

A flexible and lightweight interactive data mining tool to visualize and analyze digital citizen participation content

Sergio Bachiller
s.bachillerrubia@gmail.com
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, Spain

Lara Quijano-Sánchez*
lara.quijano@uam.es
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, Spain

Iván Cantador
ivan.cantador@uam.es
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, Spain

ABSTRACT

Addressing information overload in current e-participation platforms, we present a lightweight web application consisting of a simple HTML-based data panel that, through the use of date, location and category based filters, and several interactive graphs, allows visualizing, exploring and analyzing data obtained from public deliberation platforms in an easy and clear way. The tool, which implements natural language processing, text similarity, and graph clustering techniques to group citizen proposals, may serve as a decision support system for the municipal government, and may contribute to increase the citizens' participation and engagement.

CCS CONCEPTS

• **Applied computing** → **Document management and text processing**; • **Human-centered computing** → **Visualization**; • **Information systems** → **Data mining**.

KEYWORDS

citizen participation, data visualization, data clustering, text similarity, civic technologies

ACM Reference Format:

Sergio Bachiller, Lara Quijano-Sánchez, and Iván Cantador. 2021. A flexible and lightweight interactive data mining tool to visualize and analyze digital citizen participation content. In *Proceedings of ACM SAC Conference (SAC'21)*. ACM, New York, NY, USA, Article 4, 4 pages. <https://doi.org/https://doi.org/10.1145/3412841.3442081>

1 INTRODUCTION

The huge, ever increasing citizen generated content leads to an information overload problem for both citizens and government stakeholders in decision and policy making tasks. In addition to being overwhelmed by large amounts of data, whose exploration and understanding result challenging and frustrating, citizens could feel thwarted if their proposals do not reach sufficient visibility and

impact. In this sense, we can find proposals by different authors that address the same problem, but in different ways, for distinct city locations, or entail distinct initiatives and potential solutions. To address these problems, there is a need for information systems capable to process and mine citizen generated content, as well as to summarize, visualize and analyze relevant extracted information overtime. Motivated by this issue, we propose a flexible, lightweight data mining tool that helps unraveling public deliberation contents to both governments and citizens. The addressed goal is the development of systems and strategies that enable people to contribute to policies with an impact on their communities and own lives [7]. In this context, our tool makes these activities simpler by interactively analyzing knowledge generated from public initiatives, e.g., detecting trends in the concerns posed to policymakers, and identifying persistent demands or particular seasonal problems. The presented tool allows unifying objectives by grouping and visually monitoring proposals (that could have been accepted or remain unanswered), thus serving as a stimulus for citizens to increase quality (i.e., more informed and argued proposals) and quantity (more proposals due to ease of use) on their generated content. Our tool provides interactive mechanisms for citizen participation data visualization and analysis, and is built upon the *Tableau* data visualization software. It is lightweight and easy to configure, as well as generic, since it is reusable for other related domains and different languages as it uses data from an external database. The tool, whose code is publicly available, provides several visualization functionalities, allowing both temporal and geographical analysis by means of different diagram bars, heat maps, and time series graphs. Implementing natural language processing, text similarity, and graph theory techniques, groups of proposals related to common topics of interest are created and visualized. The performed clustering allows a better and easier extraction of patterns and insights when analyzing the published citizen generated content. Moreover, the tool has a co-production functionality based on the retrieval of existing similar proposals. Hence, a citizen who is interested in submitting a new proposal can first bring it into the tool and check if there are related ones. As a case study, we instantiated the tool with Open Government Data from Decide Madrid, the e-participative budgeting platform of the Madrid City Council.

2 CASE STUDY: DECIDE MADRID PLATFORM

Decide Madrid, active since September 2015 with 420,000 registered users, is the online platform where the Madrid City Council

*Corresponding author.

"Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)."

