개인 단위 시청률 빅데이터를 이용한 시청률 분석방법 연구 : 정량적, 정성적 데이터를 이용한 드라마 시청률 예측

백영민(연세대학교 신문방송학과)
온병원(군산대학교 통계컴퓨터과학과)
최재호(서울대학교 융합과학기술대학원)
강남준(서울대학교 언론정보학과)

1. 연구목적 및 필요성

우리를 둘러싼 매체환경이 다채널 환경으로 계속 변화하고 있다. 그래서 매일 저녁 많은 사람들이 텔레비전을 켜 수 많은 프로그램 중 어떤 프로그램을 선택할까 고민한다. 더구나 디지털 미디어 활용과 더불어 시작된 폭발적인 가시청 채널의 증가는 수용자들이 마음대로 수많은 채널을 선택할 수 있어 이러한 고민은 더욱 커졌다.

과거 시청률 예측 연구는 대부분 이러한 매체 환경의 변화를 제대로 고려하지 못하고 기껏해야 미국의 경우 ABC, CBS, NBC 3대 네트워크에 대한 시청률 예측이거나(e.g. Rust & Alpert, 1984) 또는 Fox 채널을 추가한 4채널 연구(Napoli, 2001)이었고 위성 채널 등을 추가한 연구는 Patelis et al.(2003) 등의 6채널 시청률 예측 연구밖에 없었다. 현재 우리나라의 경우 케이블을 통해 TV를 시청하는 가구 수가 85%에 이르고 있어 다채널 환경을 적극적으로 고려한 시청률 예측 모델 연구가 필요한 시점이다.

시청률 예측연구에서 기존에 사용된 분석 방법은 프로그램의 내용분류 기준을 사용한 초기의 요인분석(Ehrenberg, 1968; Gensch & Ranganathan, 1974; Swanson, 1967), 다차원 척도기법(Farley &Bowman, 1972; Lehmann, 1971) 등이 주로 사용되었다. 하지만 이러한 방법은 시청률 예측을 한 시점의 데이터만 가지고 분석하는 정적인 연구(static research)이다. 이것은 프로그램의 시계열 연속성과 시청환경 변화 가능성과 같은 동적(dynamic) 요인을 고려하지 않고 있다. 따라서 요일, 계절, 연도별과 같은 동적 변화추이를 분석하는 시계열 회귀분석방법이 제안되었고, 이후 시청률 예측연구의 분석방법은 시계열 회귀분석방법이 주를 이루게 되었다(Danaher, Dagger & Smith, 2011).

하지만 이러한 회귀분석 방법을 사용해서 얻은 예측결과를 실제 시청률과 비교했더니 편차가 다른 기법과 비교해 가장 컸고(Nikopoulos, Goodwin, Patelis & Assimakpoulos, 2007), 회귀분석 특성상 선형적 관계밖에 나타낼 수밖에 없어 다양한 시청률 예측 변인들 간에 존재하는 비선형적인 관계를 반영하지 못하는 문제점이 나타났다(박원기, 김수영, 2003; Webner, 2002). 따라서 시청률 예측에 있어 변인 간 비선형적인 관계를 나타낼 수 있는 다양한 회귀분석 기법(예: ARIMA 모델, trigonometric regression)이나 최근 복잡하게 연결된 데이터 분석에서 사용되는 데이터 마이닝(data mining) 기법, 신경망 분석 방법(neural network analysis), 복잡계 분석방법(complex system analysis) 등을 사용해 시청률을 예측할 필요성이 대두된다.

이러한 시청률 예측은 정확한 광고비 집행에서 아주 주요한 요인이 된다. 제일기획의 보고

에 따르면 2013년 국내 총광고비 규모는 이미 9조 5,000억 원을 돌파하였다(9조 5,893억여 원). 이는 2012년의 9조 3,854억 원보다 2.2% 성장한 수치이다. 지상파 광고시장은 2013년도 -5.4% 감소에서 2.2% 성장세로 돌아섰고 전년도에 비해 낮아지기는 했지만, 여전히 지상파 광고비 규모는 전체 광고시장의 18%를 차지하여 인터넷 매체에 뒤이어 2위를 고수하고 있다.

광고요금은 광고되는 프로그램 시청률과 긴밀히 연결되어 있다. 단순히 말해 시청률이 높게 나오고 시청자 규모가 큰 프로그램에는 높은 광고요금이 붙게 마련이다. 흥미로운 점은 광고요금 산정에 사용되는 시청률은 실측된(observed) 시청률이 아니고 예측된(predicted) 시청률이라는 것이다. 나폴리(Napoli, 2001)에 의하면 1995년 이미 미국 방송광고시장의 75-80%가 업프론트(up-front) 방식으로 선매되었는데 이 때 사용하는 프로그램 별 예측 시청률은 아주 부정확하고, 특히 새로 시작하는 프로그램(new PGM)에 대해서는 이전에 비교대상 프로그램이 없어 더욱 편차가 크다고 하였다(e.g. Rust & Echambadi, 1989, p. 13). 나폴리가 미국 4대 네트워크를 대상으로 예측 시청률과 실제 시청률간의 차이의 절대값(과다추정의 경우: +, 과소추정의 경우: -)을 계산한 결과 평균적으로 21% 가량 차이다 생겼다.

따라서 정확하게 예측된 시청률은 광고효과의 불확실성(uncertainty)을 감소시키고 광고시장의 효율성을 증진시킨다. 그러나 광고비 시장의 중요성과 규모에도 불구하고, 현재 통용되는 시청률 예측모형의 정확성은 그다지 높지 못한 것이 사실이다. 최근의 연구결과에 따르면 예측력이 낮은 시청률 예측모형으로 인해 미국에서는 매년 최소 2억 5천만, 최대 5억 8천 6백만 달러의 광고비가 낭비되고 있다고 한다 (Danaher, Dagger, & Smith, 2011).

