응용통계연구 제 12권 1호, 1999년, pp. 83-95

인터넷상에서의 범주형 자료분석 시스템 개발 *

홍종선[1] 김동욱[2] 오민권[3]

요 약

본 논문의 목적은 인터넷에서 범주형 자료분석에 대한 전문적인 지식이 없는 일반 분석자들에게 보다 쉽고, 간편하게 다룰 수 있는 범주형 자료 분석 시스템을 제공하는 것이다. 이 분석 시스템은 크게 세 가지 측면으로 설계하여 구현하였다. 첫째, 범주형 자료에 대한 탐색적 자료분석을 위하여 세 가지 종류의 히스토그램을 제공한다. 둘째, 범주형 변수들간에 존재하는 연관성을 측정하기 위한 여러 연관성 측도들을 제공한다. 특히, 현재 많이 사용되는 통계 패키지들에서 제공하지 못하는 모자익 그림과 연관 그림을 동적 그래픽스로 구현하여 연관성을 측정하거나 모형을 설정하는데 유용한 정보를 얻을 수 있도록 하였다. 셋째, 대수선형모형에 대한 분석을 통해 사용자가 가장 잘 적합된 대수선형모형을 선택할 수 있게 하였다.

1. 서론

범주형 자료를 분석할 수 있는 윈도우 환경의 통계 패키지들로는 SAS, SPSS, S-PLUS, STATISTICA, SYSTAT등이 있다. 그러나 사용자가 이들 패키지들을 이용하여 범주형 자료를 분석하기 위해서는 사전에 범주형 자료분석에 관한 많은 지식을 알고 있어야 하고, 패키지의 사용법도 어느 정도 알아야 한다. 이들 패키지들은 비슷한 형태의 결과를 제공하고 있지만, 범주형 자료에서 변수들의 연관성을 시각적으로 표현하는 모자익(mosaic) 그림은 Antony Unwin과 Martin Theus에 의해 개발된 MANET, SAS의 IML, INSIGHT 그리고 JMP 소프트웨어에서만 제공되고 있다. 그러나 이들 패키지에서 제공되는 모자익 그림은 동적 그래픽스로 구현되어 있지는 않다. 무엇보다도, 오늘날과 같은 정보화 시대에 월드 와이드 웹에서 사용자가 직접 범주형 자료를 분석할 수 있는 통계 패키지는 SAS와 SPSS사의 다이아몬드만이 분석을 지원하고 있는 실정이다. 본 논문에서는 범주형 자료분석에 대한 전문 지식이 없는 비특정 다수의 인터넷 사용자들이 보다 쉽고, 간편하게 범주형 자료를 분석할 수 있을 뿐만 아니라, 범주형 자료분석에 관한 내용을 학습할 수 있는 새로운 범주형 자료분석 시스템을 제안하고자 한다.

* 이 논문은 1997년도 한국학술진흥재단 자유공모 연구비에 의하여 지원되었으며, 성균관대학교의 1996년도 63 학술연구비에 의하여 연구되었음. 본 논문에서 구현된 범주형 자료분석 시스템의 사용을 원할 경우 Explorer 4.0 이상을 이용하여 http://stat.skku.ac.kr/~cshong/Categorical.html을 접속하면 된다.

1) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 교수
2) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 조교수
3) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 강사

2. 인터넷에서의 범주형 자료분석 시스템

본 연구의 목적은 인터넷상에서 웹 브라우저를 통해 사용자가 직접 범주형 자료를 분석할 수 있는 분석 시스템을 구현하는 것이다. 구현언어는 모든 컴퓨터 기종에서 사용 가능한 웹 프로그래밍 언어인 JAVA(JDK 버젼 1.1.5)를 사용하였으며 GUI(graphical user interface), DDE(dynamic data exchange), OLE(object linking and embedding)의 특성을 갖는 새로운 범주형 자료분석 시스템을 구현하기 위하여 다음과 같은 전략들을 고려하였다.

- 어떤 기종의 컴퓨터에서도 사용할 수 있는 시스템,
- 다양한 탐색적 자료분석 수행,
- 인터넷을 이용한 즉각적인 분석 수행,
- 다이나믹 그래픽스를 구현,
- 사용하기 쉽고 간편하도록 구현,
- 즉각적인 도움말 기능과 학습기능의 강화.

시스템을 보다 효율적으로 구현하기 위하여 이러한 전략들을 다음과 같은 세 가지의 구성 요소로 나누어 설계하였다.

- 분석하고자 하는 범주형 자료의 특성을 그림으로 표현하는 방법,
- 범주형 변수들간에 존재하는 연관성을 측정하는 방법,
- 주어진 분할표를 잘 설명할 수 있는 모형을 선택하고 설정하는 방법.

이러한 세 개의 구성 요소를 8개의 윈도우(File, EDA, Association, Cross Table, Mosaic, Graphical Models, Modelling, Help)로 구성하였으며, 시스템의 인터페이스는 보다 쉽고, 효과적으로 설계되었다. 즉, JAVA 언어의 장점중의 하나인 객체지향성을 최대한 활용하여 8개의 윈도우 및 각 윈도우에서 제공되는 분석내용들을 각각 객체로 구현하였으며, 사용자가 선택한 항목에 해당되는 객체(JAVA 클래스)가 실행되도록 모든 객체들이 서로 연결(link)되어있다. 시스템의 각각의 윈도우를 세부적으로 소개하면 다음과 같다.

