Ranking and Clustering Techniques to Support an Efficient E-Democracy

Marlene Goncalves[1], Maria-Esther Vidal[1], Francisco Castro[2], Luis Vidal[1,2], and Maribel Acosta[1,3]

[1] Universidad Simón Bolívar, Venezuela
[2] Universidad Central de Venezuela, Venezuela
[3] Institute AIFB, Karlsruhe Institute of Technology, Germany
{mgoncalves,mvidal,macosta}@ldc.usb.ve, maribel.acosta@kit.edu

Abstract. We focus on ranking and data mining techniques to empower e-Democracy and allow the opinion of ordinary people to be considered in the design of electoral campaigns. We illustrate the quality of our approach on Venezuelan historical electoral data; ranking results are compared to ground truths produced by an independent study. Our evaluation suggests that the proposed techniques are able to identify up to 85% of the golden results by just analyzing 35% of the whole data.

Keywords: E-Democracy, Skyline, Top-k, Clustering.

1 Introduction

Digital and Information technologies as well as Semantic Web standards and tools, facilitate the publication of governmental data and the development of applications that can impact on the citizens' quality of life. For example, e-Government initiatives as the ones supported by the *Data.gov*[1] and *GovTrack*[2] projects make available more than 450,000 datasets to empower normal people participation in governmental processes. This democratization of the information facilitates citizens' daily activities and also provides the basis to discover properties and relationships that could not be identified years before. Nevertheless, because these open datasets can be extremely large, applications need to be designed not just to meet soundness and completeness of the solutions, but to provide few relevant solutions quickly. We propose the construction of a multi-dimensional dynamic model that is able to explore large volumes of data very quickly with no limitations on data patterns or sources of information. Particularly, we apply this approach to identify electoral citizens' patterns in voting historical data. The identified patterns are used to discover electoral regions where a given electoral option has the potential of switching intended votes into actual votes. These selected regions will be the basis for the design of effective electoral campaigns. Ranking [1,2] and data clusterings [4] techniques

[1] http://www.data.gov/
[2] http://www.govtrack.us/

P. Herrero et al. (Eds.): OTM 2012 Workshops, LNCS 7567, pp. 298–301, 2012.

are combined into a two-fold solution to the problem of identifying these best regions. We demonstrate the benefits of the approach and show the following key issues: *i*) effectiveness of the approach on discovering the regions where an electoral option has good chances to win whenever the electoral champaign is intensified, and *ii*) efficiency by showing that the approach is able to identify up to 85% golden regions by just ranking 35% of the electoral data. The paper is composed of three additional sections. Section 2 defines our approach, and Section 3 reports the experimental results. Finally, we give our conclusions and future work in Section 4.

2 A Ranking and Clustering Approach

Our approach is comprised of two components: *i*) *Clustering* and *ii*) *Ranking*. The former offers different clustering algorithms, i.e., X-Means [4], to group multidimensional data according to the dataset attributes; voting centers with similar electoral patterns are placed together in the same cluster. Clusters' centroids correspond to a vector of the mean values of the electoral properties of the centers grouped in the cluster. The *Ranking* component implements techniques to identify the top-k electoral regions that best meet an electoral condition among the non-dominated regions. Non-dominated or*skyline* regions [1] are areas with at least one electoral value in the centroid that is better than the same electoral parameter of the centroids of the other areas. These areas have also at least one parameter in the centroid whose value is worse than the value of this parameter in centroids of the other non-dominated areas. Furthermore, a region is top-k if it is among the k regions with the smallest distance to the electoral condition; the ranking algorithm proposed by Goncalves et al. [3] is implemented.

3 Empirical Evaluation

The goal of our study is to analyze the efficiency and effectiveness of our approach in determining the parishes where one electoral option has the higher potential electoral value and at getting new votes. We compare the results obtained by our techniques to the results proposed by an independent study where the Venezuelan parishes were ranked according to the chances that an electoral option has to win in an electoral event. Analysis of past voting histories of two electoral events comprised our study. We analyzed an electoral dataset *ED* collected from the National Electoral Counsel[3]. This is comprised of 3,757 electoral centers distributed in 976 parishes which are grouped in 268 municipalities; 24 states comprise the municipalities. *ED* registered the electoral outcome of an event where citizens voted in a referendum to decide if an electoral option *A* was going to be included in the next national electoral event. Each electoral center of *ED* is characterized by: *i*) total number of registered voters (TNRV), *ii*) referendum election outcome (PEO), *iii*) number of votes in favor of option *A* (CFV),

[3] `http://www.cne.gov.ve/web/index.php`

iv) number of spoilt votes (NSV), and *v*) number of abstention (NA). Additionally, *ED* was enriched with two derived attributes which were computed in terms of historical data: *i*) number of potential intended voters (NPIV), *ii*) number of potential new voters (NPNV). We report on coverage measures of the percentage of parishes that are produced by our approach and that are present in the golden parishes (GP). A value of 100% indicates that all the top-k parishes produced by our technique are considered among the top-k in GP, while a value of 0%, suggests that none of these top-k parishes are in GP top-k parishes. Additionally, we measure performance as the percentage of electoral centers that were ranked to produce the top-k parishes. Evaluation steps are as follows:

Clustering of Electoral Centers: Two configurations of the X-Means algorithm implemented by WEKA[4], were used to cluster the electoral centers. [5]

Clustering 1: between 2 and 24 clusters were produced; the upper bound corresponds to the number of states in Venezuela. This configuration gives the algorithm the freedom to group the country according to their electoral characteristics and not in terms of an arbitrary lower bound.

