2012 Seventh International Workshop on Semantic and Social Media Adaptation and Personalization

Clustering Online Poll Data: Towards a Voting Assistance System

Ioannis Katakis, Nicolas Tsapatsoulis, Vasiliki Triga, Constantinos Tziouvas
Cyprus University of Technology
{*ioannis.katakis, nicolas.tsapatsoulis, vasiliki.triga, costas.tziouvas*}*@cut.ac.cy*

Fernando Mendez
University of Zurich
fernando.mendez@zda.uzh.ch

***Abstract*—Voting advice applications (VAA) are very recently developed in order to aid users in deciding what to vote in elections. Every user is presented with a set of important issues and she is asked to submit her opinion by selecting one of a predefined set of answers (e.g. agree/disagree). The VAA gathers the same information for all candidates that are about to compete in the elections. Hence, it can provide recommendation to users: the candidates that agree with the user on these selected issues. In this paper, we propose a collaborating filtering approach for providing such suggestions. Like-minded users are clustered together based on their profiles (views on the selected issues) and voting recommendation is provided to a user by the members of the nearest (to her profile) cluster. We observe that this method produces more effective recommendations by utilizing two different measures: accuracy and weighted mean rank. Furthermore, the proposed method provides with important insight and summarization information about the electorate's opinion. This research is based on new data gathered by the voting advice application Choose4Greece which was widely used for the most recent elections in Greece.**

I. Introduction

Voting advice applications (VAA) are essentially vote recommendation systems and usually exploited during elections. VAAs are of vital importance since they promote rational reasoning for voting, fill information gaps and have positive impact on voter turnout [1]. Their use has increased in recent years especially in the European zone. Choose4Greece[1] is a very recent voting advice application that launched at the latest elections in Greece.

In most voting advice applications users are presented with a set of important political/financial/social issues in form of closed set questions. The users have to submit their opinion on these topics by selecting one answer out of a predefined set (e.g. strongly disagree, disagree, neither agree nor disagree, agree, strongly agree). The same information is collected for each political party / candidate. This is achieved either: a) by having a set of experts coding this information for each party (e.g. by studying each party's programmatic statements) or b) by requesting the candidates to answer the questions themselves.

The statements are composed by a set of experts who identify a small set of issues that are considered important for the specific time and nation that is about to elect a new government. Answers are stored anonymously into the VAA's database. Having the information of users and candidates, VAAs can provide recommendations by comparing 'issue' profiles and suggesting to the user the most relevant candidates according to the selected statements.

In this work, we propose a recommendation method that is based on applying data clustering on the user-base of the VAA. This way the users are organized into groups of like-minded voters. Knowing the vote intention of the users, we calculate the distribution of each political party in each cluster. Recommendation to a new user is achieved by finding the closest cluster to this user and recommend what the majority of voters intend to vote in this cluster. We observed that this approach performs more effectively than two baseline approaches. Furthermore, we argue that clustering provides with valuable information concerning the opinion of the electorate.

The contribution of this work can be summarized in the following points: (1) The proposal of a novel, accurate, clustering-based vote suggestion method, (2) a comparative study among three approaches, (3) a discussion and demonstration on the insight that clustering provides, and (4) a new pre-processed dataset made freely available online in an attempt to promote research in the field of VAAs.

The structure of the paper is as follows: Section II provides with the problem definition while Section III reviews the limited recent related work on VAAs. Section IV contains a description on the new Choose4Greece dataset and Section V presents the approaches that are going to be evaluated in Section VI. Finally, Section VII provides with an overview of our work and summarizes the key-points of our research.

[1] www.choose4greece.com

978-0-7695-4887-6/12 $26.00 © 2012 IEEE
DOI 10.1109/SMAP.2012.19

II. Problem Definition & Notation

In the problem of voting suggestion there is a set of N users $U = \{\vec{u_1}, \vec{u_2}, \ldots, \vec{u_N}\}$, a set of M questions (or issues) $Q = \{q_1, q_2, \ldots, q_M\}$, and a set of T political parties (or candidates) $P = \{\vec{p_1}, \vec{p_2}, \ldots, \vec{p_T}\}$. Each user $\vec{u_i} \in U$ and each political party $\vec{p_j} \in P$, has answered each question $q_k \in Q$. The answers of users are recorded through on-line questionnaires like the one in Choose4Greece. The answers of political parties are either coded by experts or answered by representatives of political parties.

