로지스틱 회귀 모형을 이용한 무선인터넷 콘텐츠 서비스의 life cycle 분석 및 예측

*박지홍, 전준현**
*동국대학교 영상정보통신대학원 네트워크관리학과
**동국대학교 정보산업대학 정보통신공학과, 교수/공학박사
e-mail : *wss94@lycos.co.kr, **memory@dgu.edu

A Study on Life Cycle analysis and prediction of Contents Service in the Wireless Internet

*Ji-Hong Park, **Joon-Hyeon Jeon
*The Graduate School of Image & Information Technology
**Dept. of Information and Communications Eng. Professor

Abstract

In this paper, we proposed the technique to estimate the life cycle of Internet content services based on the logistic regression model. In this paper, to define parameters of Internet contents estimating life cycle by logistic regression model, we used market size, traffic amount, page view and session-visit number as the parameters of Internet contents estimating life cycle by logistic regression model. In this paper, to compare the performance of our proposed scheme, we estimated life cycle for the download services of bell sound & character contents in mobile network.

As a result, using our proposed logistic regression, we were able to estimate exactly the life cycle of the download services of bell sound & character contents.

I. 서론

90년대 중반 CDMA 기술의 성공적인 상용화 이후, 우리나라의 이동통신 산업은 급속하게 발전하여 이동통신사는 2004년 9월 말 현재, 3,447만 명의 가입자를 보유하고 있으며, 만 12세 이상 이동전화 보유자의 40.2%인 1,451만 명이 최근 6개월 이내 1회 이상 무선인터넷을 이용해 본 경험이 있는 것으로 나타났고, 또 전체가입자의 94%인 3,251만 명의 가입자가 무선인터넷 서비스를 이용하고 있다. 이동통신의 인프라 구축으로 언제, 어디서나 무선인터넷에 접속하여 M-Commerce, 이동위치기반 서비스를 비롯하여 MMS, 모바일 방송서비스, VOD 서비스, DMB 서비스를 이용할 수 있다. 그러나 많은 유무선 인터넷 콘텐츠 서비스가 생성·소멸되는 가운데, 사업지속을 결정할 콘텐츠의 정확한 Life Cycle을 예측하기가 어렵다.

본 논문에서는 무선인터넷 콘텐츠 서비스 유형 및 시장 규모, 서비스 사업자 경쟁도, 서비스 유형별 시장 점유율, network traffic 점유율, page_view율, 서비스 이용시간 변화율 등으로부터 파라미터를 모델링하여 이를 바탕으로 해당 서비스의 Life Cycle 예측 시뮬레이터를 제안하고자 한다.

II. 본론

본 장에서는 로지스틱 회귀 모형을 이용하여 개별(벨소리/캐릭터, 게임 다운로드, 증권, 뉴스 등) 콘텐츠의 Life cycle을 분석 및 예측을 시도한다. 로지스틱 회귀모형의 장점으로는 현상을 선형(평면 등)으로 반영하여, 추정치에 대한 해석 및 변수 간 영향력 비교가 용이하여 모형의 개발이 용이한 반면, 사전에 입력변수 선택에 대한 탐색이 필요한 단점이 있다.

2.1 로지스틱 회귀모형

변수로 적용될 개인의 연령, 이용 빈도, 이용시간, 개별 콘텐츠 서비스의 이용 요금 등을 이용하여 이에 따른 해당 콘텐츠 서비스의 Life Cycle을 추정, 이 결과에 따라 이동통신사는 특정 콘텐츠 서비스에 대해 접속(이용)할 가능성 여부의 예상, 그 콘텐츠를 더욱 활성화하기 위한 대안과 연령별, 성별 소비자의 성향을 파악할 수 있을 것이다.