2.1. FILE 윈도우

세 개의 패널(panel)을 갖는 그림 2.1은 시스템의 File 윈도우이다. 첫 번째 패널은 사용자에게 자료를 입력하는 방법을 자세하게 설명하고 있으며, 두 번째 패널은 사용자가 변수 이름, 각 변수의 수준, 자료를 직접 입력하는 텍스트 필드(text field)이다. 세 번째 패널은 사용자가 입력한 자료를 확인할 수 있도록 시스템에서 자동으로 그려주는 분할표로서 사용자는 자료가 올바르게 입력 됐는지를 확인할 수 있다. 그림 2.1에 제시된 윈도우의 두 번째 패널에 입력된 자료는 변수가 물의 부드러움 정도(Soft), 상표 선호도(Brand), 상표의 과거 사용 여부(Use), 물의 온도(Temp)인 세제상표 선호도에 대한 4차원 자료(3x2x2x2)로써 대수선형모형(홍종선(1995), 179쪽)에서 인용하였으며, 세 번째 패널에서 확인할 수 있다.

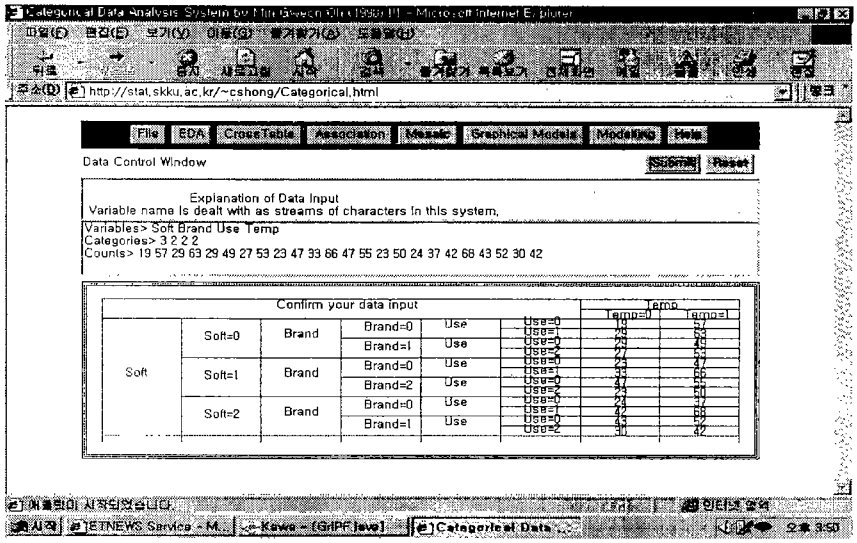

그림 2.1: FILE 윈도우

2.2. EDA 윈도우

통계 패키지에서 탐색적 자료분석의 여러 그림들은 자료의 성격을 파악하는데 매우 중요한 역할을 한다. 이 시스템에서는 범주형 자료의 전체 비율에 대한 히스토그램과 열과 행에 대한 히스토그램을 제공하고 있다. 그림 2.2에는 입력된 자료에서 사용자가 두 변수(Soft와 Brand)에 대한 전체 히스토그램을 선택한 경우의 예제로써 7개의 히스토그램이 나타나 있다. 윗부분의 4개의 히스토그램은 Soft 변수를 기준으로 Brand 변수의 각 수준과 전체에 대한 비율의 히스토그램이고, 아래 3개의 히스토그램은 Brand 변수를 기준으로 Soft 변수의 각 수준과 전체에 대한 히스토그램이다.

2.3. Cross Table 윈도우

Cross-table 윈도우에서 사용자는 입력된 자료에서 임의로 선택한 두 변수에 대한 분할표와 연관성 측도를 알아볼 수 있다. 그림 2.3의 왼쪽에는 사용자가 선택한 Soft 변수와 Brand 변수에 대한 이차원 분할표가 나타나 있고 오른쪽에는 11(2x2 분할표가 아니므로 Odds-ratio 통계량값이 나타나 있지 않다)개의 연관성 측도가 나타나 있다. 연관성 측도들에 관한 자세한 내용은 Association 윈도우를 참조하면 된다.

2.4. Association 윈도우

Association 윈도우는 11개의 연관성 측도에 관한 자세한 내용을 학습할 수 있으며, 선택된 변수에 대한 연관성 측도들의 추정값이 나타난다. 시스템에서 제공하는 11개의 연관성 측도는 Pearson's X^2, G^2, Yate's corrected X^2, Cramer's V, Gamma, Kendall's tau-b, Somer's