Clustering 2: between 12 and 24 clusters were produced; the upper bound corresponds to the number of states and the lower bound is half of this number.

Computation of the Top-k Parishes: clusters were ranked according to the cluster centroid. Since centroids are multidimensional, a set of non-dominated parishes or skyline was computed. Additionally, to determine the top-k parishes among the skyline, the election condition was defined as those with NPIV at least of 50% and NPNV less or equal than 0. The Euclidean distance metric was used to calculate the distance between the centroids of the clusters in the skyline and the electoral condition; parishes in the top-k electoral regions were ranked in terms of the number of electoral centers.

Table 1. Clusters produced by two different configurations of the X-Means clustering algorithm; NC: Number of Generated Clusters; Min: Minimal Number of Electoral Centers per Clusters, Max: Maximal Number of Electoral Centers per Clusters; NCMRC: Number of Clusters that Meet Electoral Condition; PTEC: Percentage of electoral centers in NCMRC out of the Total Electoral Centers.

Strategy	# NC	(Min;Max)	NCMRC	PTEC
Clustering 1	**4**	(221;1,378)	**3**	**94%**
Clustering 2	20	(4;1,478)	4	12%

Table 1 describes the clusters produced by each cluster configuration. We can observe that *Clustering 1* is able to group a greater number of centers in fewer clusters, and it only generates a group of clusters where more than 75%

[4] `http://www.cs.waikato.ac.nz/ml/weka/`

[5] Information about location of the center was not considered.

of clusters meet the electoral condition, i.e., three out of four clusters belong to the skyline; also, the clusters that belong to the skyline group 94% of the total of electoral centers. Furthermore, Tables(a) and (b) report on the coverage of our proposed techniques; we can observe that grouping electoral centers into a small number of clusters may better approximate the current golden ranking; *Method 1* of *Clustering 1* is enough to identify up to 85% of the relevant parishes in the ground truths. Finally, Table(c) shows the percentage of centers that were considered in the ranking process; *Method 1* of *Clustering 1* is able to reach a coverage of 85% by just analyzing 35% of the centers.

Table 2. (a) and (b) Coverage of the top-k parishes with respect to the top-k in the ground truths. *Method i* indicates that top-i clusters are identified during the ranking process; (c) Percentage of the Total Electoral Centers considered by *Method i*.

(a) Clustering 1

top-k	Method		
	1	2	3
top-7	**85%**	**85%**	71%
top-30	46%	60%	56%
top-50	53%	58%	64%
top-100	**54%**	**64%**	63%

(b) Clustering 2

Top-k	Method		
	1	2	3
top-5	28%	14%	28%
top-30	13%	2%	2%
top-50	22%	22%	2%
top-100	24%	19%	19%

(c) Performance

Strategy	Method		
	1	2	3
Clustering 1	**35%**	57%	94%
Clustering 2	5%	10%	12%

4 Conclusions and Future Work

We describe a two-fold approach that relies on clustering and ranking techniques to identify electoral regions where a given option has the potential to win. Because these techniques consider the behavior of voters in different types of electoral events, they attempt to achieve campaigns that reflect the voting patterns of their participants. Empirically we probed that these techniques are able to predict up to 85% of the golden results identified by an independent study, while only 35% of the data is considered in the ranking. In the future, we plan to extend the study with other type of data, i.e., opinion polls.

References

1. Börzsönyi, S., Kossmann, D., Stocker, K.: The skyline operator. In: Proceedings of the 17th International Conference on Data Engineering, pp. 421–430. IEEE Computer Society, Washington, DC (2001)
2. Carey, M.J., Kossmann, D.: On saying "Enough already!" in SQL. SIGMOD Rec. 26(2), 219–230 (1997)
3. Goncalves, M., Vidal, M.-E.: Reaching the Top of the Skyline: An Efficient Indexed Algorithm for Top-k Skyline Queries. In: Bhowmick, S.S., Küng, J., Wagner, R. (eds.) DEXA 2009. LNCS, vol. 5690, pp. 471–485. Springer, Heidelberg (2009)
4. Pelleg, D., Moore, A.W.: X-means: Extending k-means with efficient estimation of the number of clusters. In: ICML, pp. 727–734 (2000)