Discovering Spatio-Social Motifs of Electoral Support Using Discriminative Pattern Mining[*]

Tomasz F. Stepinski and Josue Salazar[†] , Wei Ding[‡]

ABSTRACT

Association analysis provides a natural, data-centric framework for the discovery of patterns of explanatory variables that are linked to a certain outcome. In this paper we demonstrate how such a framework can be applied for political analysis, using an expository example of discovering different spatio-social motifs of support for Barack Obama in the 2008 presidential election. Election results and thirteen different socio-economic explanatory variables, tabulated at the county level, are used as an input for calculating a collection of discriminative patterns having disproportionately large support within the counties won by Obama. These patterns are synthesized into a small number of larger socio-economics motifs using a novel pattern similarity measure that outputs a concise summary readily interpretable in terms of political analysis. The method discovers two major Obama constituencies that differ in their socio-economic makeup and in their geographical distributions. The larger constituency can be further divided into more narrowly-defined motifs.

Categories and Subject Descriptors

I.5.3 [**Pattern Recognition**]: Clustering—*Similarity measures*; J.4 [**Computer applications**]: Social and Behavioral Sciences—*sociology,economics*; H2.8 [**Database Management**]: Database applications—*spatial databases and GIS*

Keywords

Discriminative pattern mining, political analysis, summarization, visualization

1. INTRODUCTION

In our recent work [2, 10] we have been exploring a data mining-based methodology for systematic and comprehensive assessment of explanatory factors responsible for spatial change. Spatial change refers to a significant contrast in the magnitude of response variable between different geographical regions arising from spatial distribution of controlling variables. In our methodology change factors are not discovered individually, but rather as discriminative patterns [3] of multiple, potential explanatory factors. Once discovered, the patterns are synthesized into a small number of "super-patterns" that represent comprehendible and actionable form of knowledge. This method has been successfully applied previously to discover motifs (and their geographical extents) of environmental variables driving high density of vegetation over the continental United States [2, 10]. In this paper we demonstrate how our methodology can be applied to the field of political analysis by using it to discover different socio-economic patterns associated with political support for Barack Obama in the 2008 presidential election. The results of the election (response variable) and the thirteen different socio-economic indicators (explanatory variables), tabulated at the county level, are used for discovery of spatio-social patterns associated with electoral support for Obama. Two distinct socio-economic Obama constituencies, each having different geographical distribution, are discovered and analyzed.

2. METHODOLOGY

Our dataset is organized around spatial objects (counties) characterized by their geographical location (county ID #), attributes (socio-economic indicators), and labels indicating whether the county was won or lost by Barack Obama. Information about each county is structured as follows: $o = \{id; f_1, f_2, ..., f_m; c\}$, where id is county's number identifying its geographical location, f_i, $i = 1, \ldots, m$, are values of m socio-economic indicators (explanatory variables) potentially influencing a vote, and c is the binary outcome of the vote (a response variable). The entire dataset $\mathcal{O}$ is analyzed using association analysis framework [11, 1, 4]. From the point of view of association analysis, each county (after disregarding its id and c attributes) is a transaction containing a set of exactly m items $\{f_1, f_2, ..., f_m\}$. The entire spatial dataset can be viewed as a set of N fixed-length transactions, where N is the size of the dataset.

An itemset (hereafter also referred to as a pattern) is a set of items contained in a transaction. For example, assuming $m = 13$, $P = \{2, \_, \_, \_, 3, \_, \_, \_, \_, \_, \_, \_, \_\}$ is a pattern in-

[*] A portion of this research was conducted at the Lunar and Planetary Institute, which is operated by the USRA under contract CAN-NCC5-679 with NASA.

[†] Lunar and Planetary Institute, Houston, Texas, tom@lpi.usra.edu, salazar@lpi.usra.edu

[‡] Dept. of Computer Science, University of Massachusetts Boston, Boston, Massachusetts, ding@cs.umb.edu

COM.GEO 2010, Washington, DC, USA

dicating that $f_1 = 2$, $f_5 = 3$ while the values of all other attributes are not parts of this pattern. A transaction *supports* an itemset if the itemset is a subset of this transaction; the number of all transactions supporting a pattern is refereed to as a *support* of this pattern. Because transactions are associated with geographical regions (counties), there is also a spatial manifestation of support which we call a *footprint* of a pattern.

