

Understanding Social Welfare Service Patterns Using Sequential Analysis

Hye-Chung (Monica) Kum*⁺

Dean Duncan*

Wei Wang⁺

**Jordan Institute for Families, School of Social Work* ⁺ *Department of Computer Science*
University of North Carolina at Chapel Hill (kum, dfduncan@email.unc.edu, weiwang@cs.unc.edu)

1. Introduction

Classical exploratory data analysis methods in statistics and many of the earlier KDD methods tend to focus only on basic data types, (interval or categorical data) as the unit of analysis. However, some information cannot be interpreted unless the data is treated as a unit. For example, viewing the data as sequences of sets can reveal useful information that could not be extracted in any other way. Analyzing suitable databases from such a perspective can assist many social scientists in their work.

For example, all states have administrative data about who has received various welfare programs and services. Some even have the databases linked for reporting purposes [2]. Table 1 shows examples of monthly welfare services given to clients in sequential form. Using the linked data, it is possible to analyze the pattern of participation in these programs. This can be quite useful for policy analysis: What are some commonly occurring patterns? What is the variability of such patterns? How do these patterns change over time? How do certain policy changes effect these patterns?

2. Sequence Analysis

Conventional methods used in policy analysis cannot answer the broad policy questions such as finding common patterns of practice. Thus, analysts have been forced to substitute their questions. Until recently, simple demographic information (66% of those receiving Food Stamp also received TANF benefits in June) was the predominant method used. Survival analysis is gaining more popularity but still only allows for analyzing the rate of occurrence of some particular events (50% of participants on TANF leave within 6 months). In [2], they studied specific transitions of interest (15% of children who entered TANF in Jan, were in foster care before entry). These methods are very limited in their ability to describe the whole body of data.

Thus, some have tried enumeration – frequency of combinations of the first four events [2]. For example, Client A in Table 1 would be encoded as having experienced “Medicaid only” followed by “TANF” followed by “foster care”. However, the combinatoric nature of simple enumeration does not

give much useful information. In the above example, there were 5 distinct events of interest. Thus, looking at only the first four events, the number of possible patterns would be $5^4=625$. Not surprisingly, there are only a few simple patterns that are frequent. The more complex patterns of interest do not get compressed enough to be comprehensible. The problem is that almost all the frequent patterns are already known and the rest of the information is not very understandable by humans. There were a total of 179 patterns reported in the analysis.

A better method developed in sociology, optimal matching, has not gained much popularity in social sciences [1]. Optimal matching is a direct application of pattern matching algorithms developed for DNA sequencing to social science data [3]. It applies the hierarchical edit distance to pairs of simple sequences of categorical data and runs the similarity matrix through standard clustering algorithms. Then the researcher looks at the clusters to manually organize and interpret them. This is possible because researchers have only used it for fairly small data sets, collected and coded manually.

There are two problems with optimal matching. First, you are limited to strings. Thus, one could not handle multiple services received in one month very well. In real datasets, sequences of sets are much more common than sequence of letters. Second, once the clustering and alignment is done there is no mechanism to summarize the cluster information automatically. Finding consensus strings from DNA sequences have not been applied to social science data yet. Thus, the applicability needs to be investigated. Without some automatic processing to produce cluster descriptors, social scientists would be limited to very small data sets.

ApproxMAP is a new approach to sequential analysis that can detect common patterns and their variations in sequences of sets [4]. It partitions the database into similar sequences, and then automatically summarizes the underlying pattern in each partition through multiple alignment. In this paper, we describe a successful case study in which ApproxMAP was used to detect common patterns of monthly services given over time.

Table 1: Different examples of monthly welfare services given to clients in sequential form

| clientID | Sequential Data |
|----------|--|
| A | {Medicaid (M)} {TANF (A), M} {A, M} {A,M} {M, Foster Care (FC)} {FC} {FC} |
| B | {Report (R)} {Investigation (I), Foster Care (FC)} {FC, Transportation (Tr)} {FC} {FC, Tr} |

3. Case study

We investigated sequential analysis in order to answer a policy question. What are the common patterns of services given to children with substantiated reports of abuse and neglect? What are the variations? Using the data on services given to them, ApproxMAP confirmed much of what we knew about these children. These findings gave us confidence in the results. It further revealed some unknown patterns of interest to the practitioners.

