응용통계연구 제 **20**권 *1*호, *2007*년, *pp. 61–78*

정보이론과 시각화 방법에 의한 여론조사 분석의 새로운 접근방법*

허문열[1] 차운옥[2]

요 약

본 논문에서는 상호정보와 데이터 시각화를 사용하여 여론조사 결과를 분석하는 방법을 제안하였다. 여론조사의 경우, 목적변수와 이를 위한 설명변수가 있으며 설명변수는 수치형과 명목형이 혼재된 형태이다. 상호정보를 사용하면 목적변수에 대한 혼합형 설명변수의 영향을 크기순으로 순위를 매길 수 있고, 데이터 시각화 방법을 사용하여 이들 순위 매김에 대한 평가를 수행할 수 있다. 여론조사에서 목적변수에 미치는 설명변수의 영향력의 크기가 어느 정도인가를 정량화한 것은 이번 연구에 의해서만 이루어진 것이다.

주요용어: 연관성 척도, 상호정보, 데이터 시각화.

1. 서론

여론조사는 의사결정에 유용하게 사용할 수 있는 정보를 제공하기 위하여 사전에 계획된 대상을 통해 체계적, 과학적으로 자료를 획득, 분석, 해석하는 객관적이고 공식적인 과정이다. 여론조사를 행하는 분야로는 기업의 마케팅조사 및 컨설팅 분야, 사회여론조사, 정치선거조사 및 컨설팅 분야 등이 있다. 기업의 마케팅조사 및 컨설팅 분야에서는 고객만족도조사, 기존상품 이미지 및 소비행태조사, 신상품 시장가능성조사, 광고효과조사, 상권분석 및 유통조사, 기업이미지조사 등을 수행한다. 사회여론조사 및 정치선거조사 분야에서는 정당의 정책이나 정치여론조사, 선거예측조사 및 선거전략조사, 정부, 지자체, 공공단체 등 기관의 국민여론조사 및 정책만족도 조사, 언론사 여론조사 등을 수행한다. 일반 여론조사의 설문내용은 설명변수(또는 독립변수)와 목적변수(또는 종속변수)로 구성되어 있다. 대부분의 경우, 설명변수로는 수치적인 질문(예를 들면 나이, 교육연수, 수입 등)과 명목형 질문(성별, 지역, 종교 등)이 혼합적으로 이루어져 있으며, 명목형 질문도 카테고리가 두 개에서부터 십여 개 (지역 등), 또는 수 십 개에 이르는 경우가 있다. 이와 같이 각 질문(변수)들의 형식이 통일되어 있지 않고 복합적인 경우, 어떤 변수가 가장 중요한 변수인

* 본 연구는 2006년도 한성대학교 교내연구비 지원과제임.

1) (110-745) 서울시 종로구 명륜동 3가 53번지, 성균관대학교 통계학과, 교수
E-mail: myhuh123@skku.edu

2) (교신저자)(136-792) 서울시 성북구 삼선동 3가 389, 한성대학교 멀티미디어공학과, 교수
E-mail: wcha@hansung.ac.kr

지를 판단하는 기준이 매우 애매하며 이를 위한 통합된 이론이 없다. 여론조사 결과에 대한 사후 자료분석의 경우 대부분 빈도분석, 교차분석, 상관관계분석 등 간단한 통계적 방법을 사용한다. 본 논문에서는 여론조사 결과를 분석하기 위해 정보이론(information theory)을 이용하여 설명변수와 목적변수간의 연관성을 측정하고, 이를 통해 목적변수에 영향을 미치는 중요한 변수부터 순위를 부여하는 방법을 제시하고자 한다. 또한 데이터 시각화(data visualization) 기법을 통해 이들 순위에 대한 평가를 수행하는 방법을 제시한다. 본 연구에서는 이와 같은 방법을 선거여론조사를 분석하는데 적용하였다.

2. 선거여론조사 자료

본 논문에서 사용하는 자료는 문화일보가 2002년도 3월부터 10월까지 조사한 대선관련 선거여론조사 자료의 일부로서 7월과 8월에 조사한 각각 1,000명에 대한 것이며, 여러 가지 질문 중에서 중요하다고 생각되는 것만 몇 개 택하였다 (조사 시기, 성별, 연령, 학력, 소득수준, 거주지, 고향). 질문에 따라 결측값이 포함되어 있으며 이들 결측값도 분석하는데 같이 고려하였다. 또한, 소득은 다음과 같은 코드로 변환하였다.
1: 70만원 이하, 2: 71~ 100만, 3: 101 ~ 150만, 4: 151 ~ 200만, 5: 201 ~ 250만, 6: 251 ~ 300만, 7: 301 ~ 350만, 8: 401 ~ 500만, 9: 501만원 이상

이 값들을 수치형으로 생각할 수도 있고 범주형으로 생각할 수도 있다. 본 논문에서는 두 가지 경우를 모두 고려하기로 한다. 자료에 대한 개괄적인 파악을 위해 기술통계를 살펴보면 다음과 같다.

- 조사시기: 7월 1,000명, 8월 1,000명.
- 성별: 남 988명, 여 1,012명.
- 지지후보: A후보 927명, B후보 603명, 기타 470명.
- 연령분포:

20대초	20대후반	30대초	30대후반	40대초	40대후반	50대이후
236명	239명	247명	260명	223명	208명	587명

- 학력분포:

국졸	중졸	고졸	대졸이상	결측값
252명	192명	729명	815명	12명

- 소득분포:

1	2	3	4	5	6	7	8	9	결측값
247명	135명	246명	339명	215명	232명	190명	83명	97명	216명

- 거주지분포:

서울	경기	충청	전라	경북	경남	강원제주	결측값
439명	483명	207명	229명	230명	330명	82명	0명

- 고향분포:

서울경기	충청	전라	경상	강원기타	결측값
297명	353명	444명	655명	225명	26명

3. 설명변수가 목적변수에 미치는 영향의 측정

여론조사 분석을 위한 일반적인 통계적 방법에서는 간단한 기술통계를 사용한다. 가장 많이 사용하는 것은 간단한 테이블형 도표와 막대그래프, 그리고 선그래프이다. 목적변수에 많은 영향을 미치는 설명변수를 찾는 것은 기술통계만 가지고는 가능하지 않고 두 변수간의 연관성 척도를 통해 찾아낼 수 있다. 두 변수간의 연관성을 측정하는 방법에는 다음과 같은 것이 있다.