2.1 Discriminative patterns

A discriminating pattern X [3] is an itemset consisting of the values of socio-economic indicators that has much larger support within a set of transactions $\mathcal{O}_\mathrm{p}$ stemming from the counties that were won by Obama ($c = 1$) than within a set of transactions $\mathcal{O}_\mathrm{n}$ stemming from the counties that were won by John McCain ($c = 0$). For a pattern X to be accepted as a discriminating pattern, its growth rate, $\frac{\sup(X,\mathcal{O}_\mathrm{p})}{\sup(X,\mathcal{O}_\mathrm{n})}$, must exceed a predefined threshold δ, where $sup(X, \mathcal{O})$ denotes the support of X in $\mathcal{O}$.

We mine only for *closed* discriminative patterns that are relatively frequent in $\mathcal{O}_\mathrm{p}$. Mining for frequent patterns reduces computational cost. In addition, infrequent patterns, even if highly discriminative, do not represent major spatio-social motifs. Further significant reduction in computational cost is achieved by mining only for frequent closed patterns [8]. A closed pattern is a maximal set of items shared by a set of transactions. A closed pattern gives the most detailed motif of indicators in a given set of counties; there is little point in considering less descriptive motifs in the same set.

2.2 Information synthesis

Like all methods based on association analysis our technique finds thousands of patterns, each representing a nugget of specialized information about a spatio-social motif in a single county or a small set of counties. A set of all discriminative patterns gives many such nuggets of specific but localized knowledge, making it difficult to use them to gain insight about the entire country. We cluster the patterns into a small number of "super-patterns" using a similarity (distance) measure between two patterns. Our measure of similarity between two patterns is defined as a similarity between two sets grouping objects belonging to the footprints of the corresponding patterns.

We define the similarity between patterns X and Y as $S(X,Y) = \sum_{i=1}^{m} w_i S_i(X_i, Y_i)$, where X_i, Y_i indicate the ith attribute, w_i indicates the ith weight, and m is the number of attributes. The similarity between ith attribute in the two patterns $S_i(X_i, Y_i)$ is calculated using group average, a technique similar to the UPGMA (Unweighted Pair Group Method with Arithmetic mean) [7] method of calculating linkage in agglomerative clustering. The UPGMA method reduces to $S_i(X_i, Y_i) = s(x_i, y_i)$ for attributes which are present in both patterns; x_i and y_i are the values of attributes X_i and Y_i, respectively, and $s(x_i, y_i)$ is the similarity between those values. If the ith attribute is present in the pattern Y but absent in the pattern X the UPGMA method reduces to

$$S(-, Y_i) = \sum_{k=1}^{n} P_X(x_k) s(z_k, y_i)$$

where $P_X(x_k)$ is the probability of ith attribute having the value x_k in all objects belonging to the footprint of X and n is the number of different values the ith attribute can have. The UPGMA reduces to an analogous formula if the ith attribute is present in the pattern X but it's absent in the pattern Y. Finally, if the ith attribute is absent in both patterns the UPGMA gives

$$S(-_i, -_i) = \sum_{l=1}^{n} \sum_{k=1}^{n} P_X(x_l) P_Y(y_k) s(x_l, y_k)$$

We calculate the similarity between the two values of ith attribute using a measure inspired by an earlier concept of measuring similarities between ordinal variables using information theory [6]. The similarity between two ordinal values of same attribute $s(x_i, y_i)$ is measured by the ratio between the amount of information needed to state the commonality between x_i and y_i, and the information needed to fully describe both x_i and y_i.

$$s(x_i, y_i) = \frac{2 \times \log\ P(x_i \vee z_1 \vee z_2 \ldots \vee z_k \vee y_i)}{\log\ P(x_i) + \log\ P(y_i)}$$

where $z_1, z_2, \ldots, z_k$ are ordinal values such that z_1 is the next higher adjacent value to x_i and z_k is the next lower adjacent value to y_i. Probabilities, $P()$, are calculated using the known distribution of the values of ith attribute in $\mathcal{O}_\mathrm{p}$.

To calculate the super-patterns we first calculate a distance matrix between each pair of patterns and then perform agglomerative clustering directly from the distance matrix. We use agglomerative clustering because it aggregates counties without breaking the patterns. Sammon's map [9] is used to visualize the overall topological structure of the entire set of patterns and to suggest the number of super-patterns.