따라서 이 연구의 목적은 ① 효율적인 광고비 집행을 위한 정확한 시청률 예측 방법과, ② 다채널 시대의 복잡한 경쟁 환경을 고려한 시청률 예측 분석모델을 제안하며, 선형적 추정에만 머물러 왔던 기존 연구방법을 최신의 ③ 데이터 마이닝(data mining) 기법, 예측 알고리즘 등을 사용해 비선형적 관계까지 고려한 시청률 예측 계량모델을 개발하는데 있다.

2. 연구내용

기존 시청률 예측모형 관련연구는 일반회귀분석에 기반하고 구조적, 계절적 프로그램 특성 변수들을 발굴하고 테스트하는 것에 중점을 두어왔다. 그러나 최근 시청률 예측모형 관련 연구에서 일반회귀분석 모형의 한계점이 나타나면서 시청률 예측연구는 언론학이나 매체경

제학을 넘어 전산학이나 기계학습분야(machine learning)에서도 연구되고 있다. 이러한 새로운 연구경향을 크게 세 가지로 압축될 수 있다.

① 데이터마이닝(data-mining): 데이터에 기반한 모형의 가능성이 지속적으로 제기되고 있다. 기존의 일반회귀모형에서는 모형에 대한 이론적 고찰을 바탕으로 예측모형을 구성하였지만, 최근에는 서포트 벡터 머신(support vector machine), 인공신경망 등의 기계학습 기반의 예측모형이 테스트되고 있다. 연구결과를 정리하기는 이르지만, 대개의 연구들은 데이터에 기반한 예측모형의 예측력이 일반회귀분석모형과 비교하여 비슷하거나 약간 나은 수준임을 보여주고 있다. 데이터마이닝 기반 예측모형의 특성상 데이터가 보다 많이 축적될수록 예측력은 조금씩 나아질 것이라고 예상할 수 있다.

② 의사결정 도우미 시스템(decision support system): 학술연구의 입장에서는 예측모형에서 도출된 예측시청률이 "예측된 시청률"이지만, 업계의 입장에서 예측시청률은 과거 데이터의 패턴에 근거한 "참고용 시청률"에 불과하다. 즉, 광고주와 방송사의 입장에서 예측시청은 고려할 여러 사항들 중의 하나에 불과하다. 물론 과학적 기법에 기반한 예측시청률이 가장 '객관적(objective)' 시청률임에는 틀림없는 것도 사실이다. 이러한 입장에서 데이터베이스 연구자들은 시청률 예측모형을 "의사결정 도우미 시스템(Decision Support System)"으로 이해하려한다(Yin et al., 2013). 다시 말해 불확실한 상황에서 '예측시청률'을 설정하기 위한 참고자료로서의 "예측된 시청률"을 의미한다. 방송환경이 급변하고 있기 때문에, 환경적 요소(어떤 맥락에서 방영되는가 등) 그리고 제작 전문가에 의한 정성적 요소가 반영될 수 있다는 점에서 예측시청률을 '참고자료'의 하나로 접근하는 것은 합리적일 수 있다.

③ 추천 시스템(recommendation system): 아마존(Amazon.com)에서 책을 구매하면 판매자가 유사한 몇 몇 책들을 추천해주고 있다. 특히 인터액티브 시스템(이를테면 소셜 미디어)의 등장으로 추천시스템은 상당히 광범위하게 퍼져있다. 시청률 예측이라는 측면에서 최근 전산학에서는 추천시스템과 IPTV 시청률 간의 관계를 탐구하고 있다(Bambini, Cremonesi, & Turrin, 2011). 시청률 예측 관점에서 이러한 추천시스템은 상당히 흥미로운 사례일 수 있다. 백영민(2012)은 추천시스템의 아이디어를 바탕으로 k순위 최인접사례(kNN) 짝짓기 기법을 이용하여 시청률 예측을 시도한 바 있다. 특히 데이터의 축적이 가속화될수록 추천시스템의 효과가 증진된다는 점에서 향후 발전가능성이 높다고 하겠다.

따라서 이 연구에서는 아래와 같은 점들을 감안하고 진행하였다,

- 빅데이터(Big data)의 축적 및 처리: 현재 피플미터(People-meter)에서 수집되는 데이터만 해도 상당한 용량이다. 특히 모바일이나 온라인 환경과의 데이터 연동성을 고려할 때, 대용량

데이터의 효과적 축적(storing)과 처리(processing)는 필수적이다.

- 일반회귀분석을 넘어 대안적 예측모형의 타당성 검증: 현재의 시청률 예측모형의 분석단위(unit-of-analysis)는 프로그램 단위인 경우가 많으며, 예측모형은 일반 선형회귀분석을 쓰는 경우가 많다. 그러나 피플미터에서 데이터를 수집할 때의 분석단위는 '시청자'이며, 방송환경의 변화로 일반선형회귀분석의 선형성가정을 충족하지 못하는 경우도 적지 않을 것이다. 따라서 분석단위를 다변화한 시청률 예측모형을 시도할 필요가 있으며, 대안적 예측모형들(예를 들면, kNN, HMM 등)과 일반회귀모형의 예측력 비교도 매우 필요하다.
- 정성적 정보의 활용: 정량적 자료에 근거한 대부분의 예측모형이 그러하듯, 프로그램의 정성적 특징(이를테면 드라마의 경우 연출가, 작가의 능력, 주/조연의 유명세 등)은 계량화 대상이 못된다. 출연자, 작가 또는 연출자의 미묘한 특성이나 영향력이 시청률을 예측하는데 결정적 요인이 될 수 있다는 사실은 프로그램 제작 실무자들은 모두 알고 있다. 하지만, 현재의 예측모형 프레임에서는 정성적 지식을 모형에 반영하기가 쉽지 않다. 베이지안 통계학 전통에서는 객관적 데이터와 주관적 판단의 조합을 중시하고 있는데, 실제로 최근의 연구에서는 현장제작자의 정성적 데이터를 모형에 투입하는 것이 모형의 예측력을 증진시킨다고 나타났다(백영민, 2012).