이를 위해 몇 가지 가정을 하였는데, 첫째, 각 이동통신 3사의 IR자료와 한국인터넷진흥원에서 조사한 자료를 바탕으로 무선인터넷 콘텐츠 서비스 중 가장 많이 이용되는 벨소리/캐릭터 다운로드 서비스(82.9%, 최근 6개월간의 주이용 서비스는 게임다운로드 서비스)를 대상으로 하였으나, 각 이동통신사가 콘텐츠 서비스(여기서는 벨소리/캐릭터 다운로드 서비스)의 매출액을 공개하지 않는 이유로 무선데이터 ARPU를 벨소리/캐릭터 다운로드 서비스 매출액이라 가정하였다. 둘째, 사용된 매출액이 특정 이동통신사의 매출액지만, 모든 이동통신사의 공통으로 해당되는 매출액이라 가정하였다. 셋째, 이동통신3사에서 발표하는 IR자료가 2001~2002년의 자료는 분기별로 또는 반기별로, 2003년 이후 IR자료는 월별 발표된 자료이기에 자료의 통일성을 위해 2003년 이후 자료만 참고하였다. 마지막으로 한국인터넷진흥원에서 조사한 무선인터넷 사용자(만 12세 이상의 무선인터넷 가입자 40.2%)의 조사결과(중복응답 포함)를 토대로 연령별로 적용하였고, 전체 가입자 대비 비율로 발표된 자료를 가입자로 가정하여 로지스틱 함수에 적용하고자 하였다.

일반화 로짓모형에 근거하여 i번째 부 모집단에서 종속변수 Y가 j번째 범주일 확률 π_{ij}를 다음과 같이 정의한다. x는 독립변수이고, β는 추정될 파라미터이다.

$$\pi_{ij} = \begin{cases} \dfrac{\exp(x_i'\beta_j)}{1+\sum_{l=1}^{k-1}\exp(x_i'\beta_l)}, & j=1,\cdots,k-1 \\ \dfrac{1}{1+\sum_{l=1}^{k-1}\exp(x_i'\beta_l)}, & j=k \end{cases}$$

위의 식을 로짓(logit) 형태로 표현하면,

$$\ln\left(\frac{\pi_{ij}}{\pi_{ik}}\right) = x_i'\beta_j,\ j=1,\cdots,k-1$$ 이 된다. 로짓 변환 이유는 로짓변환은 독립변수 x에 관하여 선형이며, 연속이 되므로 x의 범위에 따라 와 사이의 임의의 값을 가질수 있기 때문이다. 이때 (k-1)(p+1)×1 벡터 $\beta=(\beta_1', \beta_2', \cdots, \beta_{k-1}')$ 는 추정파라미터이다. 각 결과범주의 조건적 확률은 $\pi_j(x) = P(Y=j|x),\ j=0,1,2$ 라고 표현하면 파라미터 벡터 $\beta=(\beta_1', \beta_2')$ 의 함수가 된다.

종속변수가 세 개의 범주인 로지스틱 회귀모형에서 조건적 확률의 식은 다음과 같다.

$$P(Y=j|x) = \frac{e^{g_j(x)}}{\sum_{k=0}^{2} e^{g_k(x)}},\ j=0,1,2$$

이때 로짓 함수 $g_0(x)=0$ 이다.

2.2 로지스틱 회귀모형의 적합

종속변수의 범주가 세 개일 때 파라미터를 추정하기로 하자. 가능도함수를 유도하기 위해 관측값의 그룹을 나타낼 세 개의 변수를 만든다. 이 변수들은 가능도함수를 계산하기 위해서 사용된다. 독립인 n개의 관측값에 대한 조건적 가능도 함수는 다음과 같이 표현할 수 있다.

$$l(\beta) = \prod_{i=1}^{n}\left[\pi_0(x_i)^{y_{0i}} \cdot \pi_1(x_i)^{y_{1i}} \cdot \pi_2(x_i)^{y_{2i}}\right]$$

위의 식에 로그를 취하면 로그 가능도함수는 다음과 같이 유도되는데, $l(\beta)$를 최대화시킬 β의 추정값을 찾기 위해서는 $l(\beta)$에 자연로그를 위한 로그가능도함수를 최대화시킬 β의 추정값을 찾는 것과 동일하기 때문에 로그가능도함수를 이용하게 된다.