1. 두 변수 모두 수치형이고 두 변수 모두 정규분포를 따른다면 : Pearson의 상관계수를 사용
2. 두 변수 모두 수치형이고 정규성에 의심이 가는 경우 : Spearman's rho를 사용
3. 두 변수 모두 범주형이고 범주의 수가 2개인 경우 : Kramer의 Phi 등을 사용

두 변수가 모두 2진 범주형인 경우에는 수 많은 척도가 있으며 각 척도에 따라 다른 정보를 제공하게 된다(Tan, Kumar와 Srivastava, 2002). 두 변수의 범주 수가 2개가 아니면 일반적으로 통용되는 연관성 척도가 알려져 있지 않다. 또, 본 논문에서 다루고자 하는 경우는 두 변수가 모두 범주형이거나, 설명변수는 수치형이고 목적변수는 범주형인 경우이다. 이러한 문제를 해결하기 위한 몇 가지 연구가 있었으며 대표적인 것 두 가지를 다음에 기술한다.

3.1. MDI에 의한 연관성 척도

편리상 설명변수를 X, 목적변수를 Y라고 하자. MDI(measure of departure from independence)는 X와 Y가 독립적인가에 대한 검정결과의 유의확률(p-값)을 기준으로 연관성을 측정해주는 척도이다(Lee와 Huh, 2003). 즉, 두 변수 X와 Y가 독립에서 멀어질수록 p-값은 작아지며, 따라서 p-값이 작을수록 두 변수의 연관성이 강하다고 할 수 있다. 검정방법은 X가 수치형인 경우 Kruskal-Wallis 검정을 사용하고, X가 범주형이면 카이제곱검정을 사용한다.

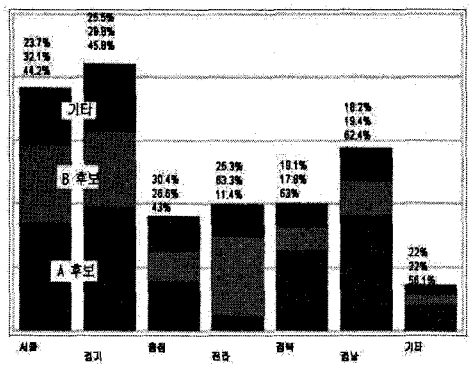


그림 3.1: 각 지역별 지지후보의 범주별 막대그래프(각 지역별로 지지후보의 3가지 범주를 상대비율로 나타내었다. 아래 부분이 A후보 지지, 중간은 B후보 지지, 그리고 윗 부분은 기타 후보 지지이다.)

본 논문에서 사용하는 자료에 대해 거주지와 지지후보의 분할표와 범주별 막대그래프를 그리면 표 3.1과 그림 3.1과 같다.

그림 3.1에 의하면 전라지역과 경북, 경남은 후보지지 성향에 매우 다른 양상을 보이는 것을 알 수 있다. 즉, 전라지역의 경우 B후보 지지가 63.3%인 반면, 경상지역은 B후보지지가 20% 미만을 보이고 반대로 A후보 지지가 63% 수준을 보이고 있다. 즉, 지역에 따라 후보자의 지지도 성향이 많이 다르며 이는 지역과 후보자 간의 연관성이 강하다는 것을 의미한다. 연관성의 정도를 파악하기 위해 거주지와 지지후보와의 분할표에 대한 카이제곱 통계량을 계산하면 $\chi^2 = 218$이며, 여기에 해당되는 p-값은 $7.3e-40$, 즉 거의 0이다. 따라서 거주 지역은 지지후보와 매우 긴밀한 연관성을 가지고 있다. 연령과 지지후보와의 연관성에 대해서는 p-값이 $4.1e-25$로 역시 거의 0이다(그림 3.2). 따라서 연령도 지지후보에 매우 강한 영향을 미치는 것을 알 수 있다. 그림 3.3을 보면 30대 초반 까지는 세 후보 간 지지비율이 크게 차이가 없으나 30대 초반 까지는 B후보 지지가 많고, 30대 후반부터 A후보 지지비율이 크게 늘어나는 것을 알 수 있다. 따라서 나이와 지지후보간의 연관성이 큰 것을 알 수 있다. 참고로 모든 변수들에 대한 p-값을 그림 3.2에 나타내었는데(p-값의 순서에 따라) 이 그림에서는 p-값 대신 $-\log$(p-값)을 사용하였다. 그 이유는 p-값이 작을수록 연관성이 많은데 이는 일반적인 관념하고는 상반되며, p-값이 매우 작은 경우 p-값 자체로 연관성

표 3.1: 지역별 지지후보 분포

	서울	경기	충청	전라	경북	경남	강원/제주
A후보	194	221	89	26	145	206	46
B후보	141	139	55	145	41	64	18
기타	104	123	63	58	44	60	18

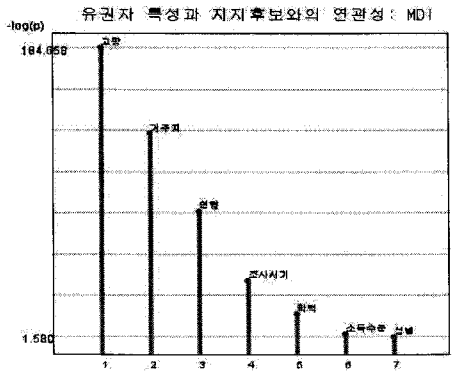


그림 3.2: 각 성향별 $-\log(MDI)$

들의 척도를 비교하는 것이 어렵기 때문에 $-\log$ 변환을 시킨 것이다. 이렇게 하면 연관성이 매우 커서 p-값이 매우 작을 때 (0에 가까울 때) $-\log$(p-값)은 매우 커지며, 연관성이 거의 없어 p-값이 1에 접근하면 $-\log$(p-값)이 0에 접근하여 쉽게 해석할 수 있다.

그림 3.2에 의하면 소득과 성별은 지지후보에 별로 영향을 미치지 않는 성향임을 알 수 있다. 소득을 수치형 변수로 고려하여 카이제곱 검정 대신 Kruskal-Wallis 검정법을 사용하면 결과는 0.8로 역시 지지후보에 영향을 미치지 않음을 알 수 있다.