  통합적 프로그램 시청률 예측모형을 완성하기 위해서는 프로그램 제작과 관련된 다양한 정보에 대한 정성적 판단이 필요하다. 따라서 이러한 정성적 판단을 가능케 해주는 시스템적 접근 필요성이 제기되고 이것은 프로그램과 관련된 여러 정보 단위의 Linked Data system을 구축함으로서 해결할 수 있다. Linked Data는 시맨틱 웹(semantic web)이 표방하는 데이터의 웹(Web of Data) 세상을 만들기 위한 구체적인 방법이다. 시맨틱 웹은 단지 웹을 데이터로 만드는 것을 넘어서서 데이터 간의 링크(link)를 만들면 데이터의 상호운용성 극대화 및 데이터 통합을 통한 데이터 리소스(data resource)의 자유로운 접근 및 이용을 쉽게 해 정성적 판단의 기본 자료로 유용하게 사용할 수 있다

마지막으로 이 연구에서 예측을 시도한 시청률은 드라마 장르로 한정하였다. 현재 우리나라 방송 편성구조상 광고주들이 가장 관심 있는 장르가 드라마이고 방송사 수익구조면에서도 가장 중요하다. 또한 최근 '별에서 온 그대' 신드롬에서도 확인된 것처럼 어떤 드라마가 '뜰' 것이냐에 대한 비교적 정확한 예측은 방송사, 광고주, 제작자 모두에게 아주 중요한 정보가 될 것이다. 따라서 최종 예측은 방영 예정인 드라마의 초회분(1회분) 시청률을 대상으로 하였다.

3. 국내외 연구동향

시청률 예측 연구는 사실상 광고주와 광고회사들이 TV 광고를 구매하는데 결정적 영향을 미치기 때문에 아주 중요하고, 그래서 다수 이러한 목적으로 수행되었다고 왔다고 추정된다. 하지만 이것이 학술적 논문으로는 거의 발표가 되지 않고 발표가 되더라도 예측기법의 중요한 부분은 제외하고 일반적 결과만 보여준다(e.g.; Nikopoulos, Goodwin, Patelis & Assimakpoulos, 2007). 일찍이 Gensch & Shaman(1980)은 이러한 경향은 시청률 예측연구가 대부분 광고회사들의 청탁연구이기 때문에 연구결과(자신의 1980년 연구결과를 포함해)가 사적소유(propriety)로 취급받기 때문에 나타난 현상이라고 하였다.

이러한 비밀에 싸인 연구들은 대부분 내부적 know-how에 의존해 시청률 추정공식을 만들어 사용하고 있으며, 여기서 가장 큰 영향을 주는 변인들은 프로그램의 내용에 대한 전문가의 판단이라고 하였다(Napoli, 2001). Napoli는 이러한 비밀주의가 시청률 예측 편차를 평균 20% 이상 나오게 하는 주된 원인이라고 비판하였다(p. 58)

시청률 예측연구에 대한 기존 연구결과를 정리하면서 데이나허 등도 비슷한 견해를 나타냈다(Danaher, Dagger & Smith, 2011, p.5). 아래 표에 나타난 것처럼 1980년 대 초에 마케팅, 광고학 저널에 집중적으로 실리던 시청률 예측 논문들이 사적 소유물 문제 때문에 더 이상 나타나지 않다가 다시 2000년대 들어서 약간 활기를 띤다고 하였다(p. 6-7).

아래 <표 1>에서 나타난 그 동안 이루어진 계량적 시청률 예측 연구의 특징은 분석 대상 채널이 아주 적다는 것이다. 그리고 분석 방법은 대부분 회귀분석 기법을 사용하였고, 피플미터 등에서 수집한 시청로그 데이터를 사용한 경우는 드물었으며, 서베이나 다이어리를 사용한 예가 많았다. 예측 변인으로는 주로 시청률의 계절적, 주간 변동 변인을 사용해 시계열적 추이를 분석하였다.

데이나허 등(Danaher, Dagger & Smith, 2011)은 이러한 기존연구의 문제점을, 다채널 상황에서 벌어지는 다양한 경쟁 환경을 고려하지 않고 적은 채널만 대상으로 예측한 점, 다양한 시청예측 변인간의 비선형적 관계를 고려할 수 있는 신경망분석이라던가 데이터 마이닝 기법과 같은 최신 분석기법을 사용하지 않은 문제(Webner, 2002만 제외), 그리고 전반적으로 예측 추정 값과 실제 시청률간의 차이가 커(MAPE와 MAD의 값) 예측이 정확치 않다 비판하였다.

그러나 다채널 상황이 가속화하면서 소수 지상파채널의 독과점 현상이 붕괴되는 과정을 통해 시청 상황이 최근 급격히 변하였다. 수십, 수백 채널의 선택 가능성을 가진 시청자들에

게는 더 이상 프로그램 장르, 편성구조, 시청시간대와 같은 구조적 요인들이 중요한 변수가 되지 못한다. 그 결과 개인의 프로그램 선호요인이 시청률 연구에서 중요한 변인으로 대두되었다(Eastman, 1998). 따라서 프로그램 장르를 시청자의 관점에서 재구성하려는 시도나 시청자들을 지역적, 인구통계학적 변인을 사용해 다양한(heterogeneous) 시청 집단으로 분류(segmentation)하려는 시도는 개인 선호요인을 시청률 예측 연구에서 사용하려는 시도라고 볼 수 있다.

시청률 예측연구에서 피플미터 등에 의해 수집된 정확한 시청로그(viewing log) 데이터를 사용할 경우 예측모델에서 편성시간대, 프로그램 장르와 같은 구조적 변인밖에 사용할 수 없다. 개인적 속성변인을 사용하기 위해서는 서베이와 같은 방법으로 직접 데이터를 수집해야하는데 피플미터에서 수집한 시청률자료를 분석하는 경우에는 이와 같은 개인적 변인의 사용이 인구통계학적 정보로 제한된다. 또한 이러한 데이터가 주로 가구단위의 집합적(aggregated) 데이터이기 때문에 개인적 변인을 사용하기 힘들었다. 그러나 피플미터에 의해서도 단순한 개인별 인구통계학적 변인이 수집되기 때문에 개인 분석단위의 연구를 수행하기도 했다(Rust, Wagner, Kamakura & Alpert, 1992).