Based on their answers, every political party or user can be represented in a vector space model:

$$\vec{u_i} = \{u_{(i,1)}, u_{(i.2)}, \ldots, u_{(i,k)}, \ldots, u_{(i,M)}\} \quad (1)$$

$$\vec{p_j} = \{p_{(j,1)}, p_{(j,2)}, \ldots, p_{(j,k)}, \ldots, p_{(j,M)}\} \quad (2)$$

where $u_{(i,k)}, p_{(j,k)} \in L$ are the answers of the i-th user and j-th party, respectively, to the k-th question. Usually, vectors $\vec{u_i}$ and $\vec{p_j}$ are named *profiles*.

A typical set of answers is a 6-point Likert scale: $L = \{1$ (Strongly disagree), 2 (Disagree), 3 (Neither agree nor disagree), 4 (Agree), 5 (Strongly agree), 6 (No opinion)$\}$ but in practice the sixth point it is not taken into consideration since does not correspond to a particular stance. As a result the set L, in the context of this work, becomes: $L = \{1,2,3,4,5\}$.

The *task*: Given the answers of a specific user $\vec{u_a}$ suggest a ranking of political parties based on user-party relevance. In machine learning terms, the task is to approximate the hidden function $h : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$, where $h(\vec{u}, \vec{p})$ is the estimation of the relevance of user $\vec{u}$ with political party $\vec{p}$. Typically $h(\vec{u}, \vec{p}) \in [0, 1]$. In each case, the top suggestion p_a for user u_a should be:

$$p_a = \underset{p}{\operatorname{argmax}}[h(\vec{u_a}, \vec{p})] \quad (3)$$

Similarly, we could consider a function $r(\vec{u}, \vec{p}) \in [1, T]$ that returns the *rank* of the political party p for the user u, if all political parties are ranked according to relevance (similarity) with this specific user. Having learned function $h(\vec{u}, \vec{p})$ it is straightforward to calculate $r(\vec{u}, \vec{p})$.

In order to produce vote recommendations, the most simple approach is to define $h(\vec{u}, \vec{p}) = d(\vec{u}, \vec{p})$ where d is a distance function between $\vec{u}$ and $\vec{p}$. A number of such distance measures are discussed in [2].

In many voting assistance systems, the information of *vote intention* v_i of many users $\vec{u_i}$ is available as it is included as a supplementary question in the online surveys. It is this kind of information we utilize in this work to provide collaborating filtering based voting recommendation and evaluate the proposed approaches.

III. Related Work

There is a lot of work on VAAs in the political science discipline [3]. However, very few scientists approached VAAs from the recommendation systems perspective and the majority of them consider only the problem of selecting an appropriate metric to compute the similarity between user and party/candidate profiles.

A. Similarity measures in Voting Advice Applications

In [2], Fernando Mendez compares four models for calculating the user-party profile distance. The first two are based on how close the answers of the party and the user are (proximity models) and are implemented either by Euclidean or City Block distance. The third metric is the Scalar Product which is a directional model. Directional models take into consideration the polarity of the opinions (i.e. if the answer of the voter and the candidate lie at the same side (disagree - agree) of the Likert scale). The last one is a Hybrid model. The basic claim of the paper is that the directional inspired models perform better. In [4] the authors share their concern that the output of voting assistance tools might be strategically manipulated by political actors and that VAAs might be most advantageous to non-programmatic political parties. Finally, Walgrave *et. al.* [5] study the effect of the selection of statements and its impact on the recommendations that are produced. The paper suggests that certain configurations might favor certain parties.