표 3.2: 연령별 지지후보 분포

	20대초	20대후반	30대초	30대후반	40대초	40대후반	50이후
A후보	37.7	31.4	29.6	41.5	49.3	56.2	60.5
B후보	42.8	44.8	41.3	33.8	23.3	20.2	18.9
기타	19.5	23.8	29.1	24.6	27.4	23.6	20.6

3.2. 상호정보에 의한 연관성 척도

앞에서 제시한 MDI 척도가 나름대로 의미를 가지고 있지만 이미 설명한 바와 같이 이 방법에는 한계점이 있다. 첫째, MDI는 수치형 변수와 범주형 변수에 대해 다른 검정통계량을 사용하기 때문에 일관적인 결과를 기대하기 어렵다. 앞에서 계산한 바와 같이 소득을 수치형으로 볼 때는 p-값이 0.8로 나타났지만, 이를 범주형 변수로 고려하면 p-값이 0.1로 나타났다. 둘째, 해당 검정통계량의 유의수준에 따라 p-값을 계산하여 MDI 척도를 측정하고 있다. 사용하는 검정통계량의 분포가 근사적인 카이제곱분포를 따른다는 성질을 사용하기 때문에 계산되는 유의수준은 근사적인 값이 된다.

상호정보(mutual information, I(X;Y))는 1946년 Shannon에 의해 제안된 정보이론에 근거하고 있으며 이는 매우 일반적인 성질을 가지는 척도이다 (Cover와 Thomas, 2004). 즉,

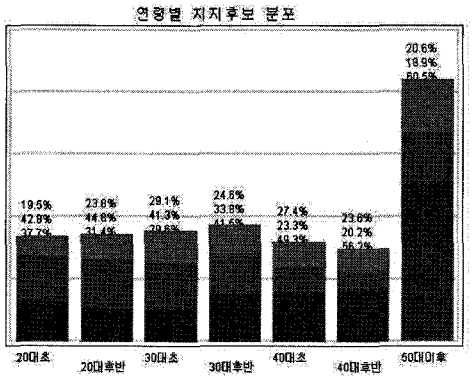


그림 3.3: 연령별 지지후보 상대비율(젊은층은 B후보지지가 많고 30대 후반부터 A후보 지지가 많은 것을 알 수 있다.)

X와 Y가 어떠한 형태라도 가능하며 두 변수에 대한 어떠한 가정도 하지 않는다. 상호정보는 다음과 같이 정의된다.

$$\begin{aligned} I(X;Y) &= \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_i p_j} \\ &= \sum_{i,j} p_{ij} \log p_{ij} - \sum_i p_i \log p_i - \sum_j p_j \log p_j. \end{aligned} \tag{3.1}$$

여기서 log는 밑을 2로 하고 X는 r개의 범주를 갖고 Y는 c개의 범주를 가지며, p_{ij}, $i = 1,\ldots,r$; $j = 1,\ldots,c$는 분할표(contingency table)의 (i,j) 칸의 확률이고, $p_i = \sum_{i=1}^{r} p_{ij}$, $p_j = \sum_{j=1}^{c} p_{ij}$ 이다. $I(X;Y)$에 관한 몇 가지 특성을 기술하면 다음과 같다.

1. $I(X;Y)$는 항상 0 보다 같거나 크다.
2. 만약 두 변수가 독립적이라면 $p_{ij} = p_i p_j$가 성립함으로 I(X;Y)=0 이다.
3. 두 변수간의 연관이 클수록 $I(X;Y)$는 커진다.
4. X나 Y는 여러 개의 변수집단이 될 수도 있다. 예를 들어 X를 학력, 지역, 소득과 같은 세 개의 변수집합으로 생각할 수 있고 Y는 지지후보자가 될 수 있다. 이 경우, $I(X;Y)$는 (학력, 지역, 소득)과 지지후보자 간의 연관성을 측정해 준다.

최근 Hutter와 Zaffalon(2005)이 베이지안적 접근방법에 의한 $I(X;Y)$의 추정값으로 다음을 제시하였다.

$$I(X:Y) \doteq J + \frac{(r-1)(c-1)}{2(n+1)}, \quad J = \frac{1}{n} \sum_{ij} n_{ij} \log \frac{n_{ij} n}{n_i n_j}. \tag{3.2}$$

여기서 $\doteq$는 좌변을 추정하기 위해 우변을 사용한다는 뜻이고, n_{ij}는 (i,j) 칸에 속하는 도수, $n_i = \sum_j n_{ij}$, $n_j = \sum_i n_{ij}$ 이며, (n_{ij}/n), (n_i/n), (n_j/n) 는 각각 p_{ij}, p_i, p_j의 최대가능도(maximum likelihood) 이다. 또한 분산은 다음과 같이 추정한다.

$$\mathrm{Var}[I(X:Y)] \doteq \frac{1}{n+1}(K - J^2), \quad K = \sum_{ij} \frac{n_{ij}}{n} (\ln \frac{n_{ij} n}{n_{i+} n_{+j}})^2. \tag{3.3}$$

$I(X;Y)$의 추정값은 중심극한정리에 의해 평균이 식 (3.2)와 같고 분산이 식 (3.3)과 같은 정규분포를 따른다. 따라서, $\hat{I}(X;Y) \pm 2\sqrt{\widehat{Var}(X;Y)}$이 0을 포함하면 $I(X;Y)$를 0으로 판단하고, X와 Y는 연관이 없는 것으로 결정한다.

대개 선거여론조사 같은 경우, n은 1,000명 정도로 매우 큰 편이므로 식 (3.2)의 두 번째 항은 무시할 만하다. 따라서 여기서는 $I(X;Y)$의 추정값으로 다음을 사용하기로 한다.

$$\hat{I}(X;Y) = J = \frac{1}{n} \sum_{ij} n_{ij} \log \frac{n_{ij} n}{n_i n_j}. \tag{3.4}$$

본 연구에서 사용하는 자료에서, X가 성별이고 Y가 지지후보인 경우 분할표는 다음과 같다.

	A후보	B후보	기타	합
남	452	290	246	988
여	475	313	224	1,012
합	927	603	470	2,000

이 표로부터 구한 p_{ij}, p_i, p_j의 추정값은 다음과 같다.

	A후보	B후보	기타	합
남	0.226	0.145	0.123	0.494
여	0.237	0.157	0.112	0.506
합	0.463	0.302	0.235	1.000

이 표에 의하면, (n_i/n) , $i = 1, 2$은 0.494, 0.506 이고, $(n_{11}/n), (n_{12}/n), \ldots, (n_{23}/n)$은 0.226, 0.145,..., 0.112이다. 이를 적용하면 $\hat{I}(X;Y) = 0.001$이다 (소숫점 이하 3자리까지만 계산).