메이어 등(Meyer & Hyndman, 2005)은 시청집단 분류(segmentation)를 사용한 시청률 예측과 전체집단의 통합시청률(aggregated rating)을 적용한 예측 결과를 비교하였다. 이들은 전체시청률 예측모델(population rating model), 분화된 집단별 시청률 예측모델(segment rating model), 개인별 시청행태 분석모델(individual viewing behavior model) 등을 분류해 선형적 회귀분석 방법, 비선형적 관계를 분석하는 결정나무모형(decision trees model), 신경망분석 기법 등을 사용해 비교하였다. 연구결과 연령별, 시청시간대별, 교육 수준별, 선호채널별 등 시청률을 예측하는데 중요한 변인으로 기존 연구에서 밝혀진 것들을 기준으로 해서 중범위 분류(medium class segmentation)를 반영한 추정이 가장 정확했고, 추정방법은 선형적 회귀분석 방법보다는 비선형적 기법이 더 우수한 것으로 나타났다.

베버(Weber, 2002)는 다채널 환경으로 급격하게 변화하고 있는 독일 시청환경에서 TV 시청률을 예측하는 모형에 대해 논의하였다. 이러한 예측은 선형적 모델로도 비선형적 모델로도 추정 가능하다고 주장하였다. 통상 커뮤니케이션 학자들은 시청률을 예측(forecasting)하기보다는 설명(explanatory)하기를 원한다. 따라서 복잡하고 다양한 시청행태 모델링을 통해 시청률을 분석하고자 시도한다. 하지만 시청률을 예측할 때는 이 모든 설명변인들이 필요하지 않고, 경우에 따라서는 이론적으로는 유용한 설명변이이지만 예측 정확도를 떨어뜨리는 변인도 있을 수 있다. 이러한 예로베버는 날씨 변인을 들었는데 이 변인은 야외활동

이 불가능한 날씨는 시청률을 상승시킴으로 아주 중요하다. 하지만 날씨는 사전에 어떤

<표 1> 기존의 시청률 예측 조사연구 비교

Comparison of the previous television ratings forecasting literature.

Study	Chans	Country	Date	Time	Predict. object	Model dep. var.	Data points Est	Data points Val	R^2 %	MAPE[a]	MAD[b]	Forecast length	Data method	Samp size	Genre	Forecast method	Important variables	Opt schd
Cattin et al. (1994)	4	France	10/92–7/93	All day	Qtr hrs	PUT × shr[c]	7056	480	n/a[d]	13.0	1.1	5 wks	*p*-meter	n/a	Y (n/a)	Historical	Prog type, day, time, week	N
Darmon (1976)	3	Canada	1/73 2 wks	6 pm–12 am	Progs	Ratings	180	180	72	n/a	3.4	2 wks	Diary	250	Y (11)	Regression	Channel, prog type	N
Danaher and Mawhinney (2001)	3	NZ	6/94	6–10 pm	Progs	Ratings	108	6	n/a	45.7	1.8	2 wks	Diary	164	N	Logit model	Lead-in, program dummies, program length	Y
Gensch and Shaman (1980)	3	US	11/66–8/69	7:30–10 pm	Qtr hrs & progs	PUT × shr	916	20	89	n/a	2.5	4 wks	Diary	n/a	N	Regression	Day, time, month	N
Henry and Rinne (1984)	3	US	10/81–3/82	8–11 pm	Progs	Shr only	525	102	81	18.2	2.35	9 wks	Diary & survey	1500	Y (12)	Logit model	Lead-in, lead-out, years on TV, changed time, prog type & preference	N
Horen (1980)	3	US	9/69–4/74	8–11 pm	Half hr ratings	PUT × shr	4130	0	70	n/a	n/a	n/a	Diary	n/a	Y (3)	Regression	Last year's ratings, competing progs, day, time, lead-in	Y
Kelton and Schneider-Stone (1998)	3	US	81–89 12 wk	7–10 pm	Progs	Ratings	53	55	70	n/a	3.6	1 wk	Diary	n/a	Y (8)	Regression	Lead-in, prog type, day, time, channel, prog attributes	Y
Napoli (2001)	4	US	1/93–12/98	8–11 pm	Progs	Shr only	140	0	19	21.4	15.2	N/A	Diary	n/a	N	Regression	Lead-in, lead-out, year	N
Patelis et al. (2003)	6	Greece	4/99–3/00	All day	Progs	PUT × shr	n/a	36 269	n/a	13.6	n/a	12 wks	n/a	n/a	Y (8)	Time series	Day, time, month, holidays, prog type	N
Reddy, Aronson, and Stam (1998)	1	US	1/90–3/90	8–11 pm	Progs	Ratings	338	0	93	n/a	n/a	n/a	n/a	n/a	Y (5)	Regression	Day, time, prog type, duration of show, attractiveness	Y
Rust and Alpert (1984)	3	US	9/77–11/77	5–11 pm	Half hr & progs	Ratings	34	34	93	n/a	2	2 days	Survey–Simmons	5434	Y (5)	Logit model	Prog type, audience flow	N
Tavaloli and Cave (1996)	4	UK	3/90 1 wk	5:30–10:30 pm	5 mins	Shr only	500	24	92	12.1	2.0	2 wks	*p*-meter & diary	n/a	Y (29)	Logit model	Lead-in, prog type, channel, audience	N
Van Meurs (1994)	5	Holland	12/92–1/94	All day	n/a	PUT × shr	4200	500	n/a	31.0	0.9	1 month	*p*-meter	n/a	Y (n/a)	Historical	Prog type, channel, day, time, month	N
Weber (2002)	1	Germany	1/95–4/97	6–11 pm	Progs	Ratings	n/a	0	81	25.2	n/a	2 months	*p*-meter	12000	N	Neural net	n/a	N
Yoo and Kim (2002)	4 US 3 Kr	US Korea	9/94–9/97	8–11 pm	Progs	Shr only	0	105 US 88 Kor	n/a	20.2 US 31.2 Kr	n/a	12wks	n/a	10 US 6 Kor	N	Judgment	Lead-in, day, prog type	N

[a] Mean absolute percentage error.
[b] Mean absolute deviation.
[c] PUT = People using Television, and shr = channel share.
[d] n/a = not available or not reported in the study.

