Supporting the Exploration and Analysis of Public Participation Process Data with Rankings and Clusterings

Lars Schütz[1,2], Korinna Bade[1], Andreas Nürnberger[2]

1 Anhalt University of Applied Sciences, Department of Computer Science and Languages, 06366 Köthen
2 Otto von Guericke University Magdeburg, Faculty of Computer Science, 39106 Magdeburg

Abstract

Public participation processes enable the inclusion of diverse perspectives and allow people to engage in various real-life areas. These processes involve a wide range of data, with a major component being individual contributions that consist primarily of natural language text. Due to the complexity of public participation process data, exploring and analyzing it is highly challenging. To address this challenge, we propose two different approaches for facilitating the exploration and analysis of contributions. Specifically, we revisit the ranking of contributions to support their assessment, and we consider the clustering of contributions. We also describe two approaches that identify cluster representatives for explaining a clustering. Throughout our work, we take into account the perspectives of both public administrations and citizens.

1. Introduction

Public participation processes in the e-participation domain offer opportunities for people to engage in various aspects of real-life. This paper specifically focuses on public participation processes in local urban land-use planning and decision processes, such as planning new public green spaces in cities. These processes can be categorized broadly as formal and informal planning and decision processes. Formal processes are institutionalized activities that are carried out to achieve social objectives (Briassoulis, 1997). The primary planning instruments and results in these processes are mandatory plans that outline the desired future state of the planned area (Blotevogel et al., 2014). Informal processes, on the other hand, are not institutionalized and are not bound to pre-defined procedures or specific instruments. They focus on both public and private interests (Briassoulis, 1997). The primary goal of public participation processes is to ensure that the perspectives and needs of all affected individuals are taken into account when making decisions that could affect them. This approach can lead to more equitable and effective decision-making and can help build trust in the decisions made.

Public participation processes involve a diverse range of participants, including citizens, planners, moderators, public administrations, and IT service providers. In this work, we focus primarily on two major groups: public administrations and citizens. Each group has distinct tasks, such as citizens submitting their ideas and complaints, while public administrations assess these contributions, commonly accepting or rejecting them. An exemplary contribution is shown in Figure 1. However, both groups also share common

tasks at least on an abstract level, such as exploring and analyzing public participation process data.

> Low-ropes courses are a sporty and visual enrichment for young and old on a relatively small space, e.g., as a circle with trees, small children, handicapped people up to seniors can have fun. Double-track would be even better, because then you could build in different levels of difficulty. The...
>
> #128 – BWK18 – 19.8.2018, 10:37:40

Figure 1: An excerpt of an exemplary contribution

In the e-participation domain, public participation process data is characterized by heterogeneity in terms of data types, measurement levels, dimensions, and structural aspects. This data can include various types such as natural language text data (e. g., plan documents, forum posts, comments or text references), time-oriented data (e. g., timestamp information), spatial data (e. g., marker positions on geographic maps), images (e. g., email-attached camera photos) and popularity information (e. g., like status or ratings). Moreover, these data types can be interrelated, with one piece of data referring to multiple others, as in the case of a comment that refers to both a text document and a specific location on a map. This leads to the creation of a complex and constantly evolving network of interconnected data, which requires the consideration of structured, semi-structured, and unstructured data, as well as possible relationships between them. To illustrate this, we present exemplary public participation process data and some interconnections in Figure 2.

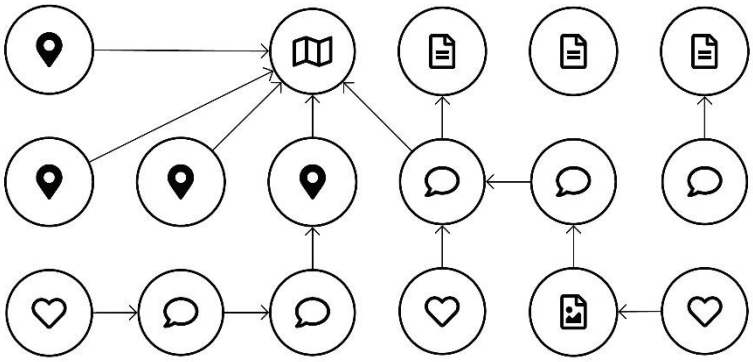

Figure 2: Exemplary planning process data. Contributions/comments (speech balloon), geographic maps, map markers, ratings (heart), images, and texts might relate to each other.