마찬가지로 다른 성향들에 대해 계산하면 그림 3.4와 같은 결과가 나타난다. 그림 3.2의 MDI와 그림 3.4의 MI 를 비교해 보면 우연히도 두 방법에 의해서 구한 연관성 척도가 유사한 결과를 보임을 알 수 있다. 이미 MDI 에서도 나타난 바와 같이 MI의 경우도 연관성 척도를 구할 때 상대적인 크기가 중요하다. 즉, I(고향, 지지후보)는 0.075이고, I(거주지, 지지후보)는 0.056, I(소득, 지지후보)는 0.007이다. 이 값들을 보면 고향은 거주지에 비해

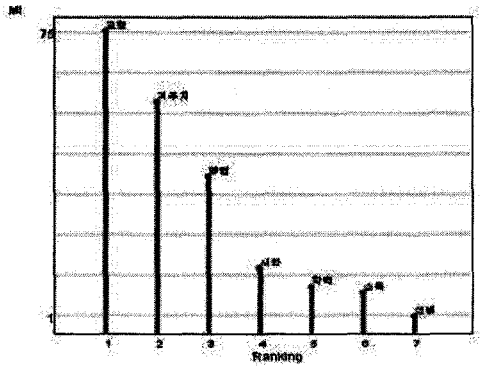


그림 3.4: 각 성향별 MI 값 1,000을 곱한 결과

50% 정도 더 큰 영향을 미치고, 소득에 비해 10배 정도 더 큰 영향을 미친다. 여기서는 연령과 소득을 모두 범주형 변수로 고려하였다.

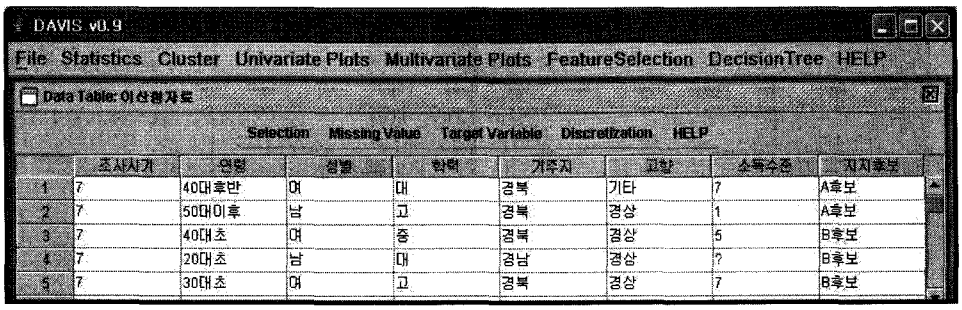
X가 수치형 변수일 때, $I(X;Y)$는 식 (3.1)의 합 공식을 적분으로 바꾸어야 한다.

$$I(X;Y) = \sum_{j=1}^{c} \int_x f(x,j) \log \frac{f(x,j)}{f(x)f(j)} dx. \tag{3.5}$$

여기서 $f(x,j)$는 수치형 변수 (유권자의 성향) X와 $Y = j$번 째 범주(j 번째 후보)와의 결합밀도확률이고, $f(x)$는 수치형 변수 X의 확률밀도함수, $f(j) = p_j$이다. 이 경우의 $I(X;Y)$를 추정하기 위해서는 $f(x,j)$와 $f(x)$의 추정이 필요하며, 이는 Parzen filter(Parzen, 1962) 방법이나 혼합정규분포 등에 의해 해결할 수 있다. 이들 방법에 대한 구체적인 추정방법은 이 논문의 핵심이 아니기 때문에 생략하고, 다만 다음 절에서 제시하는 소프트웨어에 대한 설명에서 이들에 의한 추정방법을 제시한다. 참고로, 연령과 소득을 수치형변수로 고려하여 $I(X;Y)$를 추정한 결과는 모두 0으로 나타났다.

4. 데이터 시각화 방법에 의한 여론조사 분석

여론조사의 결과를 도표로 표현하면 조사 결과를 빨리 이해할 수 있고 수치로는 알 수 없는 정보를 파악할 수 있는 장점이 있다. 특히 최근 많이 연구하고 있는 데이터 시각화(data visualization) 방법을 이용하면 자료 속에 내재된 정보를 파악하는 것이 더욱 용이해진다. 데이터 시각화의 기본적인 원리는 데이터 컨디셔닝(data conditioning)이다. 데이터 컨디셔닝은 데이터의 일부를 삭제(delete)하거나, 선택(select)하거나 집중(focus)하여 이 결과가 통계적 모형이나 도형에 미치는 영향을 분석하는 과정을 의미한다. 데이터 컨디셔닝은 통계적 도형위에서 컴퓨터의 마우스를 사용하여 데이터의 일부를 하이라이트(highlight) 또는 브러싱(brushing)하고 이를 다른 도형과 링크(link)함으로서 이루어진다. 범주형 자료의 경우, 데이터의 부분집합은 막대그래프에서 일부 범주를 선택하거나 일부 범주를 제거한 것이 될 수도 있고, 수치형 자료의 경우 히스토그램 등에서 일부를 선택하거나 일부를



	조사시기	연령	성별	학력	거주지	고향	소득수준	지지후보
1	7	40대후반	여	대	경북	기타	7	A후보
2	7	50대이후	남	고	경북	경상	1	A후보
3	7	40대초	여	중	경북	경상	5	B후보
4	7	20대초	남	대	경남	경상	?	B후보
5	7	30대초	여	고	경북	경상	7	B후보

그림 4.1: DAVIS에 선거여론조사자료를 입력시킨 후 화면

제거한 것이 될 수도 있다. 또는 군집분석을 수행하여 어떤 특정한 군집을 선택하거나 이를 제거한 후 나머지가 될 수도 있다. 하이라이트와 브러싱은 마우스를 데이터의 해당 부분에 갖다 대고 클릭을 하거나 데이터의 일부분을 빗질을 함으로서 이루어진다. 본 논문에서는 이러한 모든 과정을 DAVIS(Data VISualization system)(Huh와 Song, 2002)를 사용하여 수행한다(이 시스템은 참고문헌에 기술한 웹사이트에서 무료로 다운받아 사용할 수 있다).

DAVIS는 5개의 모듈로 이루어져 있다. 이들 5개의 모형은 데이터 관리, 군집분석, 통계적 도형, 변수선택, 결정나무이다. 각 모듈별로 간략히 설명하면 다음과 같다.

1. 데이터 관리 : 변수 선택과 표본추출, 결측값의 처리, 그리고 수치형 변수의 이산화 등을 수행해준다.