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SAC'21, March 22-March 26, 2021, Gwangju, South Korea

© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8104-8/21/03...\$15.00
<https://doi.org/https://doi.org/10.1145/3412841.3442081>

<https://www.tableau.com>

<https://github.com/sbachiller/EParticipationAnalysis-DecideMadrid>

<https://decide.madrid.es>

orchestrates the city's annual participatory budgets. In the platform, a series of initiatives and projects are proposed, discussed and supported by residents. All the content generated in the platform, i.e., proposals text and metadata, comments and votes are publicly available as open data. Throughout a year, Decide Madrid allows proposals (more than 27K as far as December 2020) to be created, discussed and supported. Once a proposal obtains the necessary support (i.e., 1% of the population in the city), it is submitted for citizen vote. Those supported proposals that achieve a simple majority of votes are subject to be implemented by Madrid City Council. Part of the municipal budget for this means is allocated.

3 MINING TOPICS OF INTEREST

Our aim is to develop a tool capable to extract from e-platforms rich information about particular interests and problems in city neighborhoods and districts, and issues that represent general concerns of majorities (minorities). In this way, it would be possible to better outline for both citizens who use a platform and local governments what inhabitants are asking for, facilitating the work of decision and policy makers, and ultimately leading to a improvement on the citizens' quality of life.

Having as input documents with the title, abstract and text of citizen proposals, the tool performs the following tasks: i) the content of the proposals is transformed using natural language processing (NLP) techniques, ii) text similarities are computed over the processed documents and used to build a document relatedness graph, and iii) a graph-based clustering is applied to group duplicate and/or similar proposals.

The final output of these tasks consists of citizen proposal clusters that are analyzed to identify, among other issues, general topics of interest. Hence, the tool not only allows visualizing the raw data contained in e-platforms, but summarize and facilitate the understanding of underlying problems and proposed solutions.

Next, we present the text processing techniques carried out prior to visualizing information by the developed tool.

3.1 Text processing

In order to ensure that the computation of document similarity is accurate, it is necessary to treat the textual content appropriately. We accomplish this task using common tools in a NLP pipeline. A first step of text pre-processing is mistake correction. Special characters are first removed, thus avoiding possible problems in the misspelling correction. Once the texts have been corrected, two more pre-processing tasks are performed: i) *Stopwords* removal and extraction of nouns, adjectives and verbs that are valuable to find relevant similarities. ii) *Lemmatization*.

3.2 Document similarity

Estimating the similarity between two texts is an extensively studied task in the NLP field [6, 11]. In this work, we aim to find a similarity measure with two main characteristics, namely lexical and semantic similarity, that is, the words appearing in the texts are from the same context and have the same meaning, respectively.

In [12, 13], the main trends, examples, limitations and successes of the most popular methods of text similarity are reported. In order to use a method that captures both lexical and semantic similarities,

in this work we advocate for the *Word Mover's Distance* (WMD) similarity [8], since it stands out over the simplest methods and at the same time does not require a pre-labeled dataset to be executed, facilitating thus its reusability of the developed tool for different domains and languages.

The WMD similarity is inspired by the *Earth Mover Distance* transportation problem, aiming to find similarity (distance) between two texts even if they have no words in common. WMD leverages the results of advanced embedding techniques like word2vec and Glove [5]. It treats text documents as weighted point clouds of embedded words. The distance between two text documents A and B is calculated by the minimum cumulative distance that words from the text document A need to travel to match exactly the point cloud of text document B. Hence, the distance measures the dissimilarity between two text documents as the minimum amount of distance that one document's embedded words need to "travel" for reaching another document's embedded words. This measure computes distance and not similarity. For this reason, all the values of the distance matrix WMD of dimension $N \times N$ (where N denotes the number of documents) and the maximum distance are used to compute a similarity matrix as follows: $Similarity(i, j) = 1 - WMD(i, j) / max_distance$.

3.3 Document clustering

Instead of using classic clustering techniques such as K-Means and agglomerative clustering, we propose to use recent approaches applied to detecting communities of interest in urban contexts [1, 4]. For such purpose, we build a non directed, weighted graph whose nodes represent the citizen proposal documents and whose edges are assigned with the computed document similarity values. On the built graph, we apply the Louvain method [2], which locally optimizes the modularity of the graph and associates nodes until convergence, with a good execution time of $O(n \cdot \log(n))$. This clustering method, in contrast to others like K-Means, it does not need a fixed number of clusters, but rather it adapts to the problem.