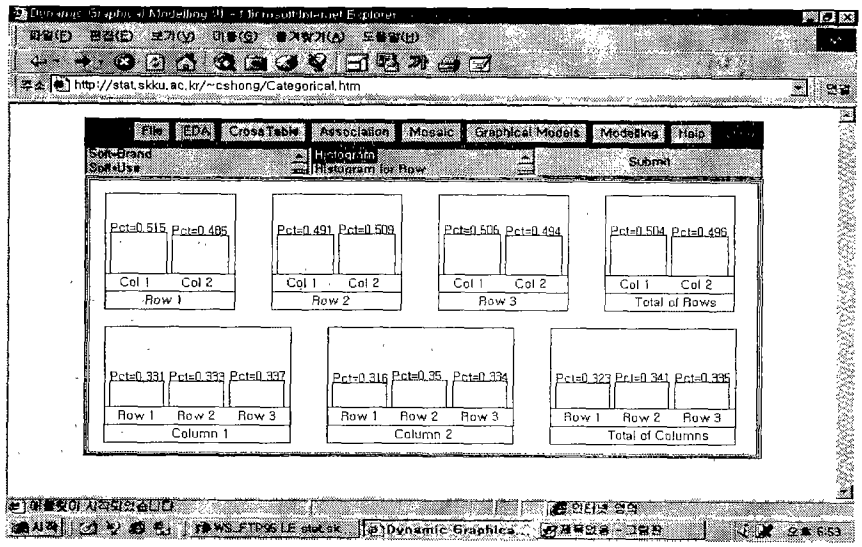

그림 2.2: EDA 윈도우

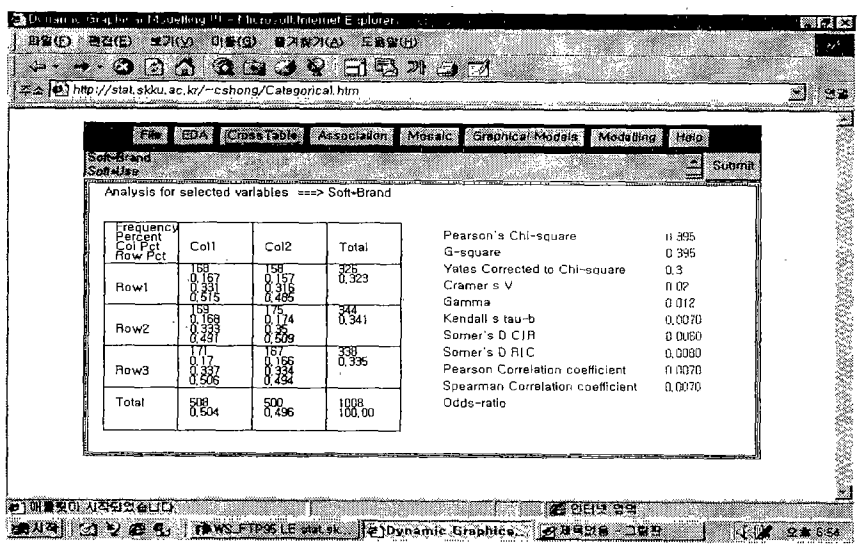

그림 2.3: Cross Table 윈도우

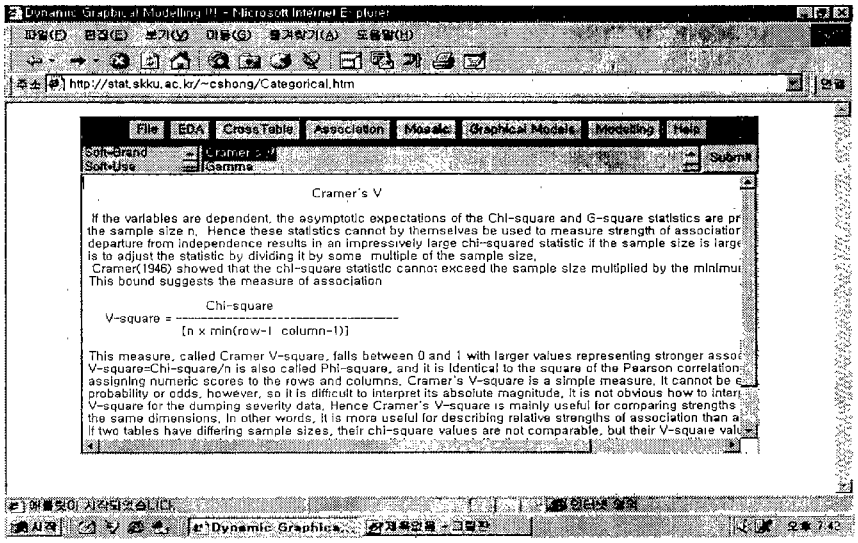

그림 2.4: Association 윈도우

D C—R, Somer's D R—C, Pearson Correlation Coefficient, Spearman Correlation Coefficient, Odds-ratio이다(보다 자세한 내용은 Agresti(1984, 1990, 1996), Christensen(1990), Fienberg(1980) 와 Plackett(1981)를 참조하기 바란다). 그림 2.4는 Association 윈도우에서 사용자가 Cramer's V를 선택한 경우의 출력 결과이다.