2. 군집분석 : 기술통계, k-평균법 (k-means), k-메도이드법(k-medoid), 분리법(divisive), EM법, 덴드로그램법(dnedrogram) 등

3. 통계적 도형 : 막대그래프, 히스토그램, QQ 도형, 상자도형, FEDF 도형, 평행좌표계, 산점도행렬, Grand Touring, 선 모자익 도형, 주성분분석도형 등

4. 변수선택 : 정보이론을 포함한 여러 가지 방법에 의한 변수중요도와 중요 부분집합 선택

5. 결정나무 : C4.5 알고리즘에 의한 결정나무

다음에는 이를 사용하여 여론조사 결과를 분석하는 간단한 시각화 방법을 소개하고, 정보이론에 의한 변수중요도 순위를 결정하는 방법과 시각화 방법을 사용하여 여론조사에 숨겨져 있는 정보를 탐색하는 방법을 설명한다.

4.1. 시각화 방법의 소개

간단한 데이터 컨디셔닝은 막대그래프를 사용하여 해결할 수 있다. 예를 들어 "지역별 A후보 지지자의 분포는?" 이라는 질문이 있다면, 이는 다음과 같은 과정에 의해 처리할 수 있다.

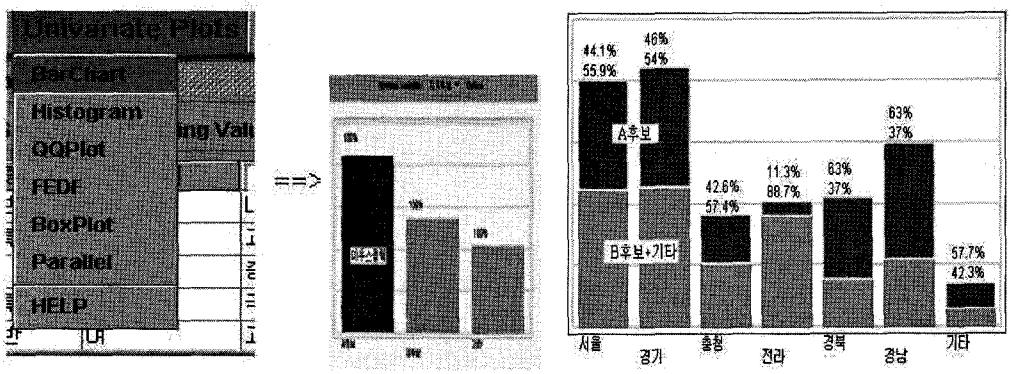


그림 4.2: DAVIS를 이용한 막대그래프 작성과정

1. 지지후보 막대그래프에서 A후보 부분을 하이라이트 한다.
2. 거주지역 막대그래프를 그린다. 막대그래프는 "Univariate Plots ⇒ BarChart"에 의해 그릴 수 있다.
3. 즉, DAVIS 메뉴에서 Univariate Plots를 택하고, 여기서 BarChart를 택한다. 막대그래프 중에, 지지후보 막대그래프를 택하고, A후보를 브러시하여 선택하면 이 부분이 붉은 색으로 하이라이트 된다. 이제 거주지 막대그래프를 선택하면 앞의 지지후보 막대그래프에서 선택된 결과가 바로 거주지 막대그래프에 전파되면서, 지역별로 A후보 지지의 분포가 나타난다.

이상의 과정이 그림 4.2에 나타나 있다.

이제 다음과 같은 좀 더 복잡한 질문이 주어졌다고 하자. "나이가 30대 이하인 유권자 중 A후보를 지지하는 사람들의 거주지역별 분포는?" 이를 막대그래프에 의해 처리하려면 다음과 같이 한다.

1. 연령 막대그래프에서 30대 이하인 유권자들을 선택한다.
2. 지지후보 막대그래프에서 A후보를 하이라이트 한다.
3. 거주지역 막대그래프를 그린다.

표 4.1: 나이 30대 이전과 40대 이후의 두 그룹에서 A후보를 지지하는 유권자들의 지역별 분포(단위 : %)

	서울	경기	충청	전라	경북	경남	기타
30대 이전	27.6	38.7	28.9	10.9	49.5	46.9	47.2
40대 이후	61.7	53.7	55.5	11.7	74.4	77.4	63.0

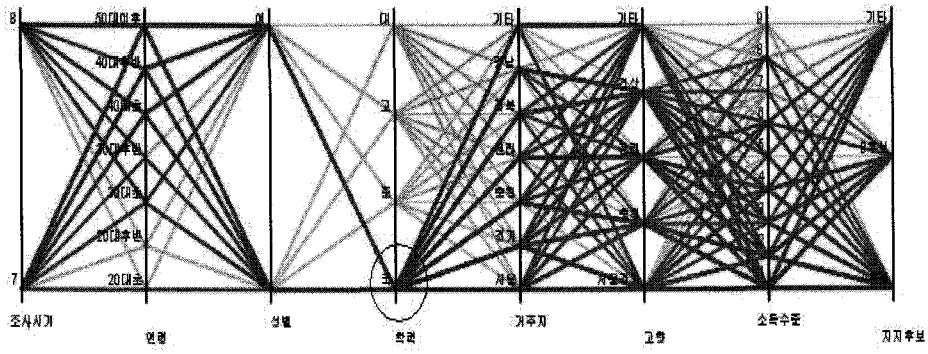


그림 4.3: 여론조사 자료의 PCP.

그러나 막대그래프를 사용하여 이러한 작업을 하는 것보다 평행좌표계(parallel coordinate plot, PCP)를 사용하면 매우 편리하고 더욱 복잡한 작업을 손쉽게 처리할 수 있다. 평행좌표계는 각 변수마다 하나의 수직선을 그리고, 각 관측값들을 선분으로 이어 표시한 것이다. 그림 4.3은 여론조사 자료에 대한 PCP이다. 여기서는 모든 변수들이 범주형이기 때문에 관측값을 나타내는 값들이 중복되어 나타나 있다. 학력 축에서 국졸이하를 브러싱한 결과 국졸이하 조사자들이 다른 축에 연결되어 나타났다. 이 그림을 보면 국졸이하는 여자쪽에 남자보다 강한 붉은 선으로 연결되고, 저소득이 많은 것으로 나타난다.

앞 질문을 PCP를 사용하여 해결하는 과정은 다음과 같다(그림 4.4)

1. PCP의 연령 축에서 30대 이하를 브러싱하고 마우스의 오른쪽 버튼을 눌러 Focus를 택한다.
2. PCP의 지지후보 축에서 A후보를 선택한다.
3. 거주지역 막대그래프를 그린다.

비슷한 과정에 의해 40대 이후를 분석한 결과는 표 4.1과 같이 나타났다. 대부분의 자료 탐색은 막대그래프와 PCP를 사용하면 해결할 수 있다. 특히 자료가 수치형인 경우, PCP를 잘 이용하면 자료의 탐색을 보다 효율적으로 수행할 수 있다.