To apply the algorithm on our graph, we removed edges with weights lower than certain value (representing the level of desired similarity within the community). Specifically, all edges with weights lower than $min_weight = 0.55$ (motivated by the range in which considerable similarities were observed) were removed. Some results are shown in Table 1. Among the formed communities we verified the citizens' concerns in the case study platform, such as '*An option of NO SUPPORT in Madrid Decide*' or '*Group similar proposals in Decide Madrid*'. Finding these proposals in such a large community verifies our hypotheses and reinforces the indications of the referred preliminary studies, summarized in [10]. We found documents similar to each other, thus achieving the desired unification and summarization objectives when it comes to understanding the large volume of proposals.

4 CITIZEN PARTICIPATION ANALYSIS TOOL

The Madrid City Council, through its open data portal, provides data collections related to the city. Among these collections, we focus on the 21,746 citizen proposals created until September 2019

We have employed word2vec on a corpus in Spanish <https://github.com/dccuchile/spanish-word-embeddings#word2vec-embeddings-from-sbwc>
<https://datos.madrid.es/portal/site/egob>

Community	Main category	Proposal examples
Blue parking zone (SER)	Mobility	Flexible allocation of SER areas for residents
		Delete reserved areas in official buildings
		More blue parking areas for people with reduced mobility
Dog poop	Animals	Significant penalties for dog owners
		Fines to people who deposit garbage on the street
		Fine dog owners for not removing their feces
Clean city	Environment	Clean the streets
		Increase cleaning in Madrid
		More trees and clean air in Madrid

Table 1: Representative communities along with their main category and some examples formed by our method.

and their associated 86,102 comments. Each proposal has a title, a summary, a description, social tags, publication date and the number of received supports. We automatically tagged each proposal with one or several of 30 existing general categories. Downloaded data also include a geographic repository with almost 1,500 streets and POIs of Madrid with their corresponding districts and neighborhoods. Using this repository, proposals were assigned to its corresponding location, i.e., a street, a neighbour (among the existing 129), a district (among the existing 21) or the whole city. The retrieved data were automatically extended with topics that refine the assigned categories (among 325 different options), and measures of popularity and controversy following [3]. We note that for reusability purposes, if the data structure is maintained, other use cases with similar scopes or in alternative languages could be loaded into the tool. Furthermore, the data filters of the tool could be easily adjusted to the characteristics of a new given database. Also, given that the techniques used for text similarity and clustering described in Section 3 are not based on particular topics, languages and corpora, the tool is extremely flexible.

4.1 Data analysis functionalities

Our tool presents a number of visual components that show distinct aspects of the database: categories (and topics), communities, districts (and neighborhoods), temporal evolution and influence distribution. Each component has several filters –date, category, topic, district, neighbourhood, community and number of proposals– that allow the user to interactively explore the information. Proposals are also displayed in real time as filters adjusted in the interface. The interface leads to two types of analysis: temporal and geographical. As for the temporal analysis, by means of bar charts, users can analyze the number of proposals presented in each district over time. This information is grouped by months and slots of trimesters per year. Users can further unravel information by selecting concrete districts or time periods. In the temporal evolution components, users can study the trends of detected topics of interest (i.e., the communities described in Section 3.3). To do so, the topics are presented as a series of bars for the months in the year, grouped by trimester. The height of the bars represent the number of proposals belonging to the communities for each month. Users can also filter by the number of communities displayed vertically, a chosen time span, and involved districts. An example of this analysis is exemplified in Figure 1. The counterpart graph is presented; in this case, analyzing the evolution of proposals belonging to the categories.

Regarding the geographical analysis, similarly to the participation by district temporal graph displayed in the participation component, bar charts representing participation by district are presented. Again, filters for desired span of time and districts to

study are available. Also to facilitate visualization, the tool presents a topographical map of Madrid divided by districts, where the influence of communities and categories for each district are highlighted (darker colours representing high production volume). This component, presented later in Figure 1, aids users to study possible geographical correlations and phenomena, as districts that are actually close are painted according to their GPS coordinates.