방법을 써도 정확히 예측하기 힘들기 때문에 이 변인을 예측모델에 넣었을 경우 오히려 예측 정확도를 떨어뜨린다.

베버는 Forward feed algorithm을 사용해 신경망 분석으로 상업 채널인 SAT1의 시청률을 예측하였다. 1995년 1/1부터 4/30까지 데이터를 추정, 학습용으로 사용하고 5/1부터 6/30까지 시청률을 단계적으로 추정, 즉 직후 2주간의 단기추정(5/1--5/14), 그리고 한 달간의 기간을 둔 장기주청(6/17--6/30)을 하였다. 신경망 분석과 비교하기 위해 일반 선형회귀분석도 수행했는데 두 분석기법간 차이는 크게 나지 않았다. 베버의 연구에서 주목할 만한 점은 MAPE(mean absolute percentage error)로 측정한 예측정확도가 추정기간 직후 투

입한 단기모델에 비해 한 달 후 시청률을 예측한 장기추정의 경우 30% 이상 커졌다는 점이다. 이러한 결과는 신경망 분석에 의한 추정모델은 시간이 경과하면서 그 정확도가 떨어짐으로 계속 데이터를 투입해 최근 추정형태로 재학습을 시킬 필요가 있다는 점을 시사하고 있다.

우리나라의 시청률 예측 연구는 대략 1980년대에 시작되어 최근까지 이어지고 있다. 피플미터 데이터를 사용해 본격적으로 시청률 예측한 최초의 연구는 서울 마케팅 서베이(SMS)의 1990년 1, 3, 5, 7, 9, 11월 6개월 치의 시청률을 이용해 방송국, 요일, 시급, 방영시간, 장르 등을 더미변수화(dummy variables)하여 회귀모형을 통해 시청률의 예측가능성을 제시한 이혜갑(1994)의 연구이다. 그 후 박원기와 김수영(2000)은 시계열모형과 회귀모형을 이용해 시청률을 예측하는 연구를 수행하였다. 시계열 모형을 이용해서는 4주 이하의 단기 시청률을 예측하였고, 회귀모형을 이용해서는 채널별 프로그램의 장르, 요일, 시간대, HUT(house using television) 등의 변수를 더미(dummy)변수화 하여 장기 시청률을 추정하였다. 위의 국내 두 연구는 모두 집합수준의(aggregate level) 시청률을 예측하는 모형에 해당한다.

경제학의 효용이론에 근거를 둔 다항 로짓 선택모형(multinomial choice model)으로 시청률을 분석한 정진욱(2003)의 연구에서는 38개의 시청자 유형별 시청률 예측모형을 구축하여 분석을 수행한 결과 주요 영향요인이 프로그램 유형, 방영시간대, 프로그램의 질, 채널 충성도로 나타났다. 이 연구에서 중요하게 발견된 내용은 바로 시청률 결정 요인 및 모형의 예측력이 시청자 유형별로 매우 상이하게 나타났다는 점이다.

박노성과 송석현(2004)은 SUR(Seemingly Unrelated Regression) 회귀모형을 사용하여 동 시점에서 발생하는 채널간, 프로그램 간 상관관계를 고려해 드라마의 시청률을 추정하였다. 이 연구에서 MBC 드라마의 경우, 각 드라마에서 전 회차 시청률과 현 회차 시청률 사이에 강한 선형관계가 존재하고 있는 것으로 나타났고, SBS 드라마의 경우도 드라마에 따라 전 회차 시청률과 현 회차 시청률 사이의 연관성이 MBC만큼 강하지는 않지만 약간 나타났다.

시청률 예측에 대한 국/내외 연구결과를 종합해 시청률 예측모형 구축에 대한 함의를 정리하면 다음과 같다.

① 다채널 경쟁환경을 시청률 예측 모형에 투입하는 방안을 고려해야 함.

② 시청률 예측은 선형적 관계뿐만 아니라 비선형적 관계도 고려해 추정해야 함.

③ 다양한 비선형 알고리즘 중 어느 것이 분석 대상 데이터에 가장 적합한지는 시뮬레

이션을 통해 최소 편차(Min. of MAPE 또는 MAE)를 가져오는 것을 찾아야 함.

④ 추정된 예측계수는 연속적으로 새로 투입되는 시청률 데이터에 의한 피드백을 통해 재학습되어(machine learning) 수정되어야 예측 오차를 줄일 수 있기 때문에 시청 예측모델을 새로운 자료의 피드백 루프(feed-back loop)를 통한 연속적 머신러닝 (continuous machine learning) 모델을 구축해야 함.

⑥ 이들 변인을 투입해 예측공식을 만들 경우 이론적 설명력이 우선인 시청률예측 설명 모델(explanatory model)과 예측 정확도가 생명인 예측모델(forecasting model)의 최적 변인 군을 각각 달리 결정해서 추정해야 할지를 사전에 결정해야 함.

⑦ 시청률 추정 시 전체 집합수준(aggregated)에서 모델을 구축하지 아니면 어떤 세분화 기준으로 하위 샘플을 분류(segmentation)해 다양하게 추정할지를 결정해야 함.

4. 시청률 데이터 전처리과정: 빅데이터 분산처리 알고리즘의 활용

시청률 예측을 위해 닐슨의 3년 치 개인 단위 시청률 데이터를 사용하였다. 이 데이터의 구조는 첨부한 <부록 1>에 나타나 있다.