4.2. 상호정보와 데이터 시각화에 의한 목적변수에 대한 설명변수의 영향분석

지지후보에 영향을 가장 많이 미치는 변수는 무엇인가? 또,조사시점(7월, 8월)에 따라 변화가 있는가? 연령층에 따라 유권자의 성향이 변하는가? 노년층 저학력의 경우 젊은 층 고학력 유권자에 비해 지역색이 더 약해지는가? 특정지역을 제외하면 지역색이 없는가? 이러한 질문들은 3절에서 설명한 연관성 척도에 의한 유권자 성향의 탐색과 앞에서 설명한 시각적 방법에 의한 데이터 컨디셔닝 과정을 통합하면 해결할 수 있다. 여기서는 이러한 과

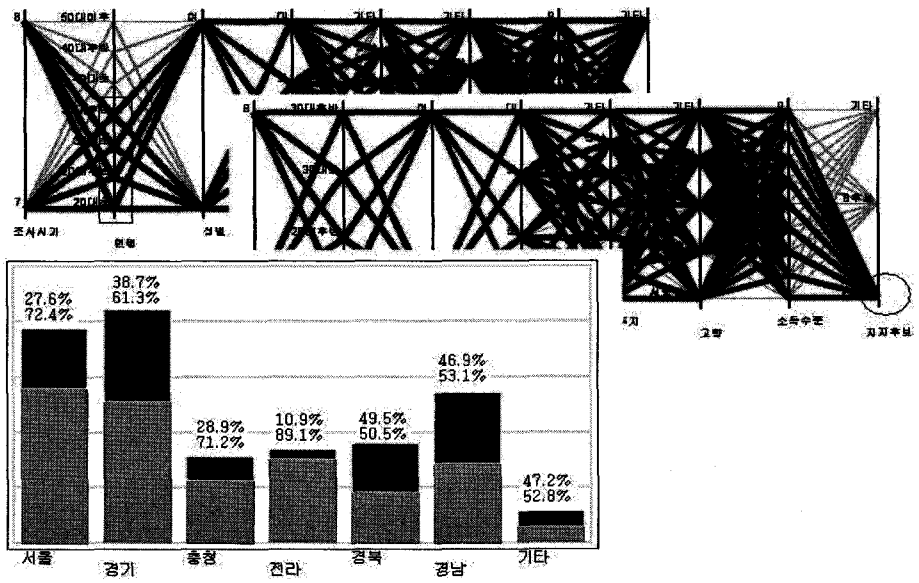


그림 4.4: PCP를 사용하여 해결하는 과정(윗 그림: PCP 의 30대 이하 선택(Focus). 가운데 그림: A후보 하이라이트. 아래 그림: 30대 이하인 사람 중에 A후보를 지지하는 사람들의 거주지 분포. 경북이 가장 많고, 다음이 경남, 기타, 경기, 충청 순이다.)

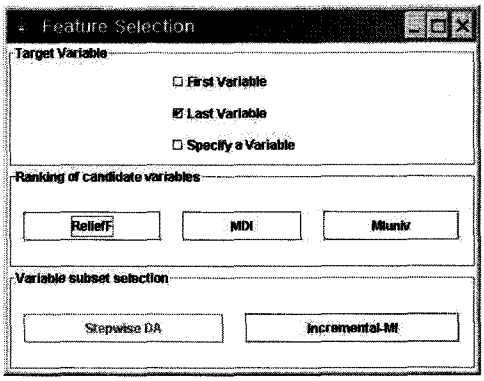


그림 4.5: DAVIS의 Feature Selection 대화상자 화면(DAVIS의 메뉴에서 Feature Selection ⇒ Feature Selection Panel을 선택하면 왼쪽과 같은 패널이 나타난다. 첫 번째 패널은 목적변수를 선택하는 패널이다. 본 여론조사의 경우, 목적변수(지지후보)가 가장 마지막 변수이다. 두 번째 패널에서 3가지 방법에 의한 변수의 중요도 평가방법을 선택할 수 있다. 두 번째와 세 번째가 이 논문에서 설명한 방법들이다.)

정에 대해 설명한다. 연관성 척도로서는 이미 설명한 바와 같이 상호정보가 매우 일반적인

개념이기 때문에 여기서는 상호정보만 사용하기로 한다.

DAVIS의 "Feature Selection" 모듈을 사용하면 상호정보에 의한 유권자 성향의 중요도를 알 수 있다(그림 4.5). 이 모듈은 변수의 중요도 외에 가장 중요한 변수들의 집합을 선택해 주는 부분이 있으나 여기서는 변수의 중요도에 따라 순위를 정해주는 부분만 사용하여 문제를 해결하고자 한다. 이를 위해서는 먼저 목적변수를 선정해야 하며, 본 논문의 경우 "지지후보"가 해당 목적변수이다. 또한 여러 가지 변수중요도 판정 기준 방법 중, 여기서는 상호정보에 의한 것만 기준으로 하여 설명한다.

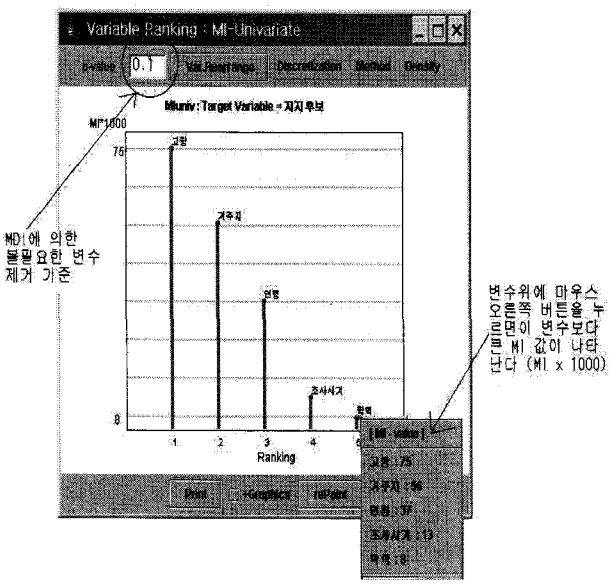


그림 4.6: MI의 여러 가지 옵션(3.2절의 이론에 의해 MI값이 0이 아닌 변수들만 나타낸 그래프이다. MI 값들을 알고 싶으면 해당 변수 위에 마우스의 오른쪽 버튼을 누른다. 이 행동에 의해 이 변수보다 더 중요한 변수들의 MI 값이 나타난다. 다른 메뉴들은 수치형 자료의 MI 계산에 필요한 것이거나, 본 논문에서는 별로 중요하지 않은 것이다.)