To assist users in information exploration, three general components are introduced with the aim of identifying proposals with desired characteristics through a chain of filters. The goal of these components is changing the filter order and the type of graphics that adapt to the specific temporal or geographical analysis. The first one, finding through category, allows users to first select a desired time span, then select categories presented in horizontal bar charts, and set filters by topics, districts and neighbours. We note that inside each filter, the number of proposals to be shown can also be established. In the last graph, a temporal analysis of the retrieved proposals is presented. The second component, finding through category, is analogous to the previous, exchanging categories by identified communities. In the last component, finding through district, after selecting a desired time span, the volume of proposals in districts is represented with the corresponding bigger or smaller bubbles. This selection allows for further tuning the search indicating neighbours to visualize inside the prior selection, and then categories and topics related to the chained filters. Again, a final temporal analysis graph of the selected information is presented. Lastly, the tool allows, given certain (new) proposal, finding similar proposals, which enable avoiding duplication and merging of proposals from the platform. Space limitation prevents the authors from showing more functionalities as screenshot figures.

5 ANALYSIS INSIGHTS

In this section, we analyze the influence of the topics of interests identified in the different districts of the city of Madrid, and their evolution over time. An in-depth social study could be carried out, in which socio-demographic variables would be used to know and understand the motivations underlying the proposals [3]. In particular, we next present a sample of insights hidden within the large volume of data, which were easily identified thanks to the data filtering and visualization functionalities of our tool.

As mentioned before, we can analyze the participation by district and its evolution, where for instance districts with the highest incomes tend to be those that have participated the least, and districts with the lowest incomes are those that present more proposals. Through the temporal visualization component (Figure 1), a deeper analysis of the evolution of participation is possible by adding two new variables: the community (at the top) or the category (at the bottom). In both cases, we can see a significant decrease on the number of proposals during the summer months. With these bar graphs, it is possible to estimate the seasonality of a problem over time, for example, in the *Animals* category, which increases considerably during the last months of the year, surely caused by the abandonment of pets or animals that are given away during Christmas. It is also possible to find a certain seasonality in the evolution of the communities and categories. Regarding the community that *Scope of public transport* represents, we can see that it is in the first trimester of each year (Figure 1 top), when it has the highest

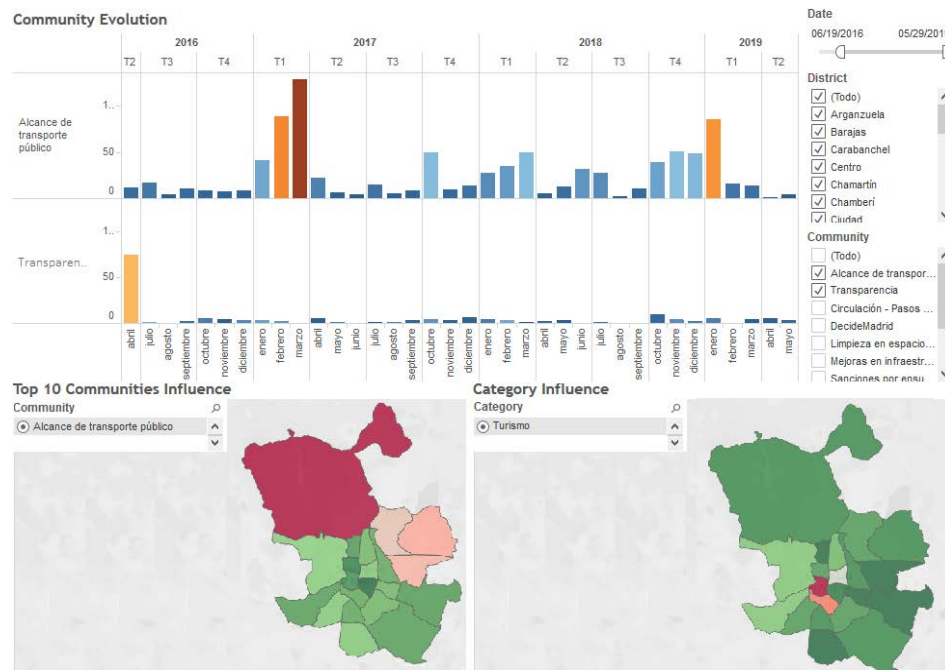


Figure 1: Screenshots of some of the visualization and analysis components of the tool.

impact, probably with the goal of being included in the objectives of the organizations responsible for that year. Thanks to this data panel, it is easy to discover the seasonality of groups of proposals, and even to detect unusual peaks, such as the one shown in Figure 1 (in the middle) for the transparency community in April 2016. In that month, the General Director of Economy of Madrid City Council presented her resignation, after documents that linked her to a illegal company in Panama were leaked.

