

Social Welfare Program Administration and Evaluation and Policy Analysis Using Knowledge Discovery and Data Mining (KDD) on Administrative Data

Hye-Chung (Monica) Kum* Dean Duncan** Kimberly Flair** Wei Wang*
University of North Carolina at Chapel Hill

**Department of Computer Science & Jordan Institute for Families (kum@cs.unc.edu)*

*** Jordan Institute for Families, School of Social Work (dfduncan, kaflair@email.unc.edu)*

Project URL: <http://ssw.unc.edu/workfirst/research.html>

Abstract

New technology in knowledge discovery and data mining (KDD) make it possible to extract valuable information from operational data. Private businesses already use the technology for better management, planning, and marketing. Social welfare government agencies have a wealth of information about the experiences of families and individuals that are the most needy in our society in their administrative databases. These data too can be mined and analyzed with proper application of KDD technology. Such social science research could be priceless for better welfare program administration, program evaluation, and policy analysis. In this paper, we discuss a successful case study involving research in computer science as well as social welfare. In a long standing collaboration between the North Carolina DHHS and the University of North Carolina, we have (1) successfully built a longitudinal information system that tracks the experiences of families and individuals on welfare in NC since 1995 (2) developed a dynamic website reporting on the various aspects of the welfare program at the county level in order to assist county staff in the administration of the welfare program and (3) developed a new method to analyze sequential data, which can detect common patterns of welfare services given over time.

1. Introduction

Our society is accumulating massive amounts of data, much of which resides in large database management systems (DBMS). Methods to explore such data would stimulate research in many fields. Knowledge discovery and data mining (KDD) is the area of computer science that tries to generate an integrated approach to extracting valuable information from such data by combining ideas drawn from databases, machine learning, artificial intelligence, knowledge-based systems, information retrieval, statistics, pattern recognition, visualization, and parallel and distributed computing. It has been defined as "The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"(Fayyad, 1996). The goal is to discover and present knowledge in a form, which is easily comprehensible to humans (Fayyad, 1996). We add one more step to this definition; timely and efficient delivery of the knowledge discovered.

A key characteristic particular to KDD is that it uses "observational data, as opposed to experimental data" (Hand, 2001). That is, the objective of the data collection is something other than KDD. Usually, it is operational data collected in the process of operation, such as payroll data. Operational data is sometimes called administrative data when it is collected for administration purposes in government agencies. This means that often times the data is a huge convenience sample. Thus, in KDD attention to the large size of the data is required and care must be given when making generalizations. In statistics, such analysis is called secondary data analysis (Hand, 2001).

Many private businesses have successfully applied KDD to operational data for better management, marketing, and planning. Some examples are retail stores finding relationships between products to design better marketing, product layout, and sales plans or mail marketing companies better targeting their solicitations using their operational data. In fact, almost all of the research in KDD is fueled by business applications.

In this paper, we discuss the huge potential for research in both KDD and social welfare. Social welfare program administration, program evaluation, and policy analysis are important but difficult functions of the numerous state and federal Department of Health and Human Services (DHHS). These agencies have huge administrative databases about the experiences of families and individuals that are the

most needy in our society. These databases can be mined for valuable information. Moreover, these administrative data can give insight on the effects of various social welfare policies. In this paper, we discuss a successful case study of using the administrative data collected by North Carolina DHHS for program administration, program evaluation, and policy analysis.

2. Case Study : Collaboration between North Carolina DHHS and UNC-CH

In this paper, we discuss a successful case study to build an information system using the administrative data from the North Carolina DHHS. In particular, we highlight two accomplishments from the project. First, we have developed a dynamic website reporting on the various aspects of the Work First program, the monthly cash assistance program in NC, at the county level. It is driven by monthly updates of summarized data extracted from various administrative data. Such accurate, fast, and comprehensive dissemination of detailed information to county officials on their welfare program is essential for proper administration of the Work First program at the county level. Second, we have developed a new method to analyze sequential data, which can detect common patterns of welfare services given over time.