상호정보에 의한 변수중요도 평가는 그림 4.5의 두 번째 패널의 "MIuniv"에 의해 이루어진다. 그림 4.6에 지지후보와 유권자 성향 간의 상호정보가 나타나 있다. 변수 중에 MDI를 기준으로 하여 p-값이 0.1보다 큰 변수들은 지지후보에 영향을 미치지 않는 변수로 판단하여 제거하였다. 또 일반적으로 연관관계가 큰 변수라 하여도 MI는 매우 작은 값으로 나타나기 때문에 해석의 용이성을 위해 MI에 1,000을 곱하여 그림에 표시하였고 앞으로 나오는 모든 MI 값은 1,000을 곱한 값이다 (MI는 상대적인 크기가 중요하기 때문이다). 참고로 모든 변수들에 대한 MI 값은 다음과 같다.

	고향	거주지	연령	조사시기	학력	소득수준	성별
MI	75	56	37	13	8	7	1

소득수준과 성별요인은 MDI의 p-값 = 0.1 기준을 사용할 때 지지후보를 결정하는 데 영향을 미치지 않는 것으로 판단되어 그림 4.6에는 이 두 변수가 나타나 있지 않다. 조사 시점에 따라 유권자 성향에 차이가 있는가를 비교해 보려면 PCP에서 7월을 선택한다.

7월과 8월의 유권자 성향은 각각 다음과 같은 순서로 나타났다.

- 7월의 성향

	고향	거주지	연령	소득수준	학력
MI	85	75	32	14	11

- 8월의 성향

	고향	연령	거주지
MI	69	47	45

여기서 특이한 내용은 7월에는 고향 다음에 거주지가 강한 영향을 미쳤으나, 8월에는 고향 다음으로는 연령이 거주지보다 더 강한 영향을 미친다는 것이다. 또한 소득수준, 학력, 성별은 두 시점에서 모두 거의 의미 없는 작은 값으로 나타났으며, 특히 8월 들어서는 이들 요인은 유의하지 않은 것으로 나타났다(그림 4.7).

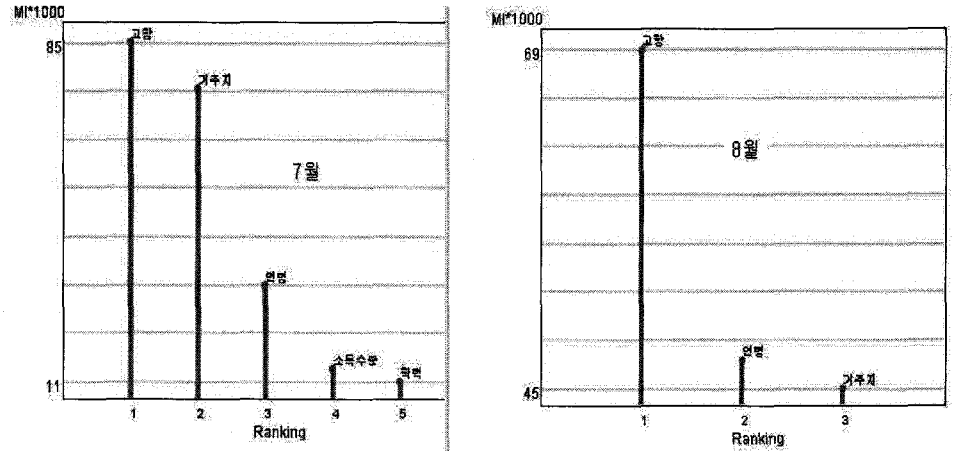


그림 4.7: 7월과 8월의 유권자의 성향(7월(좌측)과 8월(우측) 조사시기가 변함에 따라, 유권자 성향이 달라지는 것을 알 수 있다. 7월에는 고향과 거주지가 매우 많은 영향을 미치고 있으나, 8월에 들어서는 연령이 오히려 거주지보다 더 큰 영향을 미치는 것을 알 수 있다.)

연령층에 따라 유권자의 성향에 어떤 변화가 있는 가를 알아보는 방법에는 여러 가지가 있을 수 있다. 이 중에 한 가지 방법은 연령층을 두 개의 집단으로 나누고, 각 집단별로 유권자의 성향을 살펴보는 방법이다. 연령층을 두 집단으로 나누기 위해 전체 2,000명 집단을

반으로 나눈다. 그림 3.3을 보면, 30대 후반, 또는 40대를 기준으로 하여 지지후보의 경계가 뚜렷하게 나타나는 것을 알 수 있다. 여기서는 편리상 40대를 경계로 두 개의 집단으로 나눈다. 이미 앞에서 설명한 바와 같이 PCP를 사용하여 두 집단으로 나눌 수 있으며, MI 는 자동적으로 해당 집단에 대한 MI로 바꾸어진다. 30대 이전과 40대 이후의 연령층에서 유권자 성향 MI는 표 4.2와 같다.

표 4.2: 30대이전과 40대이후의 성향

	고향	거주지	조사시기
30대 이전	46	35	19
40대 이후	122	102	10

이를 참고하면, 두 연령그룹에서 모두 고향과 거주지가 강한 영향을 미치는 성향인 것으로 나타났고, 30대 이전의 연령층에 비해 40대 이후 연령층에서 지지후보에 미치는 지역적인 영향은 3배 이상인 것으로 나타났다.

이상의 분석 결과 학력은 유권자의 성향에 전혀 영향을 미치지 않는 성향이므로 저학력 노년층과 고학력 젊은 층에 대한 지지후보자의 성향분석은 노년층과 젊은 층의 성향분석과 동일한 결과를 가져올 것이다. 유권자의 성향에 지역색이 강하게 작용하기 때문에 가장 지역색이 강한 지역을 골라 이들을 제외한 다른 지역들만 대상으로 유권자의 성향을 분석해 볼 필요가 있다. 이를 위해 고향이 전라지역인 곳과 경상지역인 곳을 골라 이들 지역을 각각 제외한 후 상호정보를 구한 것이 그림 4.8에 나타나 있다.