Figure 1 (bottom), which shows the thematic influence panel of our tool, we can observe the impact of categories and communities on each district and area in the city. Specifically, in the map on the left, we can analyze the impact of the selected community, *Scope of public transport*, and find out that there are clear needs in the periphery districts. In the map on the right, we can see that in the *Tourism* category, the majority of proposals come from downtown districts, where the number of places of interest is much higher.

6 CONCLUSIONS

Nowadays, cities are implementing online platforms for citizen participation where inhabitants participate in municipal decisions and actions by means of proposals and debates. To date, these platforms are suffering problems related to the citizens' frustration for the lack of visibility and impact of their proposals, and to a low participation due to information overload and exploration difficulties. To address these problems, researchers have proposed the development of technological solutions to summarize and contextualize citizen feedback, and visualize individual and community needs and concerns [3, 9]. Following this direction, we have presented a flexible interactive tool that provides a variety of visualization and analysis functionalities for citizen generated content, and facilitates to both citizens and local governments the understanding of the underlying problems in a city and proposed solutions.

Acknowledgements This work was conducted with financial support from the Spanish Ministry of Science and Innovation (PID2019-108965GB-I00) and the Centre of Andalusian Studies (PR137/19).

REFERENCES

- [1] Monira N Aldelaimi, M Anwar Hossain, and Mohammed F Alhamid. 2020. Building dynamic communities of interest for Internet of Things in smart cities. *Sensors* 20, 10 (2020), 2986.
- [2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (2008), P10008.
- [3] Iván Cantador, María E Cortés-Cediel, and Miriam Fernández. 2020. Exploiting Open Data to analyze discussion and controversy in online citizen participation. *Information Processing & Management* 57, 5 (2020), 102301.
- [4] Nigel Francisus, Xuguang Ren, Junhu Wang, and Bela Stantic. 2019. Word Mover's Distance for Agglomerative Short Text Clustering. In *Proceedings of the 2019 Conference on Intelligent Information and Database Systems*. 128–139.
- [5] E Hindocha, V Yazhini, A Arunkumar, and P Boobalan. 2019. Short-text Semantic Similarity using GloVe word embedding. *International Research Journal of Engineering and Technology* 6, 4 (2019).
- [6] R Ibrahim, S Zeebaree, and K Jacksi. 2019. Survey on semantic similarity based on Document Clustering. *Advances in Science and Technology: Research Journal* 4, 5 (2019), 115–122.
- [7] David Rios Insua, Gregory E Kersten, Jesus Rios, and Carlos Grima. 2008. Towards decision support for participatory democracy. In *Handbook on Decision Support Systems* 2. 651–685.
- [8] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd Conference on Machine Learning*. 957–966.
- [9] Amal Marzouki, F Lafrance, Sylvie Daniel, and Sehl Mellouli. 2017. The relevance of geovisualization in citizen participation processes. In *Proceedings of the 18th Annual International Conference on Digital Government Research*. 397–406.
- [10] Directorate-General For Internal Policies. 2012. Potential And Challenges of E-Participation In The European Union. (2012).
- [11] Omid Shahmirzadi, Adam Lugowski, and Kenneth Younge. 2019. Text similarity in vector space models: a comparative study. In *Proceedings of the 18th International Conference On Machine Learning And Applications*. 659–666.
- [12] Adrien Sieg. 2018. Text Similarities: Estimate the degree of similarity between two texts. *Medium* 5 (2018).
- [13] Shuiqiao Yang, Guangyan Huang, Bahadorreza Ofoghi, and John Yearwood. 2020. Short text similarity measurement using context-aware weighted bitersms. *Concurrency and Computation: Practice and Experience* (2020), e5765.