$$L(\beta) = \sum_{i=1}^{n}\{y_{1i}g_1(x_i) + y_{2i}g_2(x_i) - \ln(1+e^{g_1(x_i)} + e^{g_2(x_i)})\}$$

여기에서 모든 i에 대해 $\sum_{j=0}^{2} y_{ji} = 1$ 이다. 2(p+1)개의 파라미터에 대해 $L(\beta)$의 미분을 구하면 다음과 같은 정규방정식을 얻게 된다.

$$\sum_{i=1}^{n} x_{ki}(y_{1i} - \pi_{1i}) = 0$$

$$\sum_{i=1}^{n} x_{ki}(y_{2i} - \pi_{2i}) = 0,\ k=0,1,\cdots,p,\ x_{0i}=1$$

여기에서 표기법을 단순화하기 위해 $\pi_{ji} = \pi_j(x_i)$ 라고 표기하였다. 최대가능도 추정량 $(\hat{\beta})$ 은 위의 식을 β 에 대해 연립하여 풀면 얻게 된다.

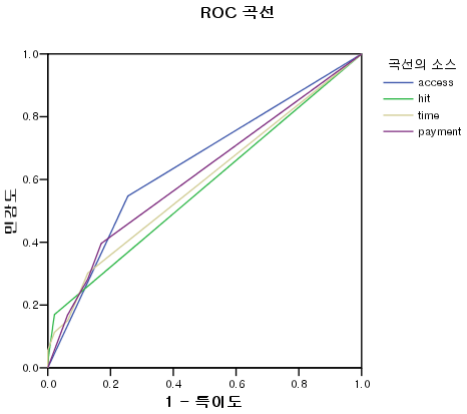
검정결과 변수	영 역	표준 오차	근사유의 확률	근사 95% 신뢰구간	
				하 한	상 한
age	0.390	0.073	0.219	0.247	0.533
time	0.590	0.096	0.351	0.395	0.772
hit(traffic)	0.574	0.041	0.000	0.761	0.921

sex		B추정값	표준 오차	유의 확률	Exp (B)	Exp(B)의 95% 신뢰구간	
						하 한	상 한
1	절편	-1.099	0.667	0.099			
	payment=1	1.296	0.708	0.067	3.656	0.913	14.646
	payment=2	0	1.054	1.000	1.000	0.127	7.893
	payment=3	0.045	0.972	0.677	1.500	0.223	10.077
	payment=4	0(b)					

위에서 설정된 변수와 빈도표에 의거하여, 위의 함수식을 풀어본 결과 값들을 위의 표에 표기하였다.

유의확률의 값이 대체로 작은 것으로 보아 성별에 따른 이용요금 정도에 따른 적합된 것을 알 수 있다. 승산비의 값은 입력변수와 목표변수 사이에서 계산된 값이 1을 넘으면 인과관계가 높은 것으로 파악되고 1보다 낮으면 인과관계가 낮은 것으로 파악되어 수식을 폐기하는 데, 결과 데이터의 값들이 1~3.6 사이의 값들을 가지는 것으로 보아 성별에 따른 이용금액의 모형에 잘 적합 되었다고 볼 수 있고 변수 payment의 값이 1인 즉, 지불요금이 3천원 미만에서는 여성들이 남성들에 비해 승산비가 3.6배나 컸다. 이는 콘텐츠 서비스에 접속하여 3천원 미만의 금액을 지불하는 여성들이 남성들에 비해 3.6배나 높다는 것을 알 수 있다. 이에 해당하는 95% 신뢰구간을 통해 승산비가 약 0.9배에서 14배가 됨을 알 수 있고, 또 이용금액이 3천원~5천원 미만인 경우는 최소 0.1배에서 많게는 7배 정도 여성이 남성에 비해 콘텐츠 서비스를 이용하는 것을 알 수 있다. 이와 같은 분포의 결과로 여성들의 이용금액 정도는 10대에서 50대 이상에 이르기까지 남성에 비해 작게는 0.1배에서 크게는 14배에 이르는 이용 정도를 알 수 있다. 로지스틱 회귀모형을 이용하여 위에서 설정한 변수를 모두를 이용하여 분석할 수 있지만, 본 논문에서는 성별 이용금액을 예로써 모형에 대한 적합도와 유의성 검정을 한다.