The complexity and diversity of public participation data present challenges for both public administrations and citizens in exploring and analyzing the data. For public administrations, maintaining consistency in their decisions when assessing contributions is crucial, i. e., similar contributions should be assessed similarly while contradicting contributions might lead to opposing decisions. This can be a time-consuming and laborious task, and grouping together contributions that address similar issues could be beneficial. The exploration of public participation process data is equally important, especially for citizens, who may want to know if their concerns have already been expressed. To address these challenges, we propose using rankings and clusterings to support the exploration and analysis of public participation process data. Both methods

can lead to different perceptions and interactions for public administrations and citizens. Therefore, it is crucial to ensure the interpretability of these methods to establish trust. We suggest that the proposed methods be designed to address the specific needs of both public administrations and citizens.

In this work, we first review related work (Section 2). Subsequently, we explore two approaches that support the exploration and analysis of public participation process data: the ranking of contributions (Section 3) and the clustering of contributions (Section 4). Finally, we present our conclusions and we discuss potential directions for future research in this area (Section 5).

2. Related work

Generally, the need for better tools and sophisticated methods for working with public participation process data has long been recognized (Schütz et al., 2015). Especially the practitioners confirm that advanced tools for the assessment of contributions would be helpful (Helbig et al., 2016). There is work that conceptualizes graphical user interfaces for working with public participation process data (Schütz et al., 2016). It considers a graph-based data structure for representing the relationships. This allows a guided exploration of the relationships when the graph's edges are followed. However, there is no description of special methods for the analysis of contributions. There is also an explanation of using the visual analytics (Wong & Thomas, 2004) approach for public participation process data (Schütz, Raabe, et al., 2017). Visual analytics simultaneously considers data analysis and information visualization methods for gaining insight into data. However, it is more or less a general view or framework. That is why we focus on explicit approaches in the remainder of this work. We also revisit previous work related to the ranking of contributions (Schütz & Bade, 2019).

3. Ranking of contributions

Ranking draws inspiration from the field of information retrieval, which refers to the process of ordering search results based on their relevance to a given query (Manning et al., 2008). For example, in the context of Internet search engines, a query typically consists of a sequence of terms in a specific natural language. Search results are then determined that match this query, and these results are ranked according to their similarities to the given query. When it comes to public participation processes, the ultimate goal of ranking contributions is to present the most relevant contributions to public administrations or citizens in order of relevance to a reference contribution that they can select. In this context, the reference contribution serves as the query. From the perspective of public administrations, this helps to ensure consistency in decision-making during the assessment process, which is crucial. A public administration worker can begin by selecting a contribution for assessment and then query a ranking of semantically similar contributions so that the most similar contributions can be assessed in a similar, if not exact, manner.

We developed and evaluated a contribution assessment system designed for use by public administrations (Schütz & Bade, 2019). The system includes a user interface with a two-column layout as shown in Figure 3. The first column contains contributions that still need to be assessed, while the second column presents a ranked list of contributions that are similar to a reference contribution selected in the first column. This allows public administration workers to assess similar contributions in one go, starting with a

selected reference contribution in the left column and continuing with the similar contributions in the second column. The system has been designed to promote consistency in decision-making, which we hypothesize will be achieved through this user interface.