B. VAAs and Recommendation Systems

A recommendation system is an information system that recommends items (e.g. books or movies) to users. The methods that have been proposed for recommendations can be organized into the following categories (for an extended review on the field see [6]):

- Content-based: Users are recommended items similar to the ones they preferred in the past [7].
- Collaborative-filtering: Users are recommended items that users with similar preferences liked in the past [8].

For the first approach, the preferences of the user for other items (e.g. political parties), cannot be applied to VAA schemes since it does not make any sense. Furthermore, the whole voting history of a user is generally not collected or not available at all (new voters). Furthermore, as we discussed in Section II, the voting recommendation problem has one more dimension than conventional content-based recommendation problems. In our case, there are the users (voters), the items (political parties) and the questions. In order to produce recommendations we need to exploit all three elements. As a result we opt for a clustering-based collaborative filtering scheme. An interesting related work that is based on fuzzy clustering and fuzzy profiles is presented in [9], [10].

IV. Dataset Description

In this section we provide information about the newly introduced dataset. By making the dataset available online we intend to promote the research in the field.

A. Choose4Greece

Choose4Greece is a non-profit collaborative academic effort involving researchers from Cyprus University of Technology, Aristotle University of Thessaloniki, University of Zurich, University of Twente and University of Oxford. It was widely used for the national elections in Greece (June 2012 and May 2012). The Choose4Greece questionnaire can be accessed at: http://www.choose4greece.com.

B. The Dataset

The dataset consists of information collected from the usage of the Choose4Greece system during the period April - May 2012 for the 2012 National Elections in Greece. There were two rounds of elections in Greece 2012 (May 6th & June 17). The dataset under study includes data collected for the elections at May 6th.

Users of Choose4Greece had to submit their opinion for 30 issue statements plus some supplementary questions asking for demographic information, voting intention and self-placement on the main political dimensions (left/right, traditional/liberal). For each issue statement, the user had to choose one of the following answers: 1) Completely agree, 2) Agree, 3) Neither agree nor disagree 4) Disagree 5) Completely disagree 6) No opinion.

The dataset is available for research purposes at http://www.choose4greece.com/datasets/.

C. "Cleaning" the Dataset

The Choose4Greece dataset had to be pre-processed in order to remove invalid records. The first step was to filter all user-entries that did not exceed the time threshold of 100 seconds during the whole session. We considered that if a user spent less that 100 seconds to complete the full questionnaire (that is approximately 3 seconds per question - not considering the supplementary questions which are not mandatory) then probably she would answered the questions randomly. Another important step was to remove user entries that did not completed the full 30-questions. Finally, duplicate user entries were removed by IP filtering. After this process there were 75294 user entries.

V. Candidate Recommendation Systems

In this section we describe the approaches that are evaluated in the experimental study.

A. Party-Coding Similarity

This is the approach most widely used in Voting Assistance Applications. In this case $h(\vec{u},\vec{p}) = d(\vec{u},\vec{p})$, where d is the Euclidean distance:

$$d(\vec{u}_i,\vec{p_j}) = \sqrt{\sum_{k=1}^{M}(u_{(i,k)} - p_{(j,k)})^2} \quad (4)$$

Naturally, normalization is necessary if h is required to be in $[0,1]$, with 0 meaning identical profiles. However, since the recommendation is $p_a = \underset{p}{\operatorname{argmax}}[h(\vec{u_a},\vec{p})]$ normalization is not required, even if a ranking of political parties is requested.

The advantage of this approach is that it provides the degree of agreement / disagreement with each political party. This information normally demands significant effort on behalf of the user. Another positive aspect of this approach is computational simplicity. The main disadvantage is that the profiles of political parties / candidates are not easy to collect. Another concern with this method is that usually users do not vote based on agreement with political parties (non-issue voters). Many citizens tend to vote based on other criteria like personal relations with party, personality of the party leader, effectiveness in solving the problems, etc (see [2] for more information).