1) 개인단위 시청률 분석을 위한 MapReduce 알고리즘 제안

이 연구의 제안방안은 (1) 원자료 데이터 전처리 (raw data pre-processing) (2) 시청률 예측 모형의 구축 (3) 드라마 시청시간 예측 단계로 구성된다. 닐슨의 시청률 데이터는 빅데이터 처리 시스템의 하둡 분산 파일시스템에 저장되어 있다. 닐슨 시청률 원자료 (raw data)는 가구(household), 패널(panel), 시청(viewing), 프로그램(drama) 등 크게 4개의 관계형 테이블로 구성되며, 첨부된 <부록 1>에 자세히 수록되어 있다.

1.1) 원자료 데이터의 전처리 과정

전처리 단계는 예측 알고리즘을 사용할 수 있도록 원시 데이터(raw data)를 가공한다. 하둡 분산 파일시스템에 있는 프로그램 테이블과 시청 테이블을 조인(join)하여 <표 2>와 같은 새로운 테이블을 생성하고 이것이 예측모형 입력 데이터로 사용된다.

<표 2> 예측모형의 입력 데이터

시청자ID	시청일자	드라마명	채널코드	진행률	러닝타임(분)	3월~5월 방영	6월~8월 방영	9월~11월 방영	월화드라마	수목드라마	주말드라마	22시 이후 방영	21~22시 방영	19~21시 방영	KBS1 시청시간(분)	KBS2 시청시간(분)	MBC 시청시간(분)	SBS 시청시간(분)
1234917aa	20100701	황금물고기	MBC	0.01	36	0	1	0	1	1	0	0	0	1	0	2	25	6
1234917aa	20100701	바람불어좋은날	KBS1	0.01	36	0	1	0	1	1	0	0	0	1	0	1	29	3
1234917aa	20100701	로드넘버원	MBC	0.05	71	0	1	0	0	1	0	1	0	0	23	0	0	0
1234917aa	20100701	제빵왕김탁구	KBS2	0.04	73	0	1	0	0	1	0	0	0	1	10	20	3	0
1234917aa	20100701	세자매	SBS	0.01	37	0	1	0	1	1	0	0	0	1	10	20	3	0
1234917aa	20100701	나쁜남자	SBS	0.09	72	0	1	0	0	1	0	1	0	0	23	0	0	0

<표 2>를 보면 시청자 1234917aa(1234917: 시청가구 ID, aa: 가구 내 시청자 식별 ID)는 2010년 7월 1일에 MBC, KBS1, KBS2, SBS에서 방영했던 여러 드라마를 시청했음을 알 수 있다. 예를 들면, 1234917aa는 MBC의 일일연속극인 '황금물고기'를 시청하였다. '황금물고기'는 2010년 7월 1일 오후 7시 ~ 9시 사이에 방영되었고, 1234917aa는 이 드라마를 25분 시청하였으며, MBC에서 황금물고기가 방영되는 동안 KBS2와 SBS를 각각 2분과 6분을 시청하였다.

<표 2>에서 시청자가 시청한 드라마는 벡터(vector)로 표현할 수 있고, 각 칼럼(column)은 피처(feature)로 나타낼 수 있다. 따라서 벡터(예: 황금물고기)는 19개의 피처들을 가진다. 특히 7번째 피처(드라마가 3월 ~ 5월 사이에 방영 유무 표시)부터 15번째 피처(드라마가 오후 7시 ~ 9시 사이에 방영 유무 표시)까지는 계절요인(seasonal factor)로 이진 값(binary value)을 가진다. 즉, '황금물고기' 드라마는 6월에서 8월 사이에 방영되었기 때문에, 8번째 피처는 1값을 가지고, 7번째와 9번째 피처는 0값을 가지게 된다. 전처리 단계에서는 프로그램 테이블의 프로그램 시작시간(start time)과 프로그램 종료시간(end time)을 사용하여 7~15번째 피처들의 바이너리 값을 계산한다. 그리고 16번째 피처에서 19번째 피처는 시청자가 '황금물고기' 시청 시간 동안 다른 방송국의 프로그램을 본 시청시간을 가리키며, 전처리 단계에서는 원시 데이터를 가공하여, 시청자가 시청한 방송사의 시청 시간

을 측정하게 된다.

이와 같이, 전처리 단계에서는 드라마의 상영 날짜를 이용하여 월, 요일, 방송 시간대에 대한 이진 값을 구하고, 드라마가 방영되는 동안, 시청자가 시청한 다른 방송국의 시청 시간을 계산한다. 이러한 작업을 수행하기 위해, 전처리 단계에서는 원시 데이터의 관계형 테이블을 빠르게 조인(join)하는 것이 필요하다. 이를 위해, 본 연구에서는 하둡 에코 시스템(Hadoop ecosystem)에서 제공하는 Hive 언어를 사용하여 병렬처리를 통해 실행시간을 크게 단축하였다. Hive는 MySQL의 SQL-like language와 유사하게 select-from-where 문을 사용하여 쉽게 데이터를 조인할 수 있다(Edward Capriolo et al, 2012). End-user는 MySQL과 유사한 질의문(query)을 통해 하둡 분산 파일시스템에 저장되어 있는 데이터를 조인하고 가공하여 <표 2>의 데이터를 빠르게 생성할 수 있다. 특히 사용자는 병렬처리에 대한 이해 없이 Hive에서 제공하는 질의문을 작성하면, 하둡 분산 시스템은 사용자의 질의문을 자동으로 MapReduce로 변경하여 실행하게 된다. Map 단계에서는 하둡 분산 파일시스템에 분산 저장되어 있는 원시 데이터에서 계절요인, 시청시간 등과 같은 예측변인 값을 계산한다. 그리고 Reduce 단계에서 이렇게 산출된 데이터 노드들의 출력 데이터를 병합(join)하는 작업을 수행하게 된다[1].