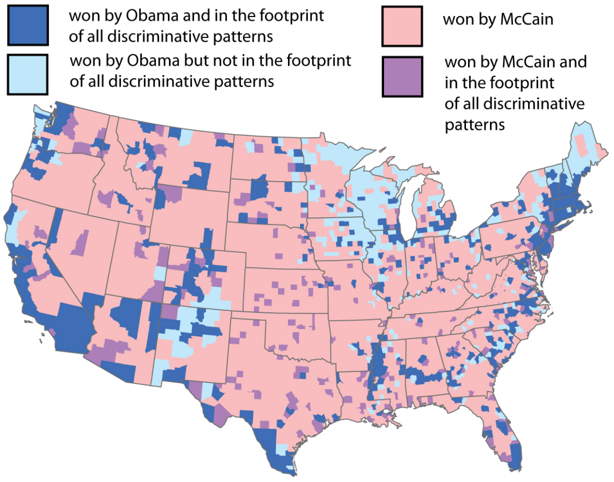

won by Obama and in the footprint of all discriminative patterns
won by Obama but not in the footprint of all discriminative patterns
won by McCain
won by McCain and in the footprint of all discriminative patterns

Figure 1: Map of the contiguous United States showing relation between the footprint of patterns discriminating between Obama and McCain and the actual results of the 2008 presidential election on the county level.

3. EXPERIMENTS

We use the county-level 2008 presidential election data for 3108 counties within the contiguous United States as an expository example of our method application to the

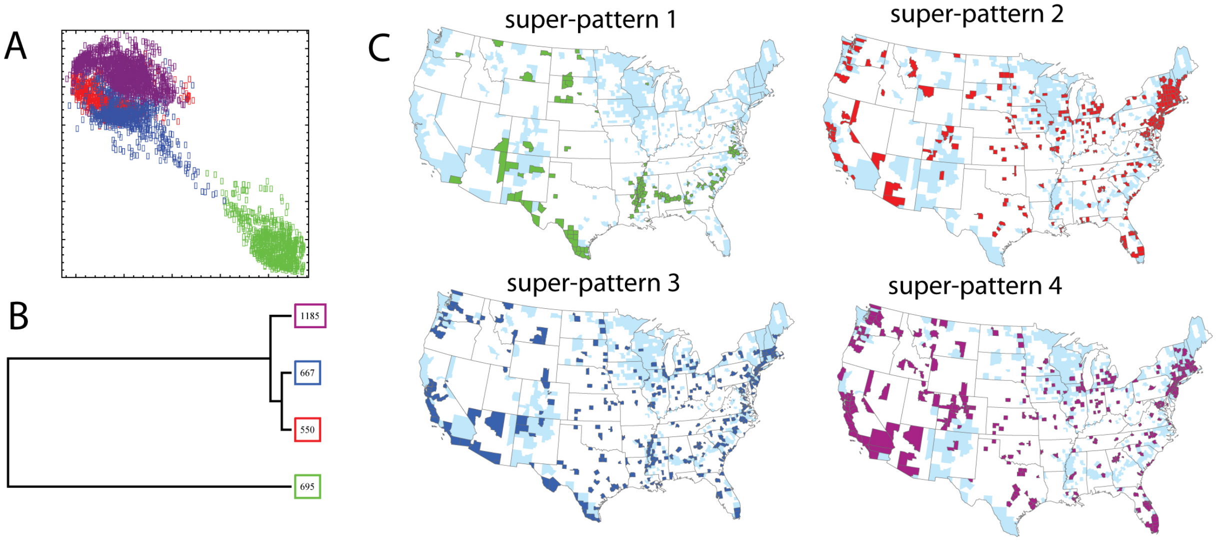


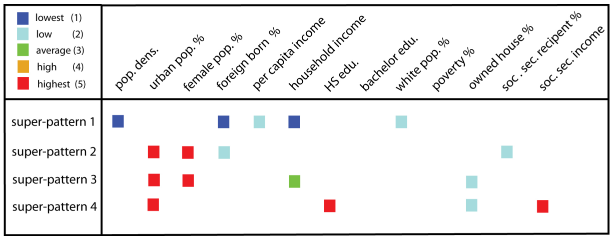
Figure 2: (A) Sammon's map showing topological relations between 3097 discriminative patterns. (B) Dendrogram showing results of agglomerative clustering of 3097 discriminative patterns into 4 super-patterns. (C) Geographical distribution of footprints of the four identified super-patterns.