The NC DHHS accumulates huge amounts of data on who is receiving which welfare services (Work First, Medicaid, Food Stamps, etc.) and is responsible for disseminating relevant information from these administrative data on the status of the welfare programs to various stakeholders such as:

- the legislature for informed policy making
- federal government agencies for monitoring purposes
- county officials for proper administration of services at the county level
- researchers, journalists, and the general public as a matter of public information

However, all the administrative data are on mainframe DBMS designed for optimal operation, providing the services, not for information gathering or dissemination. Thus, it is extremely time consuming and expensive to access the administrative data. It is difficult to obtain useful and accurate information dynamically and quickly from these systems. In fact, often times DHHS cannot respond in a timely manner to even the basic questions using the raw data unless it is a canned query. Furthermore, it is impossible to keep all information over time, thus the data in these systems are updated and maintained to provide current services appropriately. Consequently, no analysis can be done over time directly from these data without archiving it. State agencies with huge administrative data, such as DHHS, could benefit immensely from proper deployment of KDD technology. Yet, lacking in funds even today many state social welfare agencies have not become properly IT enabled.

In 1997, federal funds were available for evaluating the cash assistance welfare program under Welfare Reform. The North Carolina DHHS used the funds to contract with the Jordan Institute For Families at the School of Social Work at the University of North Carolina at Chapel Hill (UNC-CH). The objective of the contract was to properly evaluate and track the Work First program. Work First is the cash assistance welfare program, also called TANF, in North Carolina. As a result, the Jordan Institute built an information system for effectively archiving and tracking various data related to the Work First Program using the administrative data.

The Work First project started with the development of an information system that links files across multiple administrative databases in order to provide comprehensive evaluation measures for the Work First program. One of the challenges was to design a datawarehouse that could properly use the semantics of the administrative data but also is suitable for analysis of diverse policy and program questions. For example, to determine participation in Work First, the check history database and the application database is used. First, the check history database is used to determine all cases that received a check in any particular month. Then by linking this information with the application database, we are able to separate out diversion cases from the regular participants. Diversion cases are those that are assessed to need only short-term assistance in an emergency situation. These cases are diverted from entering Work First by giving them a one-time cash assistance. This participation information is then transformed into a longitudinal database (monthly participation over time) that can track changes over time. Furthermore, since welfare programs in NC are state supervised and county administered, it was important to be able to do customized analysis and share the results with all 100 counties in NC.

The Institute successfully built a longitudinal information system that tracks the experiences of families and individuals on welfare in NC since 1995. It further allows for the analysis of Work First program outcomes at different levels ranging from the individual recipient to the entire county and state caseloads. Thus, county program managers and staff are able to explore outcomes for participants in their own county. Workers and supervisors are able to track the impact of Work First services on their clients over time. In addition, county directors are able to gauge their performance in the Work First program and compare their county with other similar counties or the entire state as well as across time.

Once the system was built at the Jordan Institute, DHHS could do many of its tasks much more effectively and accurately by contracting out various tasks to the Jordan Institute. The collaboration has lasted to date for over six years and the information system has expanded considerably to include data on Food Stamps, Medicaid, and Foster Care.

Although the main objective and tasks of the collaboration was data analysis, the Jordan Institute has been able to build an extensive infrastructure for:

- large scale data acquisition and management for tracking various welfare programs
- long term archiving of the numerous administrative datasets acquired (some up to 8 years)
- IT-enabled efficient communication between state (DHHS) and county agencies (100 county department of social services in NC)
- Effective dissemination of government information to the public as well as the county agencies via the Internet.

2.1 Management Assistance for Work First via a Dynamic Website

Once the information system was built, the next challenge was to efficiently disseminate information to the county officials. The ability to access information at the county level was key because in NC, the welfare program is state supervised and county administered. All 100 county departments of social services administer their program separately. Before our project, although the individual counties were responsible for administering the programs, it was quite difficult for county officials to access their county information maintained at the state. The counties had access to only a few common canned queries.

In fact, such a gap between the funding source and service delivery is common in social work. Much of the funds for social services come from federal, state, or foundation sources. However, the most effective development and delivery of social services have to be at the local level appropriate to the local conditions. Proper use of information technology can provide a means of effective and comprehensive communication between the service providers and funders with minimal effort. Furthermore, the capacity to easily customize the content of the communication to the needs of the users make it even more appealing since the information requirements for the funder and service provider are vastly different.

In our project, we built a dynamic website reporting on the various aspects of the Work First program at the county level. It is driven by quick monthly updates of summarized data extracted from multiple administrative databases. County officials as well as the general public can access numerous updated county and state level reports on the North Carolina Work First program from our public website. This allows the county directors to understand the current status of their local program in a timely manner and plan accordingly. In addition, they can easily access information on other similar counties, the entire state, and their past information for comparison.