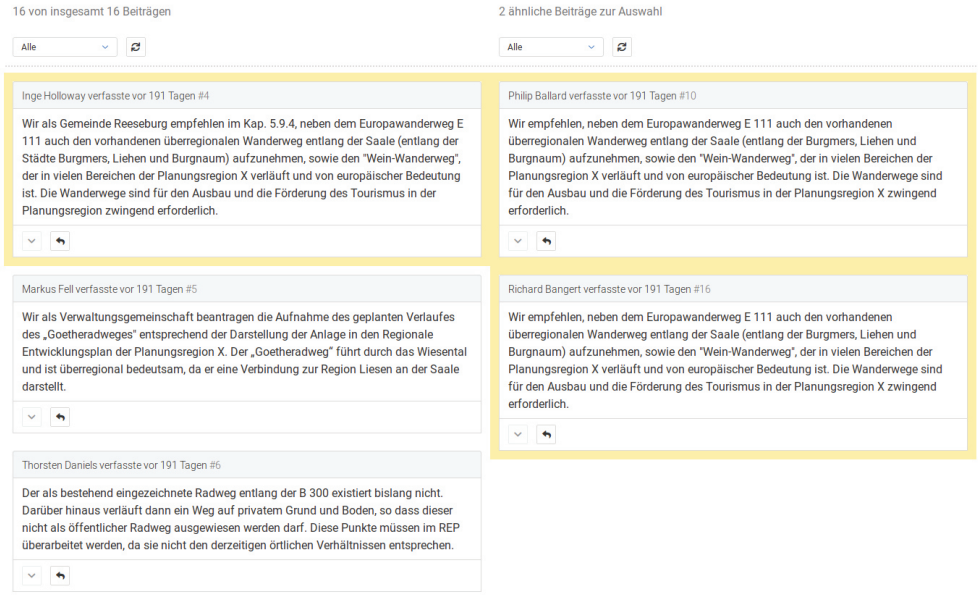

Figure 3: Schematic view of the assessment user interface. On the left hand side, there is a list of contributions. It shows a selected contribution (filled background color). On the right hand side, there is a list of ranked contributions for the selected contribution. These ranked contributions are sorted by semantic similarity.

We conducted a user study with 21 participants for evaluating the proposed assessment user interface. We exclusively focused on public administration workers. They had to complete a task that is related to the task of assessing contributions. They needed to find all contributions that are similar to a specific contribution. We opted for this proxy task due to time constraints for a single experiment. Overall, we tested two system configurations: one with the assessment user interface (system 1) and one without (system 2). System 2 only employed a one-column layout with the raw lists of not yet assessed contributions. The evaluation followed a within-subject design. We used different data sets of contributions so that the participants did not become too familiar with the contributions. Furthermore, we manually created the rankings for every contribution of the data set to ensure perfect rankings were used avoiding any negative impact on the participants due to bad quality rankings.

We found that participants who used the system 1 performed the task with higher accuracy, recall, and precision scores compared to those who used system 2, and only needed an extra 18 seconds to complete the task. Additionally, the majority of participants (16 out of 21) preferred using system 1 over system 2, and 17 participants found the ranking of contributions useful at all. However, there were also critical findings from the user study. Specifically, public administration workers did not fully understand the ranking of contributions, including the position of a contribution in the ranking. Only six participants understood that the higher a contribution is in the ranking, the more similar it is to the reference contribution. Furthermore, there were general concerns about the method used to compute the rankings, with some participants asking how the system ranked the contributions. Some participants assumed that the rankings were likely

based on the textual content of the contributions. Perhaps most significantly, the majority of participants (16 out of 21) stated that they generally do not trust the proposed rankings. We suggest that this lack of trust may be because no explanation was provided for the rankings, such as why a contribution in the ranking is similar to the reference contribution. Additionally, the participants did not know that we had manually created the similarities between the contributions. As a result, we argue for the need for interpretable ranking methods that provide more information and transparency to users.

4. Clustering of contributions

Clustering is a powerful method utilized in data analysis and machine learning to group data instances that exhibit similar characteristics. The primary objective of clustering is to segregate a dataset into different clusters or groups in such a way that the data instances within a cluster are more similar to each other than they are to instances in other clusters (Jain et al., 1999). When used for contributions, clustering automatically groups together contributions with similar concepts and ideas. This enables public administrations to assess similar contributions consistently and efficiently. It can lead to a more streamlined and efficient process of assessing contributions and improving the quality of decision-making. Citizens can also benefit from clustering as pre-computed groups of similar contributions can be easily explored.