B. Average Voter

This is a simple approach that calculates the distance between the user and the average voter of each party. The party with the nearest average voter comprises the recommendation in this approach. The average voter of party p_j is defined as:

$$a(\vec{p_j}) = \frac{1}{N_j}\{\sum_{i=1}^{N_j} u_{(i,1)},\ldots,\sum_{i=1}^{N_j} u_{(i,k)},\ldots,\sum_{i=1}^{N_j} u_{(j,M)}\} \quad (5)$$

where N_j are the total number of voters of political party p_j. In this approach $h(\vec{u_i},\vec{p_j}) = d(\vec{u_i}, a(\vec{p_j}))$ where d is the Euclidean distance. As discussed previously depending on the application requirements h should be normalized.

The advantage of this approach is that it does not require the profile of each political party and that it is computationally undemanding. However, a sufficient number of users is necessary in order to calculate the average voters. In recommendation system literature this issue is known as "cold-star" problem [6].

C. Clustering

Our approach is based on data clustering [11]. Given a set of data points in a multi-dimensional space, a clustering algorithm is able to organize data points into similar groups (*clusters*). Partitioning algorithms, like the widely known k-means, organize data based on feature space distance. Points that are close are organized into the same group. In this work, we exploit clustering in order to organize voters into clusters: Voters will be similar in terms of their feature vector which expresses their answers in Choose4Greece's issue-questions. Therefore, clustering will produce groups of like-minded users. After creating clusters, the system will be able to produce vote recommendations for new users. This is achieved by calculating the closest cluster to the new user. Then, the system suggests the political party that has the greatest number of voters in that cluster.

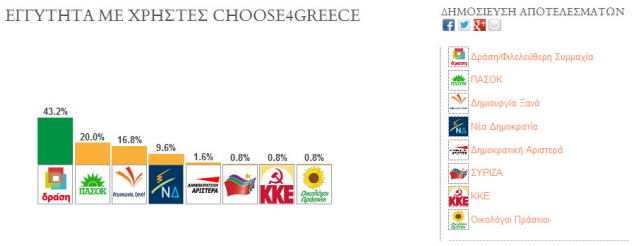

Figure 1. A screenshot of the clustering based voting recommendation

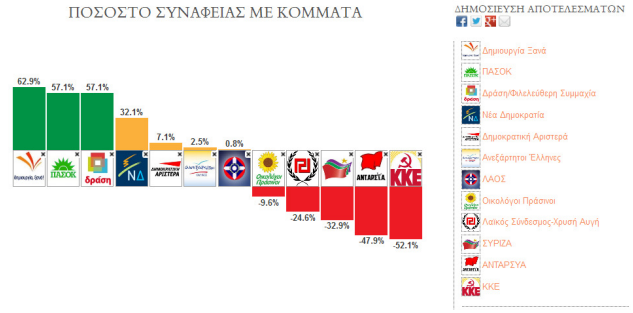

Figure 2. A screenshot of the user-candidate similarity scores for the user of Figure 1

A particular example of voting recommendation based on user clustering is shown in Figure 1. Once the user completed the set of 30 questions the closest cluster according to her profile was computed. Percentages of voting intention of the members of this cluster are used as recommendation and are illustrated in Figure 1. Note, however, that the results refer to cluster members that answered the supplementary question on voting intention. In summary: 43.2% of the cluster members that answered the question on voting intention choose to vote for 'DRASI' (the corresponding bar is colored green showing strong match), 20% choose to vote for 'PASOK', etc. Thus the voting recommendation was 'DRASI'. The corresponding recommendation based on the user-candidate similarity scheme, for the same user, is shown in Figure 2. We observe important differences both in the similarity scores and ranking of parties.

An obvious advantage of the user clustering approach is that it is not necessary to obtain the profiles of each political party / candidate. However, the most important characteristic is that it enables the organization of users into clusters. This feature will provide with three more advantages.

Firstly, it will enable the production of more accurate recommendations than the average voter and the user-candidate similarity since it will enable to create finer groups of users that hopefully will vote for the same candidate.