For example, the racial composition of the welfare recipients are different across counties and has changed significantly over the 8 years we have tracked the program (Figure 1). Graphs from the website indicate that the urban counties (Chatham) have a much higher percentage of African Americans than the rural counties (Dare). Moreover, regardless of the location of the county, the percentage of Hispanic participants has risen significantly suggesting the need of Spanish speaking caseworkers.

In addition, counties can see how changes in policy affect the program. As an example, the graph showing the two-parent caseload has a significant drop in March of 1998 (Figure 1). Welfare families with two parents are relatively rare. They constitute less than one percent of the welfare caseload. These families are watched closely, since the federal welfare reform law places stricter work requirements on

them. Therefore, the state implemented a policy of pay after performance, which requires most two-parent families to participate in work or job training activities before receiving cash benefits. This policy went into effect in March 1998 causing the sudden drop in two-parent cases.

The demo paper describes the website in more detail. When welfare programs are state supervised and county administered, properly managing welfare reform initiatives require efficient and timely communication between the state and the 100 counties in NC. A comprehensive set of recent data analysis customized to each county would assist both the state and the counties to better do their jobs. The dynamic website that tracks the experiences of families and individuals on welfare in NC by county can assist county program managers in evaluating their local program. The information technology not only provides a means for efficient dissemination of information, but by presenting complex information in an intuitive, easy to use manner, it also shows social workers a new way of thinking. The information provided through the website has resulted in many counties evaluating their performance in terms of outcomes. It has provided a common set of measures that can be compared over time and across counties.

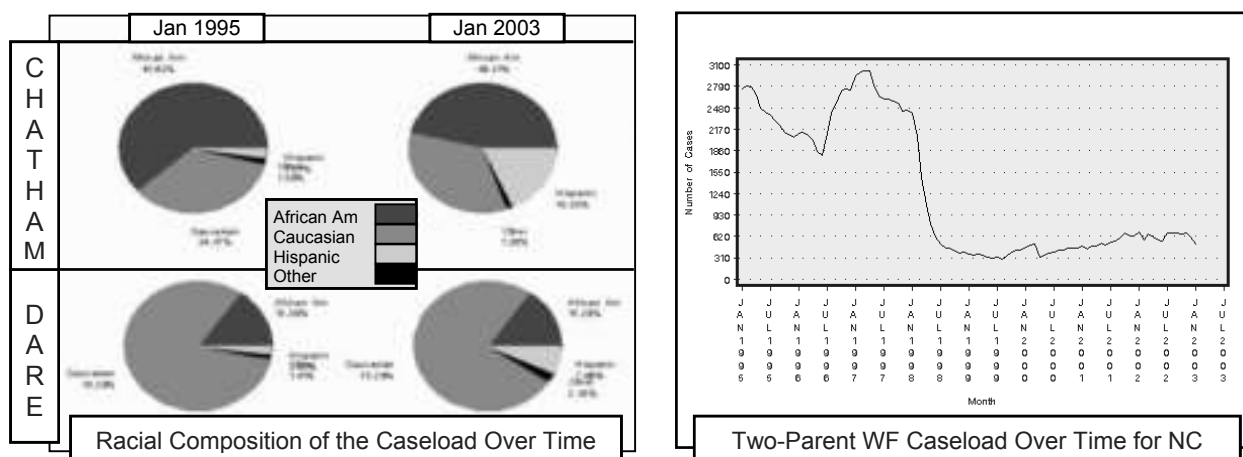


Figure 1. Graphs from the dynamic website on Work First Performance

2.2 Approximate Mining of Consensus Sequential Patterns

The main goal of the project was data analysis; so much of the technology deployed in this project was limited to existing technology that can accomplish the task as quickly and as best as possible. In fact, the website used existing SAS/IntrNet technology and the research achievements were mostly in the areas of social work. We did not have the resources to research in depth the many long-term technical issues that we encountered. However, no existing methods for analyzing sequential data were satisfactory. Thus, we had to develop a new method to analyze sequential data.

Data in the form of sequences are ubiquitous in social work. Proper analysis of many of the complex administrative data required exploring sequences of sets as a unit. For example, service pattern analysis tries to collect and analyze data on various welfare services given monthly over time (i.e. {investigation, foster care} {foster care} {foster care, transportation}). The goal is to identify common patterns and its variations as well as the subgroups that follow such patterns. That would allow us to start understanding common pattern of services, its variations, and who received them. Thus, there was a strong need for methods to analyze sequence of sets. However, conventional methods in social science, computer science, and computational biology had inherent difficulties in mining databases with long sequences and noise.