To enable the clustering of contributions, it is necessary to first transform them into a structure that can be processed by clustering algorithms. These algorithms typically operate on vectors in a vector space, requiring that contributions be converted into vectors of the same vector space. In this paper, we focus on the main contents of the contributions that consist solely of textual data, and we describe a two-part transformation process that includes (1) text pre-processing and (2) embedding. Text pre-processing involves mandatory steps, such as tokenization, followed by optional steps, such as stop word removal or word stemming, that further refine the resulting tokens. The result is a list of tokens for each contribution. Embedding these pre-processed contributions into a vector space involves selecting from a range of methods that vary in complexity and sophistication, including term frequency–inverse document frequency, (averaged) word embeddings (Bojanowski et al., 2017; Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013), and (pooled) transformer-based embeddings (Brown et al., 2020; Devlin et al., 2019). Depending on the chosen method, additional information may be gathered during text pre-processing, such as recording the sentence to which a token belongs. Once the contributions have been embedded, they can be clustered. Figure 4 provides an overview of the transformation pipeline including the clustering step.

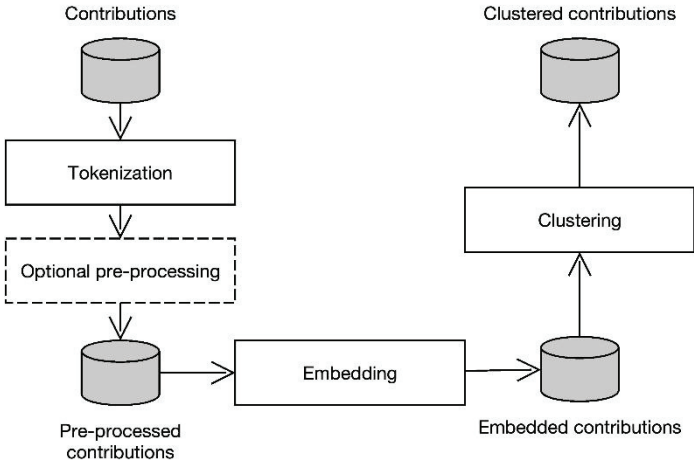

Figure 4: General transformation pipeline and final clustering

Although clustering contributions can help group similar contributions together, it also brings new challenges, such as the difficulty in interpreting clusters due to the vague definition of a cluster (Estivill-Castro, 2002). To mitigate this issue, we investigate example-based explanations (Miller, 2019) to provide clearer understanding of the clusterings. To this end, we propose two distinct methods for identifying prototypical instances of a clustering.

The first method is about using the maximum mean discrepancy (MMD). It describes the distance between two distributions by considering the distance between the mean embeddings of the data features. Regarding our specific domain, we need to consider the distribution of contributions X, the distribution of prototypes P, and a kernel function k, e. g., the radial basis function kernel or the polynomial kernel. Then we can estimate the MMD empirically as noted in the following equation:

$$MMD(X,P) = \left(\frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, x_j) - \frac{2}{nm} \sum_{i,j=1}^{n,m} k(x_i, p_j) + \frac{1}{m^2} \sum_{i,j=1}^{m} k(p_i, p_j) \right)^{\frac{1}{2}}.$$

In order to select prototypical instances of a clustering, we can utilize a variety of approaches that incorporate the MMD. While random selection is a simple and valid approach, it is not reliable. As an alternative, a greedy algorithm exists for selecting prototypes (Kim et al., 2016). This algorithm utilizes the MMD and iteratively computes the next best prototype until the desired number of prototypes have been found. This method is flexible and can be applied to any clustering algorithm. In our work, we focus explicitly on the clustering of contributions and apply the method to each cluster individually, allowing for a different number of prototypes for each cluster. This method can even be used on non-clustered contributions to identify the specified number of prototypes within the entire raw data set. However, we did not investigate this method further for non-clustered contributions due to our specific research focus.