2.5. Mosaic 그림 윈도우

Hartigan과 Kleiner(1981)는 분할표의 각 칸의 도수를 정사면체의 면적으로 표시한 모자익 그림을 제안하였다. Freindly(1992, 1994)는 모자익 그림을 대수선형모형을 적합시키는 도구로 확장시켰으며, 다차원인 경우 모자익 그림의 식별이 어렵고 두 변수들간의 관계를 구체적으로 파악할 수 없다는 단점을 보완하기 위해서 각 타일(tile)의 연관관계와 피어슨 X^2의 각 칸에 해당하는 편차의 크기를 고려하여 색상과 빗금을 친 모자익 그림을 제안하였다. 구현된 시스템에서는 삼차원 이상의 분할표를 두 변수로 이루어진 새로운 이차원 분할표로 재구성하여 모자익 그림을 제공한다. 사용자는 다차원 분할표인 경우 모자익 그림 윈도우의 왼쪽 상단부에서 두 변수를 선택하여 모자익 그림을 그려봄으로써 두 변수들간의 독립성 여부를 파악할 수 있다. 그림 2.5은 동적 그래픽스로 구현된 모자익 그림 윈도우로서 사용자가 선택한 Soft 변수와 Brand 변수에 대한 모자익 그림의 출력 결과이다. 왼쪽의 그림은 입력된 자료에 대한 모자익이고 오른쪽의 그림은 Soft 변수와 Brand 변수가 독립이라는 가정 하에서 그린 모자익이다. 만일 분석자가 왼쪽의 관찰값에 대한 모자익 그림의 임의의 타일 위에 마우스를 클릭(click)하여 끌면(dragging), 각각 6개의 타일로 이루어진 두개의 모자익 그림의 각 타일의 크기가 변화된다. 이때 변화된 모자익 그림에 대한 도수, 행과 열의 비율, 전체 비율이 실시간에 변화되기 때문에 사용자는 Soft 변수와 Brand