As proposed by Agrawal and Srikant (1995), a sequential pattern is a subsequence frequently appearing in the database. Mining sequential patterns from large databases has been studied extensively in the last few years. Many efficient mining algorithms have been proposed (Kum, 2002). However, all these methods work based on exact match that makes it impractical for this application. For sequential pattern mining to be effective and meaningful for social work, we need to find the *approximate sequential pat-*

erns, patterns approximately shared by many sequences, and present patterns only when they are coherently shared by a group of similar sequences. None of the conventional methods can do this (Kum, 2002).

We researched the theme of approximate sequential pattern mining and designed an efficient and effective algorithm, **ApproxMAP** (**APPROXimate Multiple Alignment Pattern mining**), to mine consensus patterns from large sequence databases (Kum, 2003). Our goal is to assist the data analyst in exploratory data analysis through organization, compression, and summarization. The method has three steps.

- 1) First, similar sequences are grouped together using kNN clustering.
- 2) Then we organize and compress sequences within each cluster into *weighted sequences* using multiple alignment.
- 3) In the last step, the information in each cluster is summarized into the longest *consensus pattern* best fitting each cluster from the *weighted sequences* using the user specified *strength* threshold. We use color to visualize item strength, how many sequences in the cluster contain the item.

We successfully used **ApproxMAP** to analyze the daysheet data, a database of monthly welfare services given to clients in NC, for children who had a substantiated report of abuse and were subsequently placed in foster care. In summary, we found 15 interpretable and useful consensus patterns. For example, about half of the children followed the typical behavior with consensus pattern {RPT} {INV,FC} {FC} {FC} {FC} {FC} {FC} {FC} {FC} {FC} {FC} {FC} {FC} {FC}, where RPT is Report, INV is Investigation, and FC is Foster Care Services. The pattern indicates that typically within one month of the report, there is an investigation and the child is put into foster care. Once children are in the foster care system, they stay there for a long time. This is consistent with the policy that all reports of abuse and neglect must be investigated within 30 days and with our analysis on the length of stay in foster care. For the first time we were able to start looking at and understanding service patterns for those in foster care.

We also used **ApproxMAP** to study service patterns for those in the Willie M program. The program served a certain class of children, who became state certified as being severely or chronically aggressive. **ApproxMAP** is used to examine the complex combinations of services given to different groups of children classified as Willie M in order to determine common service patterns for each group.

Our extensive experimental results on synthetic and real data show that **ApproxMAP** is very robust to noise and does well in mapping the high dimensional noisy data into approximate sequential patterns. Our study illustrates that approximate sequential pattern mining can find general, useful, concise and understandable knowledge and thus is a promising direction for sequential analysis (Kum, 2003).

2.3 Discussion

The extensive information system has been and continues to be expanded and updated for over six years. In that time, the project has assisted with numerous mandated federal reports (such as the high performance bonus report), provided ad-hoc technical reports, presentations, and data analysis on various topics to DHHS (such as child only caseload, earnings of Work First participants, and migration patterns), and supported three Ph.D. dissertations. The two dissertations in social sciences used the information system for policy analysis while one in computer science developed **ApproxMAP** discussed in section 2.2. In addition, a temporary dynamic website similar to the Work First website was built for Medicaid in order to comply with a federal mandate to review a group of individuals who might not have been properly screened. This required that DHHS identify the group, then request that counties review the group and report back on the status. The project lasted for a year, and in addition to disseminating information efficiently through the Internet, DHHS was able to gather county input efficiently as well.

As the data we are archiving is sensitive data, privacy is of utmost importance to the project. We have taken all precautions to limit access to the personal data and are looking into new technology for more privacy while still maintaining information sharing. The databases used to track the experiences of individuals currently contain only encrypted identification codes. Information from these databases is aggregated into smaller datasets that are used by the web site. In the process of aggregation, if a county or grouping has less than ten individuals that have a particular attribute (e.g., Asian-Americans participating in a particular month), that category is merged with another group (e.g., "Other").

3. Conclusion

Administrative data does not have all the answers, but with proper mining and analysis, there is much we can learn about the current policies and programs from them. The long close collaboration between NC DHHS and UNC-CH demonstrates that proper deployment of KDD technology on administrative data is possible and beneficial to both parties. DHHS has been able to meet many of its information needs more accurately, efficiently, and timely while UNC-CH has been able to use this experience and the information system to do research in both social work and computer science. The infrastructure has potential for much more research in both areas.