The second method we propose is based on the k-medoids clustering algorithm (Kaufman & Rousseeuw, 1987). This algorithm is used to cluster contributions into distinct groups, and crucially, it outputs one medoid for each cluster. We can directly use a

medoid as the prototype for the corresponding cluster. However, this approach is limited in its flexibility, as it only allows for one prototype per cluster and requires the use of the k-medoids algorithm for clustering the contributions. Alternatively, we can modify the approach by using the idea of relying on medoids with any clustering algorithm. Then this involves clustering the contributions using any desired algorithm first, and subsequently computing the medoid for each cluster of the resulting clustering. This approach affords greater flexibility in terms of the clustering algorithm used, and it enables the identification of multiple prototypes per cluster.

5. Conclusion and future work

Public participation processes involve complex data, with contributions being a crucial component. However, analyzing and exploring such data can be challenging, requiring sophisticated methods to support these tasks. In this context, we have described two such methods, namely ranking and clustering, that can benefit both public administrations and citizens.

However, while these methods offer promising results, there is still much work to be done in terms of interpretability. Specifically, we need to develop new, more interpretable methods that are targeted towards laypersons, such as public administrations and citizens. Additionally, it is important to conduct user studies to evaluate the effectiveness of these new methods. This area of research presents a significant opportunity for future work, with the potential to make public participation processes more transparent and accessible to all.

References

Blotevogel, H. H., Danielzyk, R., & Münter, A. (2014). Spatial Planning in Germany. In M. Reimer, P. Getimis, & H. H. Blotevogel (Eds.), *Spatial Planning Systems and Practices in Europe* (pp. 83–108). Taylor & Francis Group.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5(0), 135–146.

Briassoulis, H. (1997). How the Others Plan: Exploring the Shape and Forms of Informal Planning. *Journal of Planning Education and Research*, 17(2), 105–117.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), Advances in neural information processing systems (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies (Vol. 1, pp. 4171–4186). Association for Computational Linguistics.

Estivill-Castro, V. (2002). Why so many clustering algorithms: A position paper. *ACM SIGKDD Explorations Newsletter*, 4(1), 65–75.

Helbig, D., Pietsch, M., Schütz, L., Bade, K., Richter, A., & Nürnberger, A. (2016). Online-Beteiligung in Entscheidungs- und Planungsprozessen – Anforderungen aus der Praxis. *Journal Für Angewandte Geoinformatik (AGIT) 2-2016*, 508–517.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. *ACM Comput. Surv.*, 31(3), 264–323.

Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. In *Statistical Data Analysis based on the L1-Norm and Related Methods* (pp. 405–416). Elsevier Science.

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not Enough, Learn to Criticize! Criticism for Interpretability. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 2280–2288). Curran Associates, Inc.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Higher Education from Cambridge University Press; Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv: 1301.3781 [cs.CL].

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality*. arXiv: 1310.4546 [cs.CL].

Schütz, L., & Bade, K. (2019). Assessment User Interface: Supporting the Decision-making Process in Participatory Processes. In J. Filipe, M. Smialek, A. Brodsky, & S. Hammoudi (Eds.), Proceedings of the 21st International Conference on Enterprise Information Systems, ICEIS 2019 (Vol. 2, pp. 398–409). SciTePress.

Schütz, L., Helbig, D., Bade, K., Pietsch, M., Nürnberger, A., & Richter, A. (2016). Interaction with Interconnected Data in Participatory Processes. In REAL CORP 2016 – SMART ME UP! How to Become and How to Stay a Smart City, and Does This Improve Quality of Life? Proceedings of 21st International Conference on Urban Planning, Regional Development and Information Society (pp. 401–410).

Schütz, L., Helbig, D., Bade, K., Pietsch, M., Richter, A., & Nürnberger, A. (2015). Projekt partiMAN: Neue Ansätze zur aktiven Partizipation in Entscheidungsprozessen. In R. Krug, M. Pietsch, M. Heins, & E. Kretzler (Eds.), *Beteiligen * kommunizieren * partizipieren* (pp. 72–88). Shaker Verlag.

Schütz, L., Raabe, S., Bade, K., & Pietsch, M. (2017). Using Visual Analytics for Decision Making. *Journal of Digital Landscape Architecture*, 2, 94–101.

Wong, P. C., & Thomas, J. J. (2004). Visual Analytics. *IEEE Computer Graphics and Applications*, 24(5), 20–21.