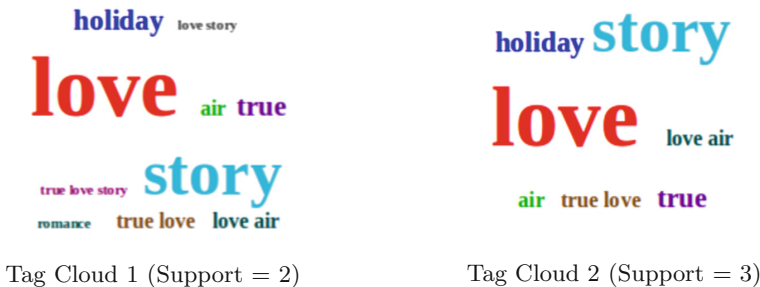
Fig. 2. Tag clouds for text in Table 1

Table 2. Coverage, overlap and disparity for tag clouds in Fig. 2

Metrics	Tag Cloud 1	Tag Cloud 2
Coverage	1	0.95
Overlap	0.393	0.365
Disparity	0.406	0.369
Balance	0.5	0.75

Table 3. Details of the balance calculation

	Tag Cloud 1	Tag Cloud 2
C	{2, 3, 4}	{3, 4}
Outliers	None	None
B	[4 3 2]	[4 3]
W_*	$[0\ 0\ 1]^T$	$[0\ 1]^T$
W^*	$[1\ 0\ 0]^T$	$[1\ 0]^T$
$F_*(C)$	$[4\ 3\ 2] * [0\ 0\ 1]^T = 2$	$[4\ 3] * [0\ 1]^T = 3$
$F^*(C)$	$[4\ 3\ 2] * [1\ 0\ 0]^T = 4$	$[4\ 3] * [1\ 0]^T = 4$
Balance	$2/4 = 0.5$	$3/4 = 0.75$

greater number of tags appears in its visualization, increasing the coverage. The bigger number of tags causes redundant information to be represented, so the overlap is increased as well. The disparity is also increased with the appearance of tags with smaller weights and the tag cloud is more unbalanced. Apparently, disparity is a softer measure than balance, due to the fact that it considers all the tags for its calculation and the balance only two.

The choice of one or another tag cloud will depend on the preferences in each case.

4   Conclusions

In order to evaluate the tag cloud as a content representation tool, three metrics have been proposed: coverage, overlap and disparity. The three take values in the interval $[0, 1]$ and use all the tags in the visualization for its calculation. In addition, a new way to calculate balance through OWA operators has been proposed.

Coverage and overlap are usually in confrontation with each other. When the coverage increases, the overlap also increases as more tags appear in the cloud, representing a large amount of information which brings more redundancies. Disparity and balance can also be affected with the variations in coverage and overlap. Disparity seems to be a softer measure than balance. The proportion between these values will be established according to the requirements in each case.

As future work we plan to continue researching with tag clouds generated over textual databases, the inclusion of fuzzy logic in tags is considered, their grouping in clusters, the adoption of OWA operators for tag cloud comparison, as well as the study of the properties of said operators. We intend also to use the tag cloud for searching entities and the introduction of other languages to create a multilingual tool based on ontologies.

References

1. Aouiche, K., Lemire, D., Godin, R.: Web 2.0 OLAP: from data cubes to tag clouds. In: Cordeiro, J., Hammoudi, S., Filipe, J. (eds.) WEBIST 2008. LNBIP, vol. 18, pp. 51–64. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01344-7_5

2. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: Losada, D.E., Fernández-Luna, J.M. (eds.) ECIR 2005. LNCS, vol. 3408, pp. 345–359. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-31865-1_25

3. Leone, S., Geel, M., Müller, C., Norrie, M.C.: Exploiting tag clouds for database browsing and querying. In: Soffer, P., Proper, E. (eds.) CAiSE Forum 2010. LNBIP, vol. 72, pp. 15–28. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-17722-4_2

4. Morik, K., Kaspari, A., Wurst, M., Skirzynski, M.: Multi-objective frequent termset clustering. Knowl. Inf. Syst. **30**(3), 715–738 (2012)

5. Skoutas, D., Alrifai, M.: Tag clouds revisited. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM, pp. 221–230 (2011)

6. Torres-Parejo, U., Campaña, J.R., Vila, M.-A., Delgado, M.: Text retrieval and visualization in databases using tag clouds. In: Greco, S., Bouchon-Meunier, B., Coletti, G., Fedrizzi, M., Matarazzo, B., Yager, R.R. (eds.) IPMU 2012. CCIS, vol. 297, pp. 390–399. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31709-5_40

7. Torres-Parejo, U., Campaña, J., Delgado, M., Vila, M.: MTCIR: a multi-term tag cloud information retrieval system. Expert Syst. Appl. **40**, 5448–5455 (2013)

8. Torres-Parejo, U., Campaña, J., Vila, M., Delgado, M.: A theoretical model for the automatic generation of tag clouds. Knowl. Inf. Syst. **40**(2), 315–347 (2014)

9. Venetis, P., Koutrika, G., Garcia-Molina, H.: On the selection of tags for tag clouds. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM, pp. 835–844 (2011)

10. Yager, R.: On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Trans. Syst. Man Cybern. **18**(1), 183–190 (1988)

11. Yager, R.: Families of OWA operators. Fuzzy Sets Syst. **59**(2), 125–148 (1993)