1.2) 방영 예정 드라마의 초회분 시청시간 예측

그 다음 '황금의 제국' 드라마에 초회분 방영에 대한 시청자들의 SBS 채널의 시청시간을 예측하기 위해, k-Nearest Neighbour (k-NN) 알고리즘을 사용한다(Sean Owen et al, 2011; Foster Provost et al, 2013). <표 3>의 기준벡터가 입력으로 주어지면, SBS에서 방영되었던 모든 드라마에 대한 벡터들과 기준벡터 간의 Euclidean distance를 구하게 된다.

SBS에서 방영되었던 모든 드라마들에 대한 벡터들을 균등하게 나누어, 여러 개의 Map 프로세스들에 전달한다. 예를 들면, SBS에 방영되었던 모든 드라마들에 대한 벡터들을 $v_1, v_2, v_3, v_4, v_5, v_6$ 라 하고, 기준벡터를 v_i라고 가정하며, 2개의 Map task(M_1, M_2)들을 실행시키면, M_1에 v_1, v_2, v_3, v_i가 저장되고, M_2에 v_4, v_5, v_6, v_i이 저장된다. 각 Map task에서는 $dist(v_i, *)$를 계산한다. 그리고 key는 시청자 ID, value는 시청시간과 distance로 설정하여 Reduce task에 보낸다. 이를 테면, M_1과 M_2로부터 Reduce task는 (key=141083ad, value={SBS 시청시간=12분, distance=1.62})와 같은 key-value 값을 받게 된다. Reduce task는 M_1과 M_2로부

1) 이 연구에서 사용한 닐슨의 시청률 원자료는 27GB로, 이러한 대용량 데이터를 한번에 join하는 것은 많은 시간이 소요된다.

터 받은 value들을 정렬하여, 가장 distance 작은 k개의 벡터를 선택하고, k개의 벡터에 있는 SBS 시청시간을 구한다. <표 3>은 k-NN을 병렬 처리하기 위한 MapReduce 알고리즘을 나타낸다.

<표 3> k-Nearest Neighbor MapReduce 알고리즘

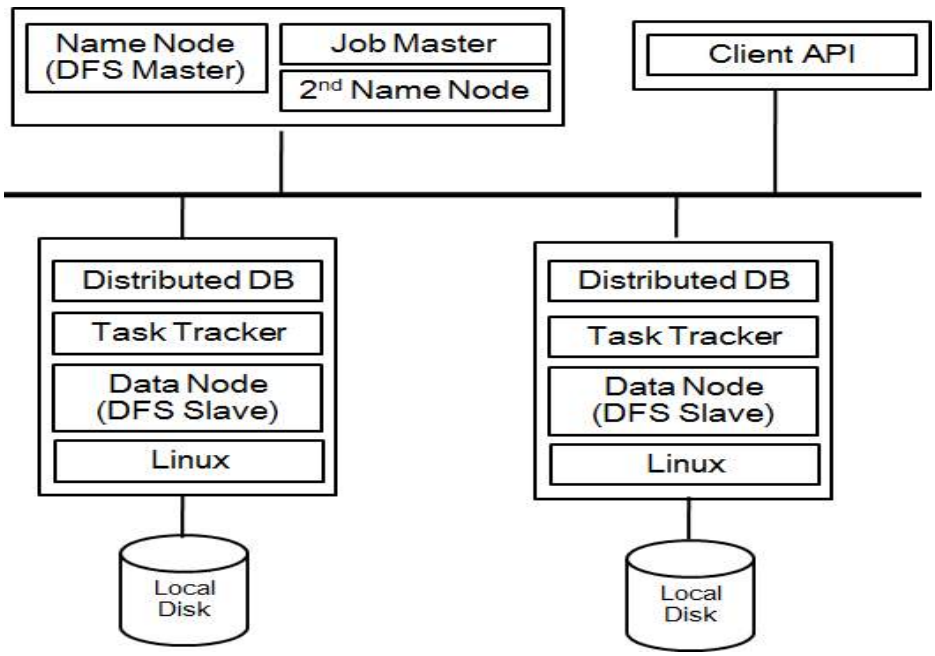
	Procedure k-Nearest Neighbor MapReduce
Map task	Input: All points Query point p Output: k nearest neighbors (local) Emit the k closest points to p
Reduce task	Input: Key = null Values = local neighbors Query point p Output: k nearest neighbors (global) Emit the k closest points to p among all local neighbors

2) 빅데이터 시스템을 이용한 분산처리과정

닐슨(Nielsen)의 3년 간 (2011년~2013년) 개인 시청률 단위 원자료(27GB)를 빅데이터 처리 시스템의 Hadoop 분산 파일시스템에 분산 저장하였다.

이 연구에서 사용된 빅데이터 처리 시스템은 4개의 노드로 구성되어 있으며, 1개의 네임 노드(name node)와 3개의 데이터 노드(data node)들로 구성된다. 또한 1개의 데이터 노드는 secondary name node 역할을 수행하며, 기존의 네임 노드에 오류가 발생 시, 대체되어 사용된다. 각 노드는 2.2GHz Hexa core 64bit 워크스테이션 서버이며, 24GB 메모리와 1TB 용량의 하드디스크로 구성된다. 빅데이터 시스템은 Cent OS 계열의 Linux가 운영체제로 설치되어 있고, Hadoop 1.2가 설치되어, 분산 컴퓨팅을 지원하고 있다(Alex Holmes, 2012). 또한 본 연구에서 제안한 시청률 예측 알고리즘을 병렬로 처리하기 위해 MapReduce 프레임워크가 설치되어 있다. 원자료 데이터 전처리를 위해서 Hive 언어를 사용하였고, k-NN 클러스터링 알고리즘은 Java 언어를 사용하여 MapReduce 프로그램을 작성하였다.

<그림 1>은 빅데이터 기술을 이용한 시청률 분석방법 연구에 대한 빅데이터 처리 시스템 구성도를 나타낸다.