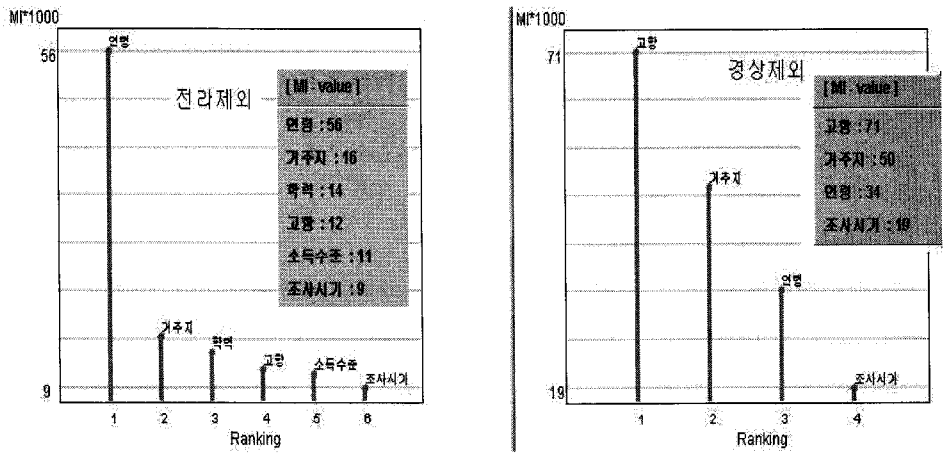


그림 4.8: 유권자 성향 MI 값의 비교(왼 쪽은 고향=전라 제외, 오른 쪽은 고향=경상 제외한 유권자 성향 MI. 이 두 그래프를 비교해 보면 고향의 전라지역을 제외하면 연령이 가장 많은 영향을 미치지만, 경상지역을 제외하여도 고향과 거주지가 강하게 영향을 미치고 있는 것을 알 수 있다.)

5. 선거여론조사 분석결과

2002년도 선거여론조사 자료를 분석한 결과, 지지후보에 많은 영향을 미치는 성향들을 순서대로 나열하면 고향, 거주지, 연령, 시간, 학력, 소득, 성별 등의 순이다. 이들 영향력의 상대적 크기는 고향 75, 거주지 56, 연령 37, 시간 13, 학력 8, 소득 7, 성별 1과 같다. 고향 및 거주지가 영향을 많이 미친다는 것은 이미 알려져 있는 것이지만, 영향을 미치는 성향을 정량화해서 영향력의 크기를 비교해 볼 수 있었다. 7월 조사에서는 고향 85, 거주지 75, 연령 32, 소득 14, 학력 11, 성별 2의 순으로 나타난 반면, 8월 조사에서는 고향 69, 연령 47, 거주지 45, 소득 10, 학력 9, 성별 1의 순서로 나타나 선거일이 다가올수록 나이가 거주지에 비해 더 강한 요인으로 나타나는 것을 알 수 있고, 소득이나, 학력, 남녀별 차이는 미미해 지는 것을 알 수 있었다.

표 5.1: 7월과 8월의 비교

7월	성향	8월	전체
85	고향	69	75
75	거주지	45	56
32	연령	47	37
14	소득	10	7
11	학력	9	8
2	성별	1	1

표 5.2: 서울지역 거주자의 경우

고소득	성향	저소득	전체
217	고향	302	88
158	연령	232	95
61	학력	59	16
37	성별	23	6
37	시간	23	25
4	소득	41	31

고향과 거주지역이 많은 연관성을 가지고 있기 때문에 거주지를 서울로 한정시켜 분석한 결과 연령 95, 고향 88, 소득 31, 시간 25, 학력 16, 성별 6으로 나타났다. 서울에 거주하는 소득이 높은 유권자와 소득이 낮은 유권자를 비교한 것이 표 5.2에 나타나 있다.

6. 결론

본 논문에서는 상호정보와 데이터 시각화를 사용하여 여론조사 결과를 분석하는 방법을 제안하였다. 상호정보를 사용하면 목적변수에 대한 설명변수의 영향을 크기순으로 순위를 매길 수 있고, 데이터 시각화 방법을 사용하면 이들 순위 매김에 대한 평가를 수행할 수 있다. 여론조사에서 목적변수에 미치는 설명변수의 영향력의 크기가 어느 정도인가를 정량화한 것은 이번 연구에 의해서만 이루어진 것이다. 본 논문에서 사용한 선거여론조사 자료에 대해, 지지후보에 영향을 미치는 유권자의 성향을 중요한 순으로 파악 하게 되면 특정 후보자의 선거 전략을 세우는데 도움이 되고 나아가 해당 후보자의 당선 여부를 더 정확하게 판단할 수 있다.

참고문헌

Cang, S. and Partridge, D. (2004). Feature ranking and best feature subset using mutual information, *Neural Computing & Applications*, **13**, 175–184.

Cleveland, W. S. and McGill, M. E. (1988). *Dynamic Graphics for Data Analysis*, Wadsworth & Brooks/Cole.

Cover, T. M. and Thomas, J. A. (2004). *Elements of Information Theory*, 2nd ed., John Wiley & Sons.

Huh, M. Y. and Song, K. R. (2002). DAVIS: a Java-based data visualization system, *Computational Statistics*, **17**, 411–423, http://stat.skku.ac.kr/myhuh/DAVIS.html.

Hutter, M. and Zaffalon, M. (2005). Distribution of mutual information from complete and incomplete data, *Computational Statistics & Data Analysis*, **48**, 633–657.

Lee, S. C. and Huh, M. Y. (2003). A measure of association for complex data, *Computational Statistics & Data Analysis*, **44**, 211–222.

Parzen, E. (1962). On estimation of a probability density function and mode, *Annals of Mathematical Statatistics*, **33**, 1065–1076.

Tan, P. N., Kumar, V. and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns, *In Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 32–41.

[2006년 10월 접수, 2006년 12월 채택]

Information Theory and Data Visualization Approach to Poll Analysis*

Moon Yul Huh[1] Woon Ock Cha[2]

ABSTRACT

A method for poll analysis using information theory and data visualization is proposed in this paper. Questions of opinion poll consist of a target variable and many explanation variables. The type of explanation variables is either numerical or categorical. In this study, explanation variables of mixed types have been ranked according to the magnitude of their effect on target variable by using mutual information. Likewise, the order of explanation variables has been evaluated using data visualization. This is the first study to quantify the impact of specific explanation variable on the related target variable.

Keywords: Association measure, mutual information, data visualization.

* This research was financially supported by Hansung University in the year of 2006.

1) Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea
E-mail: myhuh123@skku.edu

2) (Corresponding author) Professor, Department of Mutimedia Engineering, Hansung University, Seoul 136-792, Korea
E-mail: wcha@hansung.ac.kr