Secondly, it provides with valuable insight of the electorate. See for example the clusters produced after applying k-means at the Choose4Greece data (Table I). One could note some interesting observations on this outcome. For example, we observe that Cluster 5 consists mostly of Siriza voters, Cluster 8 is a group of left parties voters (KKE, Siriza, Dimokratiki Aristera), Cluster 4 consists of voters with right-conservative orientation (Nea Dimokratia, Anexartitoi Ellines, Xrisi Augi) and finally, Cluster 7 and 9 seems to have voters from various political parties.

Finally, each cluster can be represented by a centroid (average vector of each cluster). This representation is of great importance since it enables to interpret the opinions that dominate each cluster and can be exploited as data compression technique.

VI. Evaluation

A. Evaluation Setup

We separated the dataset into training and test set (70%-30%). In Average Voter, the training set is used to calculate the average vectors for each political party. In the clustering approach, the training set is used to organize the voters into clusters and calculate the vote distributions for each cluster. The evaluation of all approaches (calculation of evaluation measures) was carried out in the test set. For Clustering, the Weka implementation of k-means was exploited [12].

B. Evaluation Measure

In order to evaluate and compare the aforementioned approaches in terms of quality of prediction, we exploit the following evaluation measures:

1) Accuracy: This is a widely used evaluation measure for classification problems. It calculates the number of correct predictions. If a prediction of an approach h for user i is $p_i = \underset{p}{\text{argmax}}[h(\vec{u_i}, \vec{p})]$ and the vote intention is v_i then, accuracy for method h in dataset D is calculated as:

$$acc(h, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} e(p_i, v_i) \tag{6}$$

where $|D|$ denotes the cardinality (i.e., number of user entries) of set D, and

$$e(p_i, v_i) = \begin{cases} 1 & \text{if } p_i = v_i \\ 0 & \text{if } p_i \neq v_i \end{cases}$$

Accuracy is a strict measure that considers only the cases where the recommendation system has placed first the correct political party / candidate.

2) Weighted Mean Rank: Is a measure that evaluates how high did the recommendation system placed the correct political party. We consider this as a more fair evaluation measure. Consider two recommendation systems h_1 and h_2 with ranking functions r_1 and r_2 and a user u_a with vote intention v_a. If $r_1(u_a, v_a) = 2$ and $r_2(u_a, v_a) = 4$ then these cases will be treated equally in accuracy since none of these methods ranked first the correct political party (v_a).

Table I
THE TEN CLUSTERS CREATED USING K-MEANS ($k = 10$)

#	PASOK	ND	KKE	LAOS	SIRI	DIAR	DISI	ANEL	OP	KISI	ARPO	DRASI	ANTA	XA	DIKS
1	14%	2%	2%	0%	12%	30%	2%	2%	10%	1%	0%	16%	2%	1%	6%
2	14%	30%	1%	3%	2%	6%	8%	8%	2%	1%	0%	14%	0%	8%	4%
3	21%	9%	0%	1%	0%	5%	9%	1%	2%	0%	0%	47%	0%	1%	4%
4	1%	11%	2%	5%	8%	3%	1%	32%	2%	1%	0%	2%	1%	29%	2%
5	1%	1%	8%	1%	29%	10%	1%	26%	5%	1%	1%	4%	2%	9%	2%
6	1%	1%	14%	1%	23%	3%	0%	31%	2%	0%	0%	1%	3%	19%	1%
7	4%	9%	3%	2%	12%	14%	3%	21%	4%	1%	0%	13%	1%	9%	4%
8	1%	0%	11%	0%	54%	10%	0%	4%	7%	1%	0%	1%	8%	1%	1%
9	5%	6%	6%	1%	26%	17%	2%	18%	6%	2%	0%	3%	2%	4%	3%
10	0%	0%	32%	0%	40%	1%	0%	1%	1%	0%	0%	0%	24%	0%	0%

This problem is alleviated in the weighted mean rank which is defined as follows:

$$wmr(h, D) = \frac{1}{T} \sum_{j}^{T} w_j \frac{1}{N_j} \sum_{i}^{N_j} r(u_i^j, p_j) \qquad (7)$$

where T is the number of political parties, w_j is the percentage of voters that party j collected in the training set, N_j is the number of voters of party j in the evaluation set, r is the ranking function corresponding to recommender h and u_i^j is the i user (voter) of political party j. wmr takes into consideration the ranking of the correct political party (vote intention) and the number of the voters of each political party. Weighted mean rank firstly introduced in [2] for voting assistance applications.