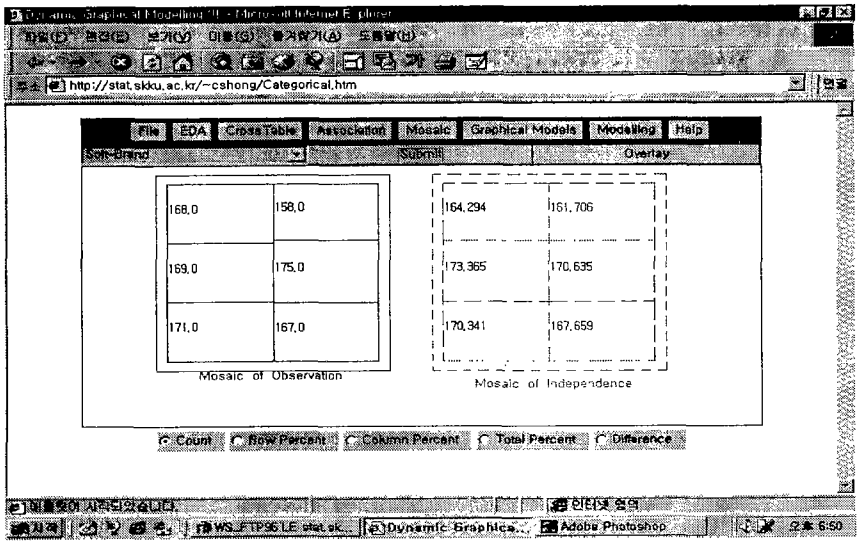

그림 2.5: 모자익 그림 윈도우

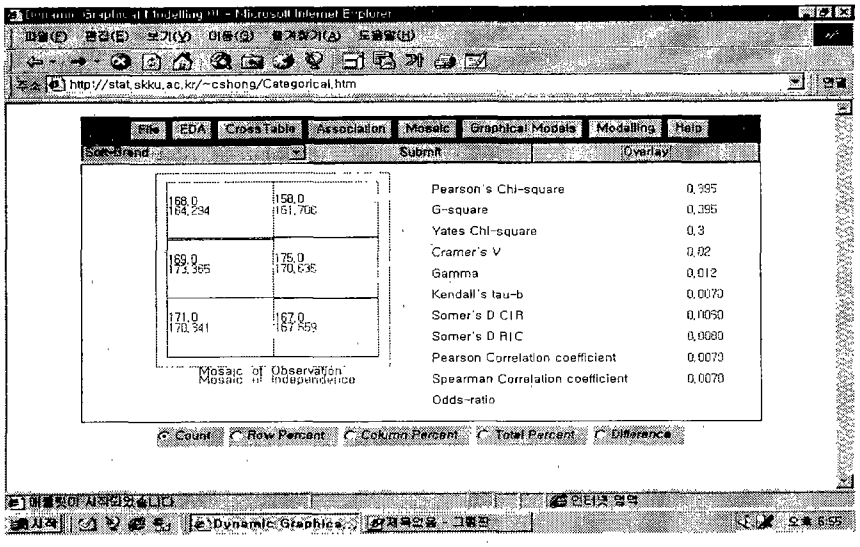

그림 2.6: 겹친 모자익 그림

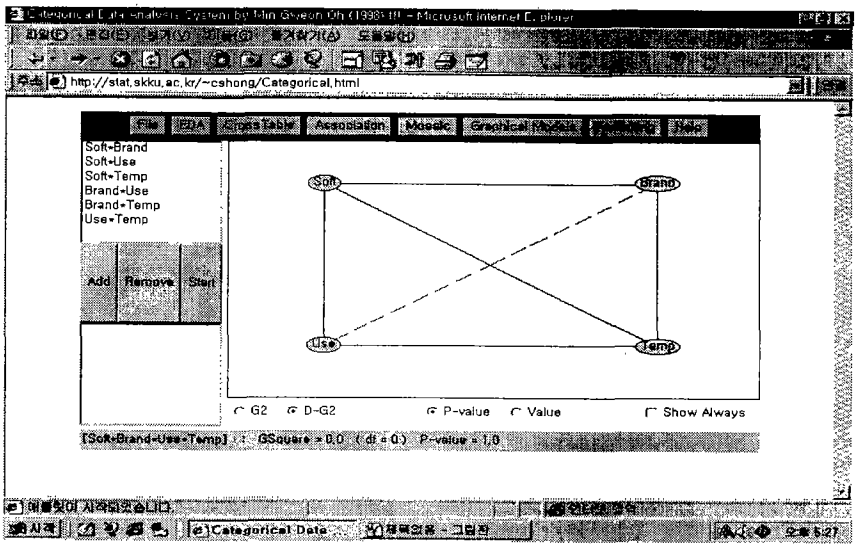

그림 2.7: Graphical Models 윈도우

변수의 독립여부를 동적으로 파악할 수 있다. 또한 그림 2.6과 같이 사용자가 Overlay버튼을 누르면, 왼쪽에는 관찰값에 의한 모자익 그림과 선택된 두 변수가 독립이라는 가정 하에서의 모자익 그림이 겹쳐져 나타나고 오른쪽에는 11개의 연관성 측도(2x2 분할표가 아닌 경우 odds-ratio 통계량값은 나타나지 않음)들이 나타난다. 겹쳐진 두개의 모자익 그림을 통해 사용자는 선택된 두 변수가 독립인 것으로부터 얼마나 멀리 떨어져 있는지를 쉽게 파악할 수 있을 뿐만 아니라 특정한 타일에서 마우스로 타일의 크기(관찰 도수)를 증가 시키거나 감소 시켜가면서 변화되는 11개의 연관성 측도와 관찰도수(행과 열의 비율, 전체 비율)를 통해 선택된 두 변수가 독립이려면 관찰값이 어느 정도여야 되는지도 파악할 수 있다. 이러한 동적 그래픽스로 구현된 모자익 윈도우를 통해서 사용자는 선택된 두 변수의 연관정도 뿐만 아니라 관찰값이 독립인 것으로부터 얼마나 멀리 떨어져 있는지, 각 관찰값이 어느 정도이면 독립이 되는지도 알 수 있기 때문에 모형을 설정하는데 아주 유용한 정보를 제공한다고 할 수 있다.

2.6. Graphical Models 윈도우

그래픽칼 모형(Graphical Models) 윈도우에서는 주어진 범주형 변수들로 구성된 그래픽칼 모형 (자세한 내용은 Goodman 와 Kruskal(1979), Edwards(1995), Haberman(1974) 와 홍종선(1995) 참조)중에서 자료에 대한 최적의 모형을 찾을 수 있다. 그림 2.7의 왼쪽에는 입력된 자료들의 모든 일차 교호작용항들을 나열하였으며, 우선 모든 일차 교호작용항들이 포함되는 그래픽칼 모형 ([Soft*Brand*Use*Temp])에 대한 연관그림(Association plot)을 오른쪽에 제시하였다. 이 연관그림에서 유의한 두 변수(일차 교호작용항)는 점선(시스템에서는 빨간 실선)으로 나타난다. 그래픽칼 모형 윈도우의 하단부에는 설정된 그래픽칼 모형식