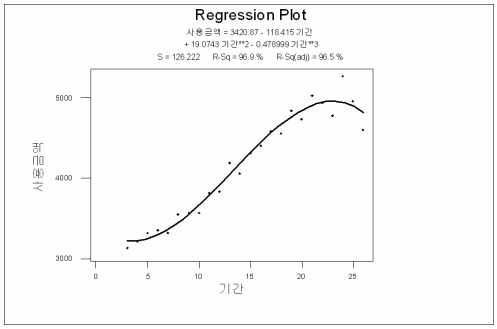
변수에 영향을 주는 공변량(covariate)의 함수로 모형화하고, 결과값을 분석, 모형 적합도를 판단하고, 경계점의 값에 따른 민감도 대 (1-특이도)의 그래프로 적합된 모형의 종속변수의 값이 1인 개체와 그렇지 않은 개체를 얼마나 잘 식별해 낼 수 있는지를 재는 척도인 ROC 곡선 모형으로 검정방법의 식별능력을 알아보았다. ROC곡선은 그래프 아래의 면적이 크면 클수록 검정법의 식별 능력이 뛰어남을 나타낸다. 아래의 결과 ROC 곡선에서는 대체로 변수를 검정방법에 있어서 그래프 아래쪽의 면적이 위쪽의 면적보다 넓은 것으로써 식별능력이 대체로 양호함을 나타내는 것을 알 수 있다.



<그림1> ROC 곡선

위의 자료로부터 개별 콘텐츠의 기간에 따른 이용금액의 변화 추이를 보면 다음과 같은 곡선을 나타낼 수 있다. <그림2>의 그래프는 기간(월수 : 본 논문에서 시작을 2003년 3월로 하였다. 따라서 기간 25가 나타내는 것은 25개월 후인 2005년 4월을 나타낸다) 에 따른(시간이 흐름에 따라) 금액의 상관계수는 96.9%로 위 회귀곡선을 따름을 알 수 있고, 96.9%의 확률로 미래예측이 가능하며 3.1% 벗어날 확률이 있다는 것을 알 수 있다. 이때의 함수식은 다음과 같다.

$$\text{사용금액} = 3420.87 - 118.415 \times \text{기간} + 19.0743 \times \text{기간}^2 + 0.478999 \times \text{기간}^3$$



<그림2>추정된 로지스틱 회귀계수의 그래프

로지스틱 회귀모형을 이용한 콘텐츠 서비스의 life cycle을 예측한 결과 벨소리/캐릭터 다운로드 서비스는 아래의 그래프에 따르면 2003년 3월을 시작점으로 23~24개월이 되는 시점인 2005년 3월~4월을 정점으로 해서 그래프의 하강이 시작되고 있다. 함수식에 데이터를 입력하면 해당 기간의 곡선이 나타나므로 한 Cycle의 그래프로 나타나게 된다. 이로써 벨소리/캐릭

터 다운로드 서비스의 life cycle은 대략적으로 48개월 정도라고 예측 가능하다고 할 수 있다. 이동통신사나 CP 들은 이에 맞추어 killer application과 같은 새로운 콘텐츠 개발을 할 수 있을 것이다. 이를 위해 표본이 되는 정확한 데이터를 더 많이 확보한다면 좀 더 정확한 예측이 가능하리라 예상된다.