C. Results and Discussion

1) Comparative Results: Table II displays the results of the three approaches in two evaluation measures: Accuracy (acc) and Weighted Mean Rank (wmr). For clustering, we use k-means algorithm with $k = 200$, we elaborate on the selection of k later on.

Table II
COMPARISON OF VOTING RECOMMENDATION SCHEMES

	acc	wmr
Party Coding	20.6%	4.03
Average Voter	32.8%	3.53
Clustering(k=200)	41.9%	2.93

We observe that the clustering approach performs better in both measures (accuracy & weighted mean rank). This fact confirms our initial intuition that the clustering will organize the users into like-minded voters who tend to vote the same political party. Average voter presents better predictive performance than the baseline of user-candidate similarity. The bad performance of user-candidate similarity proves that voters don't agree in the selected issues with the political parties that they vote. In general, acc and wmr seem to be correlated. The method with the best (highest) acc produces the best (lowest) wmr as well.

2) The effect of number of clusters: In Figures 3 and 4 we observe the variation of performance for clustering with respect to k. In both metrics the performance seems to be stable with respect to k and better than the other two approaches. However, if we observe Figures 5 and 6 we note that the performance is reduced when the number of clusters increases drastically. This can be explained by the fact that with such large values of k, small clusters (clusters with only few members) will be created. Obviously, small clusters do not contain enough number of voters to comprise a sufficient block of like-minded voters. This is a rather important conclusion, since it suggests that our approach is independent of the number of clusters (k) as long as it enables a sufficient number of users at each cluster. For the case of the Choose4Greece dataset the experiments suggest a number of clusters between 110 and 500.

VII. CONCLUSION

In this work, we introduced a collaborating filtering approach for voting aid applications. In contrary to the traditional and widely used user-candidate similarity approach we approximate voting advice as a recommendation problem. Like-minded voters are clustered together according to their profiles using the k-means algorithm and for every user the nearest cluster is selected. The percentage of voting intention of the members of this cluster indicate the recommendation towards a party/candidate. The proposed method outperforms both the user-candidate and user-average party voter similarity approaches. Our approach not only produces better predictive results but can provide with insight about voter opinion. Collaborating filtering voting recommendation enlarges the scope of traditional voting aid schemes which aim at 'issue-voters' [2], that is those whose vote choice is based on the policy stance of a candidate/party on a given set of policy issues, by including 'non-issue' voters, those whose vote choice is based on other factors including sociological/psychological ones, such as party identification, and valence factors such as the perceived competence of a candidate/party to deliver the desired goals. Finally, in this study we provide the first release of the poll dataset for voting recommendation which can be accessed at

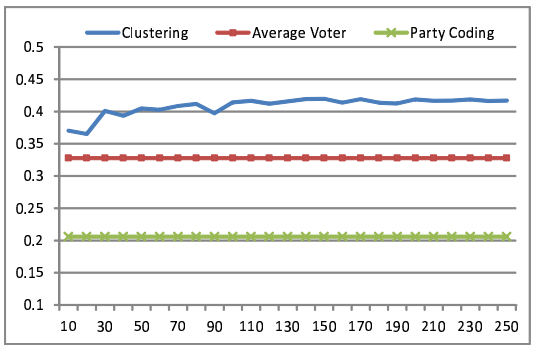

Figure 3. Clustering: accuracy(Y) vs $k(X)$ (10 to 250, step 10)

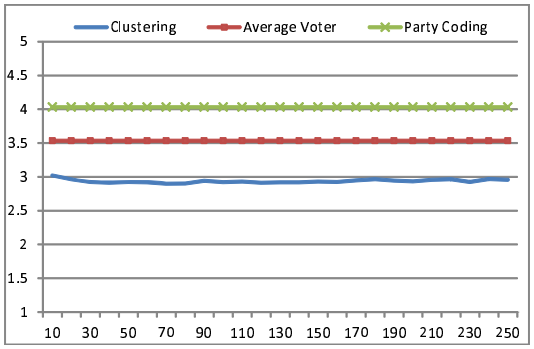

Figure 4. Clustering: weighted rank(Y) vs $k(X)$ (10 to 250, step 10)

http://www.choose4greece.com/datasets/ in order to promote research in the field.