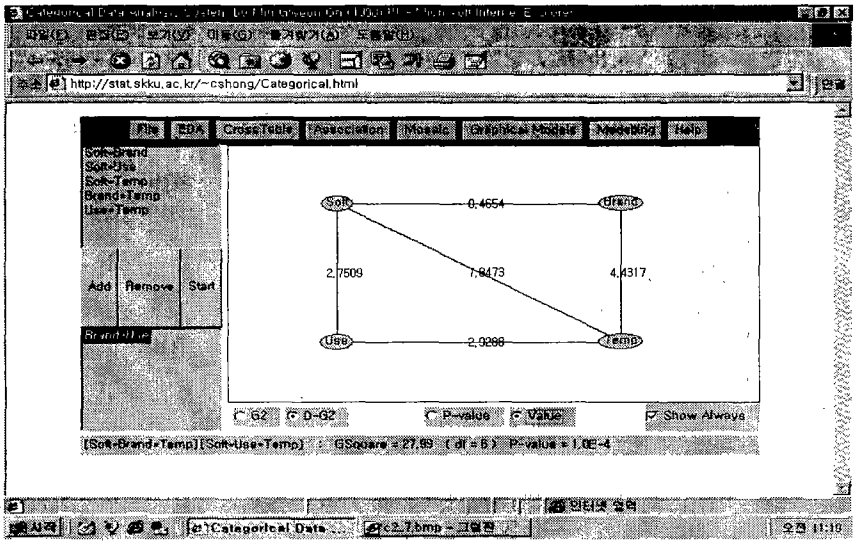

그림 2.8: [Use*Brand]이 제거된 연관그림

과 대응하는 G^2 통계량값, 자유도, p-값이 나타난다. 또한 사용자가 하단부에 있는 선택사항중 G^2과 Value(또는 P-Value)를 선택한 후 연관그림에서 두 변수를 연결하고 있는 임의의 실선(Edge)위에 마우스를 위치시키면 두 변수의 일차 교호작용에 대한 G^2 통계량값(또는 P-value)이 실선위에 나타나고 상단부에 두 변수의 교호작용항이 제거된 모형식이 나타난다. 선택사항 중 $D - G^2$은 선택된 두 변수의 교호작용항이 포함되는 연관모형의 G^2 통계량값과 제거된 경우의 G^2 통계량값의 차이이다. 사용자가 이러한 선택사항을 적절히 사용하여 이 윈도우의 왼쪽에 나열된 일차 교호작용항중에서 유의하지 않는 두 변수를 Remove 버튼을 사용하여 제거하면 오른쪽에 있는 연관그림에서 대응하는 변이 제거되고 하단부의 모형식과 대응하는 G^2 통계량값, 자유도, p-값이 실시간에 변화된다. 이러한 후진선택 과정을 반복하면서 사용자는 edge가 포함된 모형과 제거한 모형의 차이에 대하여 비교 분석할 수 있으며 사용자 스스로 최적의 모형을 찾을수 있다. 또한, 사용자는 Remove 버튼을 사용하여 왼쪽에 있는 모든 일차 교호작용항들을 선택하여 제거한 후에 단계식 선택방법과 유사하게 Add 버튼을 사용하여 유의한 일차 교호작용항들을 추가하면서 최적의 모형을 찾을 수 있다. 그림 2.8은 그림 2.7에서 설정된 부분 연관모형에 대한 연관그림에서 유의(논문에서는 점선, 그리고 시스템에서는 빨간 실선으로 표시)하게 나타난 [Use*Brand]항을 제거한 경우의 모형([Soft*Brand*Temp][Soft*Use*Temp])에 대한 연관그림이다.

2.7. Modelling 윈도우

2.6절에서는 연관그림으로 나타낼 수 있는 그래픽칼 모형들에 대하여 edge를 제거하거나 추가하면서 최적의 모형을 설정할 수 있었다. 그러나 여기에서는 계층모형(Hierarchical model)에서 최적의 모형을 찾을 수 있는 방법을 제안하고자 한다. 그림 2.9에 나타나는