There are three administrative databases used. Most of the data comes from the NC social workers' daysheet data that indicates what services were given when, to whom, and for how long. The data gives a fairly accurate picture on the various services given to clients each month. Therefore, we can convert this data into monthly services given to clients. Then, we identify children who had a substantiated report of abuse and neglect using the abuse and neglect report database. Finally using the foster care database, we can further split the children with substantiated reports into those that were placed in foster care and those that were not.

The interesting patterns were found in children who were placed in foster care because they received more services. Here we report on our results from children with a substantiated report of abuse and neglect that were placed in foster care. There were 992 such children (sequences). Each sequence starts with the substantiated report (RPT) and is followed by monthly services given to each child. The follow up time was 1 year from the report.

In summary, we found 15 interpretable and useful patterns. The most common pattern was $\langle\{RPT\}\{INV,FC\}\{FC\}\{FC\}\{FC\}\dots\{FC\}\{FC\}\{FC\}\rangle$ where INV stands for an 'Investigation', and FC stands for a 'Foster Care' service. In total, 419 sequences were grouped together to generate the above consensus pattern. The pattern indicates that many children who are in the foster care system after getting a substantiated report have very similar service patterns. Within one month of the report, there is an investigation and the child is put into foster care. Once children are in foster care, they stay there for a long time. Recall that 12 months in foster care means the child was in foster care for the full time of analysis. This is consistent with the policy that all reports of abuse and neglect must be investigated within 30 days. It is also consistent with our analysis on the length of stay in foster care. The median is slightly over one year in NC.

The rest of the sequences in this data set split into clusters of various sizes. Another obvious

pattern was the small number of children who were in foster care for a short time. One cluster formed around the 57 children which had the following consensus pattern $\langle\{RPT\}\{INV,FC\}\{FC\}\{FC\}\rangle$.

There were several consensus patterns from very small clusters with about 1% of the sequences. $\langle\{RPT\}\{INV,FC,T\}\{FC,T\}\{FC,HM}\dots\{FC,HM\}\{FC\}\{FC,HM\}\rangle$ was one such pattern of interest. HM stands for 'Home Management' services and T stands for 'Transportation'. There were 39 sequences in the cluster. The practitioners were interested in this pattern because foster care services and home management services were expected to be given as an "either/or" service, but not together to one child at the same time. Home management services were originally designed for those staying at home. Thus, we confirmed the unexpected pattern in the original data. Was this a systematic data entry error or were there some components to home management services that were used in conjunction with foster care services on a regular basis? If so, which counties were giving these services in this manner? Such investigation would not have been triggered without our analysis because no one ever assumed there was such a pattern.

4. Conclusion

As social work researchers collaborating with practitioners, we had an important policy question and the data to answer it: What are the commonly occurring monthly service patterns and their variations? Yet a thorough investigation revealed no existing method could answer such a broad question to our satisfaction. Thus, we engaged in interdisciplinary research with computer science researchers in data mining to explore a new approach to sequential analysis. As a result, ApproxMAP was developed to detect patterns in sequences of sets. Our extensive evaluation on synthetic and real data show that ApproxMAP can effectively detect the underlying patterns with little confounding information. Most importantly, we have successfully used ApproxMAP to answer our original policy question. It found useful, interpretable, and previously unknown patterns in services given to children with substantiated reports of abuse and neglect.

Reference

- [1] A. Abbott & A. Tsay. Seq Analysis and Optimal Matching Meth in Soci: Review & Prospect. In *SMR*. 29:3-33, 2000.
- [2] R. Goerge, et al. Dynamics of Children's Movement among the AFDC, Medicaid, and Foster Care Programs. *Technical report to U.S. Dept. of HHS*. 2000.
- [3] Dan Gusfield. *Algo. on strings, trees, and sequences: Comp Sci and Comp. Bio*. Cambridge Univ. Press. 1997.
- [4] H. C. Kum, et al. ApproxMAP: Approximate Mining of Consensus Sequential Patterns. *SIAM-DM*. SF, May 2003.