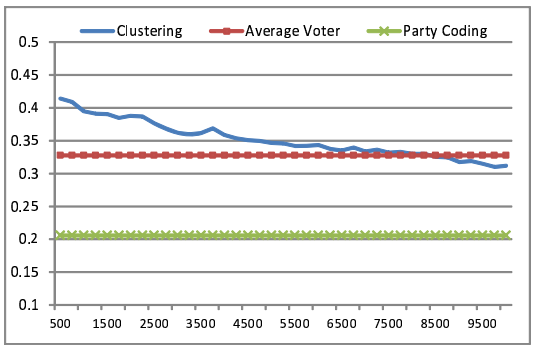

Figure 5. Clustering: accuracy(Y) vs $k(X)$ (500-10000, step 250)

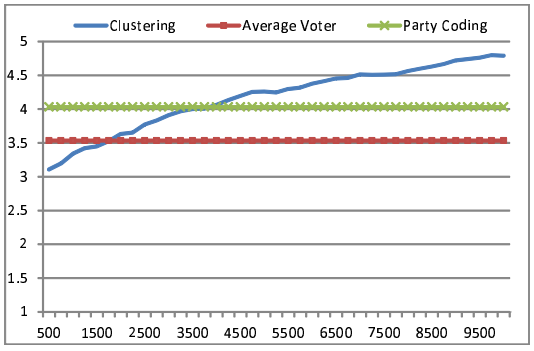

Figure 6. Clustering:weighted rank(Y) vs $k(X)$ (500-10000, step 250)

References

[1] A. Ladner and J. Pianzola, "Do voting advice applications have an effect on electoral participation and voter turnout? evidence from the 2007 swiss federal elections," in *Electronic Participation*. Springer Berlin / Heidelberg, 2010, vol. 6229, pp. 211–224.

[2] F. Mendez, "Matching voters with political parties and candidates: An empirical test of four algorithms," *International Journal of Electronic Governance*, 2012, (in print).

[3] T. Chadjipantelis, U. Serdült, and V. Triga, "Special issue on "voting advice applications and state of the art: Theory, practice, and comparative insights"," *International Journal of Electronic Governance*, 2012, (in print).

[4] A. Ramonaite, "Voting advice applications in lithuania: Promoting programmatic competition or breeding populism?" *Policy & Internet*, vol. 2, 2010.

[5] S. Walgrave, M. Nuytemans, and K. Pepermans, "Voting aid applications and the effect of statement selection," *West European Politics*, vol. 32, no. 6, pp. 1161–1180, 2009.

[6] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. on Knowl. and Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005.

[7] M. Balabanović and Y. Shoham, "Fab: content-based, collaborative recommendation," *Commun. ACM*, vol. 40, no. 3, pp. 66–72, Mar. 1997.

[8] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, "Grouplens: applying collaborative filtering to usenet news," *Commun. ACM*, vol. 40, no. 3, pp. 77–87, Mar. 1997.

[9] L. Terán, A. Ladner, J. Fivaz, and G. Stefani, "Using a fuzzy-based cluster algorithm for recommending candidates in eelections," in *Fuzzy Methods for Customer Relationship Management and Marketing - Applications and Classification*, ser. 2, A. Meier and L. Donze, Eds. IGI Global, 01/2012 2012, ch. 6, pp. 115–138.

[10] L. Terán and A. Meier, "Smartparticipation – a fuzzy-based platform for stimulating citizens' participation," vol. 4, pp. 501 – 512, 12/2011 2011.

[11] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.