area of political analysis. For these counties we have selected 13 socio-economics indicators using the Census Bureau data. These indicators are: (1) population density, (2) % of urban population, (3) % of female population, (4) % of foreign-born population, (5) per capita income, (6) median household income, (7) % of population with high school or higher education, (8) % of population with bachelor degree or higher education, (9) % of population that is white, (10) % of population living in poverty, (11) % of houses occupied by owners, (12) percentage of population receiving social security benefits, (13) average social security monthly benefit. The socio-economic indicators are transformed into ordinal-valued attributes using "natural brakes" method [5] in order to fulfill association analysis requirement for categorical variables. We use five categories (bins) denoted as "lowest", "low", "average", "high", and "highest", respectively.

We conducted two different experiments: (1) using a single transaction for each county (all counties contribute equally regardless of their population) (2) using a number of (identical) transactions for each county in proportion to the its population. The results of the two experiments differ; due to a limited length of this contribution we report here only on the results of experiment (1). Discriminative patterns were found using a growth rate threshold $\delta = 15$. With such threshold 3097 patterns were found ranging in support from as little as 9 counties to as much as 103. The summary of geographical aspect of discovered pattern are given in Fig. 1. This figure illustrates disparity between a combined footprint of all discriminative patterns and the footprint of all the counties won by Obama. In the ideal case (when the correlation between exploratory and response variables is stationary) these two footprints would be identical. In the actual case local disparities exist because some (sparsely populated) counties do not follow the national trend. There are 479 counties (constituting 91% of the population living in the counties won by Obama) that are in the footprint of discriminative patterns. There are also 327 counties (constituting, however, only 9% of the population living in the counties won by Obama) which are left out of the footprint of discriminative patterns. These counties are located mostly in the Midwest and New England. There are 214 counties won by McCain (constituting 37% of the population living in the counties won by McCain) that are, nevertheless, in the footprint of patterns discriminating for Obama.

No discriminative patterns consist of all 13 indicators, the most descriptive patterns involve only 11 indicators; there are two such patterns each being supported in 9 counties. Our analysis reveals that descriptive patterns, those consisting of a large number of indicators, tend to describe counties that are characterized by low population density, low income levels, and low education levels. Thus, it appears that these indicators are highly correlated to each other in counties that Obama won in 2008. In other counties, which Obama also has won, but are characterized by other set of indicators, there appear to be less correlation between the indicators resulting in less descriptive patterns. The single least descriptive discriminative pattern involves only one indicator, % of population that is white = lowest, and is supported in 103 counties. Most patterns consist of four (818 patterns) or five (1024 patterns) indicators.

In-depth analysis of 3097 patterns of different level having specificity and support is not practical. We use our pattern similarity measure to cluster the patterns into a small number of clusters of patterns (called super-patterns). Fig. 2 shows the result of such clustering. Panel A is the Sammon's map that visualizes in 2-D the "distances" between the patterns - similar patterns are close to each other on the Sammon's map. The map reveals that all patterns could be naturally divided into just two large clusters. The four different colors of points corresponding to patterns on the map represents four clusters found using the agglomerative clustering (see panel B). The hierarchy of clustering is terminated (arbitrarily) at four clusters; three of these (closely related) clusters correspond to the agglomeration seen in the left-upper corner of the Sammon's map. Panel C shows geographical distribution of footprints corresponding to the four super-patterns.