2.3 추정된 계수의 유의성 검정 및 해석

여러 변수들의 유의성을 검정하기 위해서는 가능도비(likehood ratio) 검정을 사용해야 한다. 각각의 로짓 함수에 대해 변수를 포함한 모형의 로그 가능도와 상수항만을 포함한 모형의 로그가능도를 비교한다. 계수들이 "0"이라는 귀무가설 하에서 로그 가능도의 변화량에 -2를 곱한 것은 자유도가 "2"인 카이제곱 분포를 따르고 아래의 식과 같다. 검정통계량 G는

G = -2 [(변수를 포함한 모형의 로그 가능도) - (상수항만을 포함한 모형의 로그가능도)] 와 같다. 여기서 구해진 결과값이 1보다 크면 귀무가설을 채택하고 아니면 기각한다, 그리고 변수에 대한 계수들의 유의성을 검정하는 가능도비 검정에 대한 자유도는

[(종속변수의 범주의 수 - 1) * (각 로짓에서 변수에 대한 자유도)] 와 같다.

효 과	축소모형의 -2log 가능도비	카이제곱 α	자유도	유의확률
절 편	16.172	2.690	1	0.101
payment	18.932	5.450	1	0.020

만약 완전한 모형 적합이라면 가능도 함수의 값은 “1”이고 로그 가능도 값이 “0”은 되지만, 보편적으로 좋은 모형이라 함은 관측된 결과의 가능도함수 값이 매우 클 때이며, 로그가능도함수 값은 양수의 값을 가질 때이다. 위 표의 가능도비 검정결과를 보면 검정통계량 G의 값이 18.932로 “0“보다 큰 것으로 나왔기 때문에 추정된 모형이 데이터를 잘 적합된 좋은 모형이라고 볼 수 있다. 카이제곱이 5.450이고 유의확률이 0.020 이므로 α= 0.05 에서 기각된다. 따라서 성별에 따른 이용금액에 대한 차이가 있다고 분석할 수 있다. 유의확률이 0.05보다 크다면 성별에 따른 서비스 이용금액에 대한 남녀 차이가 없다는 것을 나타낸다. 이에 남성보다는 여성이 콘텐츠 서비스에 이용부분에 있어서 이동통신사 관점에서 질적, 양적인 면에서 우수하다고 볼 수 있다.

III. 결론 및 향후 연구 방향

본 연구는 이동통신사에서 제공하는 무선인터넷 콘텐츠 서비스에 대하여 각각의 콘텐츠 서비스에 대한 Life Cycle을 예측하기 위한 것으로, 무선인터넷 콘텐츠 서비스 유형 및 시장 규모, 서비스 사업자 경쟁도, 서비스 유형별 시장 점유율, network traffic 점유율, page_view율, 서비스 사용시간 변화율 등으로부터 파라미터를 모델링하고, 이를 이용하여 정량적으로 평가할 수 있는 해당 서비스의 Life Cycle 예측 시뮬레이터를 설계하고자 하였다. 콘텐츠 서비스 중 가장 높은 점유율을 차지하고 있는 벨소리/캐릭터 다운로드 서비스에 대한 정보를 이용하여 변수를 설정, ROC 곡선을 통한 변수설정에 있어서 검정방법의 식별 능력이 뛰어난지 여부를 알아보았다. 로지스틱 회귀 모형의 적합여부를 평가하기 위해 가능도함수를 이용하여 최대가능도 추정값들을 추정하여, 성별에 따른 콘텐츠 서비스의 이용 빈도, 이용시간, 이에 따르는 트래픽율 등 관련된 요인을 평가하여 해당 서비스의 Life Cycle을 예측할 수 있는 그래프를 추출할 수 있었다.

참고문헌

[1] DEA와 로지스틱 회귀분석을 이용한 정보화촉진기금 융자사업의 효율성 분석. 지유나, 문태희, 손소영,

[2] Hosmer, D. W. and Lemeshow, S. (2000). Applied Logistic Regression. New York : Hohn Wiley and Sons

[3] Kleinbaum, D. G. (1994). Logistic Regression: A Self-Learning Text. Springer-Verlag, New York

[4] SPSS를 활용한 로지스틱 회귀모형의 이해와 응용. 김순귀, 정동빈, 박영술 SPSS아카데미

[5] 2004 대한민국 모바일 연감

[6] 2004 인터넷연감

[7] http://www.mic.go.kr

[8] http://www.nida.or.kr

[9] http://www.ktf.co.kr

[10] http://www.lgtelecom.com

[11] http://www.sktelecom.com