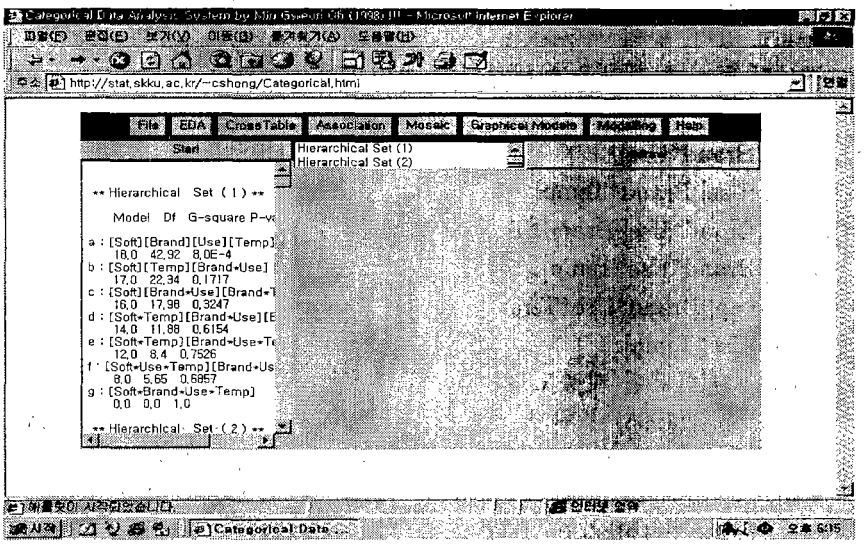

그림 2.9: 검정 결과

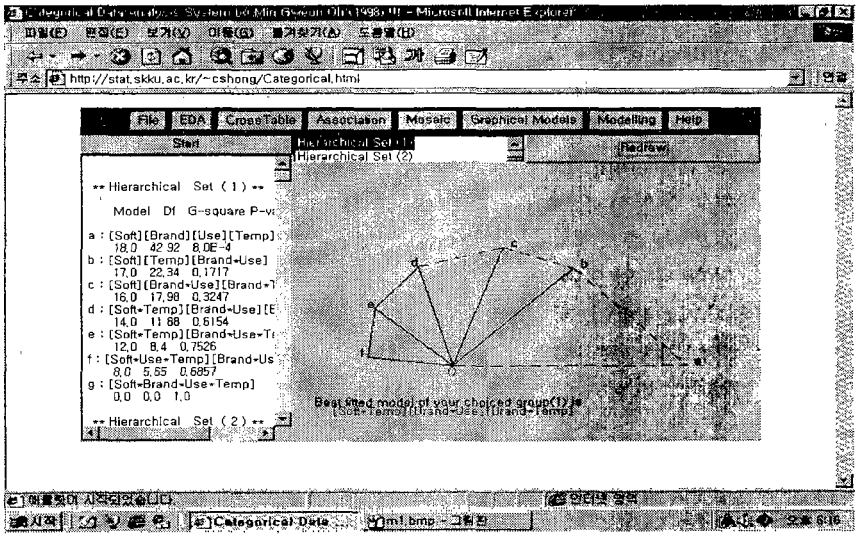

그림 2.10: 펼쳐진 다면체 그림

Modelling 윈도우의 왼쪽에는 본 논문에서 제시한 4차원 자료에 대해 계층구조하에서 고려할수 있는 24개의 계층구조 집단의 모든 대수선형모형과 대응하는 G^2 통계량값, 자유도, p-값를 나타내고 있다. 만일 사용자가 특정한 계층구조하에서 다음과 같은 일곱개의 계층모형으로 구성된 첫번째 집단을 선택했다고 가정하자.

[Soft][Brand][Use][Temp],
[Soft][Temp][Brand*Use],
[Soft][Brand*Use][Brand*Temp],
[Soft*Temp][Brand*Use][Brand*Temp],
[Soft*Temp][Brand*Use*Temp],
[Soft*Use*Temp][Brand*Use*Temp],
[Soft*Brand*Use*Temp].

이러한 일곱개의 선택된 계층모형에 대하여, 사용자는 그림 2.10와 같이 윈도우의 오른쪽에 홍종선과 최현집(1995)이 제안한 "펼쳐진 다면체 그림(Stretched polyhedron plot)"을 사용하여 모형을 비교할 수 있다. 모형의 검정통계량이 유의한 값을 가지면, 모형의 G^2 통계량값을 의미하는 선이 점선(시스템에서는 빨간색) 으로 나타나고 검정통계량이 유의하지 않는 값을 가지면 모형이 자료에 적합하다는 것을 의미하며 실선(시스템에서는 파란실선)으로 나타난다. 이러한 계층모형의 분석결과로부터 다음 다섯개의 모형은 유의하지 않으므로 실선(시스템에서는 파란실선)으로 나타나고 있다.

[Soft][Temp][Brand*Use],
[Soft][Brand*Use][Brand*Temp],
[Soft*Temp][Brand*Use][Brand*Temp],
[Soft*Temp][Brand*Use*Temp],
[Soft*Use*Temp][Brand*Use*Temp].

두 모형([Soft] [Brand*Use] [Brand*Temp] 와 [Soft*Temp] [Brand*Use] [Brand*Temp])의 차이가 유의하게 점선(시스템에서는 빨간 실선)으로 나타나고 있지만, [Soft*Temp] [Brand*Use] [Brand*Temp] 와 [Soft*Temp] [Brand*Use*Temp]의 G^2 통계량값의 차이는 유의하지 않게 실선(시스템에서는 파란실선)으로 나타나고 있다. 그러므로 사용자는 선택한 첫 번째 계층구조 집단의 일곱개의 계층모형들 중에서 [Soft*Temp] [Brand*Use] [Brand*Temp]을 최적의 모형이라고 판단할 수 있으며, 펼쳐진 다면체 그림 밑에 최적 모형식이 사용자에게 제공된다. 만약 사용자가 24개의 계층구조 집단에서 다른 집단의 계층모형들을 선택한다면, 선택된 모형들에 대한 새로운 펼쳐진 다면체 그림을 얻을 수 있으며 그 구조하에서 최적의 모형을 구할 수 있다.

2.8. HELP 윈도우

Help 윈도우는 사용자가 시스템을 쉽고 편리하게 사용할 수 있도록 도와주는 도움말 기능을 제공하고 있다. 사용자는 이 윈도우를 이용하여 시스템의 사용상의 문제뿐만 아니라 통계적 사전 지식에 관한 정보를 학습할 수 있다. 또한 범주형 자료를 분석하는데 필요한 11개의 연관성 통계량, 분포, 모형의 적합도 검정 통계량, 연관그림이나 모자익 그림에 대