🟦 lowest (1)
🩵 low (2)
🟩 average (3)
🟧 high (4)
🟥 highest (5)

	pop. dens.	urban pop. %	female pop. %	foreign born %	per capita income	household income	HS edu.	bachelor edu.	white pop. %	poverty %	owned house %	soc. sec. recipient %	soc. sec. income
super-pattern 1	🟦			🟦	🩵	🟦			🩵				
super-pattern 2		🟥	🟥	🩵								🩵	
super-pattern 3		🟥	🟥			🟩					🩵		
super-pattern 4		🟥					🟥				🩵		🟥

Figure 3: Meaning of the four super-patterns in terms of socio-economic indicators

Each supper-pattern agglomerates a large number (from as little as 550 to as much as 1185) individual patterns; the table shown in Fig. 3 gives a brief socio-economic interpretation to each super-pattern. Note, that a super-patterns, just like individual patterns, are not described in terms of all potential indicators. Supper-pattern 1 is found in sparsely populated, low income counties with large minority populations. The other three super-patterns are fund in counties dominated by urban populations. In addition, super-patterns 2 and 3 are associated with counties with disproportionately large female populations, and the super-patterns 2 and 3 are associated with low percentage of home ownership. Other details can be found in the table given in Fig. 3. The footprints of different super-patterns overlap, however, there are only six counties where all four patterns are found. There is little geographical overlap between the footprint of super-pattern 1 and the other super-patterns. There are 125 counties (mostly associated with major cities) where footprints of all three urban super-patterns overlap.

4. CONCLUSIONS

In this paper we have demonstrated the utility of our association analysis-based methodology for discovery of change factors to the domain of political analysis. Our methodology departs fundamentally from the bulk of computational methods presently utilized in political analysis inasmuch as it based on the data mining paradigm and not on the regression paradigm. The core contribution of our method is the similarity measure between two patterns that allows for pattern clustering and thus provides an efficient tool for summarizing the results of association analysis.

The expository example presented here focuses on finding spatio-social motifs of electoral support for Barack Obama in the 2008 presidential election. We refer to the presented calculations as an "expository example" because we have selected socio-economic indicators without in-depth research of what indicators are most appropriate from the point of view of the political analysis. Nevertheless, our calculations have discovered interesting segmentation of the electoral support in the 2008 election. A similar analysis, but with transactions weighted by population of the counties reveals a similar segmentation, but does not find a spatio-social motif associated with the super-pattern #1. This is because weighting transactions by population decreases the relative importance of sparsely populated counties. Another interesting observation is provided by geographical distribution of counties that Obama has won, but which are not in the footprint of discriminative patterns. These counties, located predominantly in New England and Midwest states of Wisconsin, Minnesota, Iowa, and Illinois, went for Obama despite having socio-economic motifs that are not associated with Obama-leaning counties. Evidently, factors other than those considered here were responsible for the outcome in those counties. Similarly, a number of counties won by McCain fulfill Obama-leaning motifs of electoral support. Again, factors other than those considered here may play the role in the outcome. Calculations with more, carefully selected indicators may reveal more intricate structure of electoral support.

There is a large number of problems in the field of political analysis, as well as in other domains that can benefit from our methodology. We have now demonstrated that the method works well with raster-based datasets [2, 10], and shapefile-based datasets (this work). In its present form, the method is not intended to provide electoral *prediction*, but future versions will incorporate prediction capabilities via the Classification by the Aggregating Emerging Pattern or CAEP technique [3].

5. REFERENCES

[1] R. Agrawal and A. N. Swami. Fast algorithms for mining association rules. In *In Proc. VLDB*, page 487Ű499, 1994.

[2] W. Ding, T. F. Stepinski, and J. Salazar. Discovery of geospatial discriminating patterns from remote sensing datasets. In *IN Proceedings of SIAM International Conference on Data Mining*, 2009.

[3] G. Dong and J. Li. Efficient mining of emerging patterns: discovering trends and differences. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 43–52, San Diego, California, United States, 1999.

[4] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8(1):53Ű87, 2004.

[5] G. F. Jenks. The data model concept in statistical mapping. *International Yearbook of Cartography*, 7:186–190, 1967.

[6] D. Lin. An information-theoretic definition of similarity. In *International Conference on Machine Learning*, Madison, Wisconsin, July 1998.

[7] L. McQuitty. Similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological Measurement*, 26:825–831, 1966.

[8] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In *ICDT '99: Proceedings of the 7th International Conference on Database Theory*, pages 398–416, 1999.

[9] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18(5):401–409, 1969.

[10] T. F. Stepinski, W. Ding, and C. F. Eick. Controlling patterns of geospatial phenomena. *Geoinformatica*, 2010.

[11] M. Zaki and M. Ogihara. Theoretical foundations of association rules. In *In 3rd ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 1998.