- **Step 0:** Given the sequence database in Table 1.
- **Step 1:** ApproxMAP first divides the sequences into 2 clusters (k was set to 2 for kNN clustering).
- **Step 2:** Then ApproxMAP aligns the sequences in each cluster and compresses each cluster into one weighted sequence per cluster as given in Tables 2 and 3.
- **Step 3:** Finally, ApproxMAP generates a consensus pattern for each cluster using the weighted sequence. Cluster strength was set to 40%. Thus, the consensus pattern is generated by selecting all items in the weighted sequence that have a weight more than 40% of the cluster (Table 4). Color is used to visualize item strengths.

ID	Sequences
seq1	{A} {B C Y} {D}
seq2	{A} {X} {B C} {A E} {Z}
seq3	{A I} {Z} {K} {L M}
seq4	{A L} {D E}
seq5	{I J} {B} {K} {L}
seq6	{I J} {L M}
seq7	{I J} {K} {J K} {L} {M}
seq8	{I M} {K} {K M} {L M}
seq9	{J} {K} {L M}
seq10	{V} {K W} {Z}

Table 1. Sequence database

seq1	{A}	{}	{B, C, Y}	{D}	{}	
seq4	{A, L}	{}	{}	{D, E}	{}	
seq2	{A}	{X}	{B, C}	{A, E}	{Z}	
Weighted sequence	{A:3, L:1}:3	{X:1}:1	{B:2, C:2, Y:1}:2	{A:1, D:2, E:2}:3	{Z:1}:1	3
Consensus sequence $\{w \geq 2\}$	{A}		{B, C}	{D, E}		
Weighted Consensus sequence $\{w \geq 2\}$	{A:3}:3		{B:2, C:2}:2	{D:2, E:2}:3		3

Table 2. Cluster 1 {min_strength = 40% = 1.2 ≤ 2 sequences}

seq9	{J}	{}	{K}	{L, M}	{}	
seq5	{I, J}	{B}	{K}	{L}	{}	
seq3	{A, I}	{Z}	{K}	{L, M}	{}	
seq7	{I, J}	{K}	{J, K}	{L}	{M}	
seq8	{I, M}	{K}	{K, M}	{L, M}	{}	
seq6	{I, J}	{}	{}	{L, M}	{}	
seq10	{}	{V}	{K, W}	{}	{Z}	
Weighted sequence	{A:1, I:5, J:4, M:1}:6	{B:1, K:2, V:1, Z:1}:5	{J:1, K:6, M:1, W:1}:6	{L:6, M:4}:6	{M:1, Z:1}:2	7
Consensus seq $\{w \geq 3\}$	{I, J}		{K}	{L, M}		
Wgt Consensus seq	{I:5, J:4}:6		{K:6}:6	{L:6, M:4}:6		7

Table 3. Cluster 2 {min_strength = 40% = 2.8 ≤ 3 sequences}

Pattern Consensus Seq 1	support = 40% = 1.2 2 sequences	{A}{B, C}{D, E}
Variation Consensus Seq 1	Not appropriate in this small set	
Pattern Consensus Seq 2	support = 40% = 2.8 3 sequences	{I, J} {K} {L, M}
Variation Consensus Seq 2	support = 20% = 1.4 2 sequences	{I, J} {K} {K} {L, M}

Table 4. Consensus sequences (100%: 85%: 70%: 50%: 35%: 20%)

Figure 2. An example of ApproxMAP

Reference

- [1] R. Agrawal and R. Srikant. "Mining Sequential Patterns". In Proc. IEEE ICDE, March 1995.
- [2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. The KDD process for extracting useful knowledge from volumes of data. In Communications of ACM Vol 36 Issue 11, pages 27-34. NY, NY: ACM Press, Nov. 1996.
- [3] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. Cambridge, MA: MIT Press, 2001.
- [4] Hye-Chung (Monica) Kum, Susan Paulsen, and Wei Wang. (2002). Comparative Study of Sequential Pattern Mining Frameworks: Support Framework vs. Multiple Alignment Framework. *Proc. of the 2002 IEEE ICDM Workshop on The Foundation of Data Mining and Discovery*. Maebashi, Japan, Dec 2002.
- [5] Hye-Chung (Monica) Kum, Jian Pei, Wei Wang, and Dean Duncan. (2003). ApproxMAP: Approximate Mining of Consensus Sequential Patterns. *Proc. of the 3rd SIAM Intl. Conf. on Data Mining*. San Francisco, CA, May 2003.