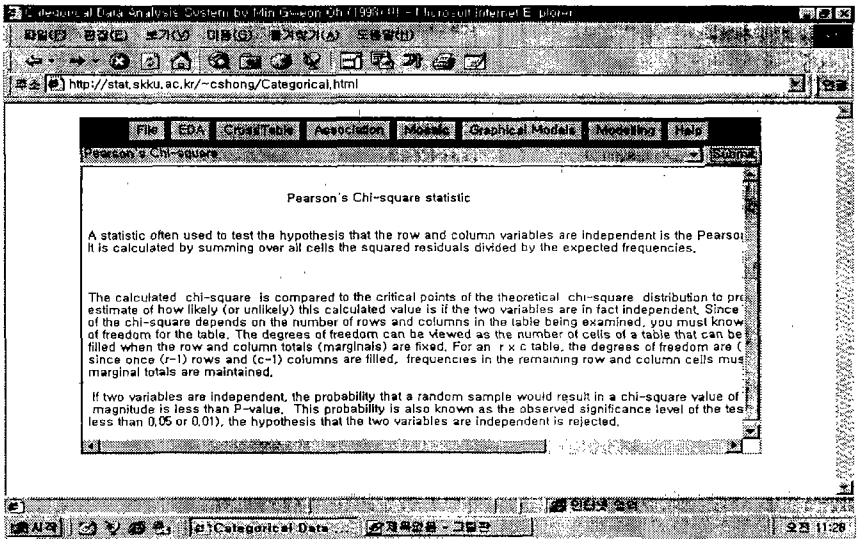

그림 2.11: HELP 윈도우

한 자세한 설명, 참고문헌 등을 필요할 때마다 즉시 찾아볼 수 있다. 그림 2.11은 Help 윈도우에서 사용자가 피어슨의 X^2 통계량을 선택한 경우의 출력 결과로써 사용자는 피어슨의 X^2 통계량에 관한 자세한 내용 뿐만아니라 피어슨의 X^2 통계량을 소개한 참고문헌에 관한 정보도 얻을 수 있다.

3. 결론

본 논문에서는 월드 와이드 웹에서 누구나 쉽게 사용할 수 있는 범주형 자료에 관한 학습 시스템과 분석 시스템을 구현하였다. 구현된 시스템은 월드 와이드 웹에서 구동되는 시스템이므로 구현 언어인 JAVA의 특성상 자료의 호환이 용이하지 않고 속도면에서 다소 느리다는 단점이 있지만, 방대한 크기의 다른 통계 패키지를 구입할 필요가 없으며 다른 통계 패키지를 구입하였다 하더라도 이 시스템은 프로그램의 크기가 아주 작으며, 웹 브로우져가 있는 어떤 기종의 컴퓨터에서도 쉽게 사용할 수 있다는 장점이 있다. 특히, 기존의 다른 통계 패키지에서 제공하고 있지 않는 모자익 그림이나 연관그림 그리고 펼쳐진 다면체 그림을 동적 그래픽스로 구현하여 사용자가 스스로 모형을 설정하고 최적의 모형을 찾을 수 있도록 하였다. 또한 통계학이나 범주형 자료분석에 대한 사전지식이 없는 일반 사용자라도 구현된 시스템의 Help 윈도우를 통해 범주형 자료분석에 대하여 지식을 습득할 수 있으며, 자료를 실시간에 분석할 수 있을 것이다.

참고문헌

[1] 홍종선 (1995). 〈 대수선형모형〉, 자유아카데미.

[2] 홍종선, 최현집 (1995). Graphical Descriptions for Hierarchical Log Linear Models, 〈 한국통계학회 논문집〉, 제2권 2호, 310-319.

[3] Agresti, A. (1984). *Analysis of Ordinal Categorical Data*, John Wiley & Sons, New York.

[4] Agresti, A. (1990). *Categorical Data Analysis*, John Wiley & Sons, New York.

[5] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*, John Wiley & Sons, New York.

[6] Christensen, R. (1990). *Log-Linear Models*, Springer-Verlag.

[7] Edwards, D. (1995). *Introduction to Graphical Modeling*, Springer-Verlag.

[8] Fienberg, S. E. (1980). *The Analysis Cross-Classified Categorical Data*, MIT Press.

[9] Friendly, M. (1992). Mosaic displays for log-linear models, *Proceedings of the Statistical Graphics Section, American Statistical Association*, 61-68.

[10] Friendly, M. (1994). Mosaic displays for multi-way contingency tables, *Journal of the American Statistical Association*, Vol. 89, 190-200.

[11] Goodman, L. A. and Kruskal, W. H. (1979). *Measures of association for cross classifications*, Springer-Verlag.

[12] Haberman, S. J. (1974). *The analysis of frequency data*, Chicago and London Press.

[13] Hartigan, J. A. and Kleiner, B. (1981). *Mosaic for contingency tables*, Springer-Verlag.

[14] Plackett, R. L. (1981). *The Analysis of Categorical Data*, Charles Griffin and Company Ltd.

[1998년 4월 접수, 1998년 9월 최종수정]

Categorical Data Analysis System in the internet *

Chong Sun Hong[1] Donguk Kim[2] Min Gweon Oh[3]

ABSTRACT

A categorical data analysis system in the World Wide Web is proposed with an easy-to-use environment. This system is composed of four components. First, this system presents several graphical displays for Exploratory Data Analysis for categorical data. Second, it provides some measures of association including dynamic graphics for mosaic plots of Hartigan and Kleiner (1981) and Friendly (1994). Dynamic graphics for mosaic plots give some useful informations. Third, this system can analyze categorical data with loglinear models. So we can select the best fitted loglinear model interactively.

* This paper was supported by Non Directed Research Fund, Korea Research Foundation and 63 Research Fund, SungKyunKwan University, 1996(http://stat.skku.ac.kr/~cshong/Categorical.html).

1) Professor, Department of Statistics, SungKyunKwan University, MyungRyun-dong 3 Ga, JongRho-Gu, Seoul, KOREA

2) Professor, Department of Statistics, SungKyunKwan University, MyungRyun-dong 3 Ga, JongRho-Gu, Seoul, KOREA

3) Department of Statistics, SungKyunKwan University, MyungRyun-dong 3 Ga, JongRho-Gu, Seoul, KOREA