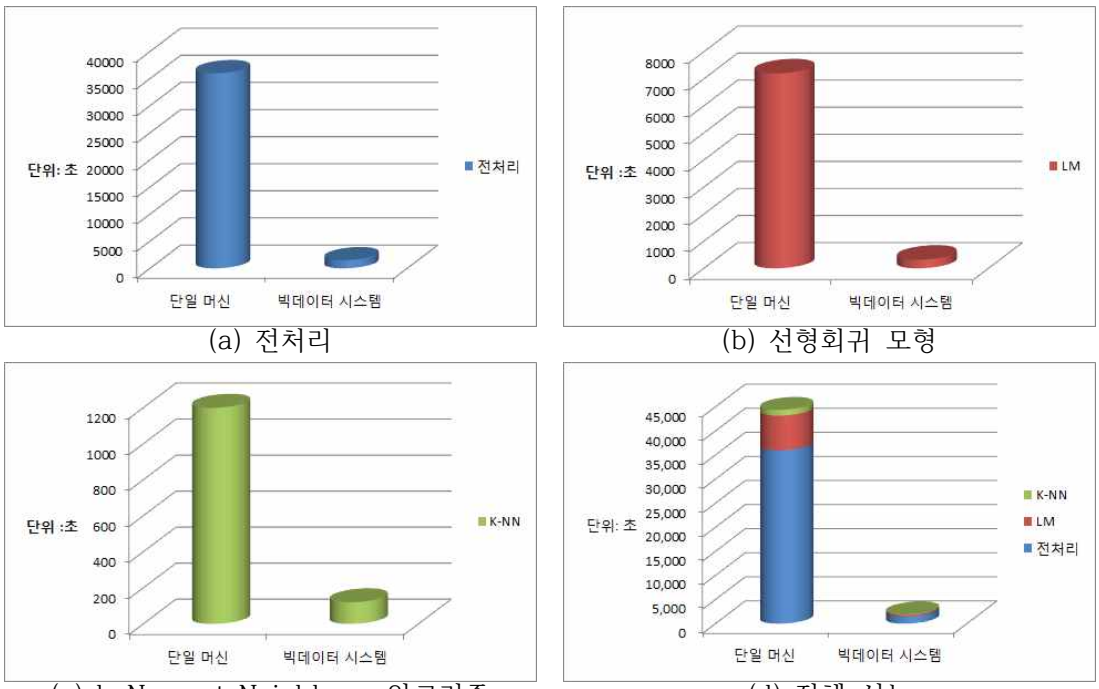
<그림 1> Hadoop 기반의 빅데이터 처리 시스템

네임 노드는 파일을 분할하여 데이터 노드에 전송한다. 이때 부하가 크지 않은 데이터 노드를 선정하여 chunk(분할된 파일)를 전송하며, 네임 노드는 chunk들이 어떤 데이터 노드에 저장되어 있는지에 대한 메타 정보를 저장한다. 데이터 노드는 약 200MB 크기의 chunk들을 대략 수백GB 씩 저장하고 있으며, Map과 Reduce 태스크들을 수행한다. 빠른 검색을 위해서 chunk의 수가 많아지면, chunk들을 병합하고 chunk의 사이즈가 크면, 성능 개선을 위해 chunk들을 분할한다.

3). 빅데이터 처리 결과

빅데이터 처리 시스템의 데이터 노드와 동일한 성능을 가지는 단일 머신에서 R 언어로 작성한 개인단위 시청률 예측 프로그램을 실행하고 성능을 측정하였다. 또한 빅데이터 처리 시스템에서 개인단위 시청률 예측을 위한 MapReduce 프로그램을 실행하고 성능을 측정하였다. <그림 2>는 단일 머신(single machine)과 빅데이터 처리 시스템의 성능을 비교한 그래프이다. dl 연구에서 제안한 방안이 시청률 예측 단일 프로그램보다 훨씬 빠른 성능을 보

였다. 시청률 데이터 전처리 단계에서는 제안방안이 23배 빠른 성능을 보였고, 선형회귀 모형(타 채널의 예견된 시청률 예측)에서는 22배, k-Nearest Neighbor 알고리즘에서는 9배, 그리고 전체 성능에서는 22배 빠른 성능을 보였다.



(a) 전처리 (b) 선형회귀 모형

(c) k-Nearest Neighbour 알고리즘 (d) 전체 성능

<그림 2> 단일 머신과 빅데이터 기술을 이용한 데이터 처리 시간 비교

5. 개인 시청률 기반 시청률 예측 모형: 정량적 접근

앞에서도 논의한 것처럼 여기서는 시청률 예측 장르 중에서 현실적으로 가장 관심정도가 큰 신작 드라마의 1회분 시청률을 예측하는 것으로 한정하였다.

예측변수로는 계절적 요인(예측대상 프로그램이 몇 월(月)에 방영되었는가를 변수화 함), 요일(7개의 요일을 변수화함), 방영시간대(24시간 중 어떤 시간대, 이를테면 22-23시에 방영되었다면 '밤 10시대' 등으로 변수화 함) 등이 있고, 경쟁채널 시청환경 요인을 고려하였다. 경쟁채널의 시청환경 요인은 경쟁중인 다른 3개의 지상파 채널의 시청률, 즉 KBS2의 프로그램을 예측할 경우 KBS1, MBC, SBS의 경쟁프로그램의 예견된 시청률을, MBC의 프로그램을 예측하는 경우 KBS1, KBS2, SBS의 경쟁프로그램의 예견된 시청률을, SBS의 프로그램을 예측할 경우 KBS1, KBS2, MBC의 예견된 동시간대 시청률을 경쟁요인으로 모

델에 투입하였다(예견된 경쟁프로그램의 시청률에 대해서는 백영민, 2012 참조).

예측모형에 투입되는 예측변수가 모두 개별시청자 각각의 개체 내(within-subject) 변화이기 때문에 프로그램 수준의 기존 예측모형과는 달리 인구통계학적 요인이나 시장상황과 같은 개체 간(between-subject) 요인은 투입되지 않았다. 개체 내 변화에 초점을 맞추었기 때문에, 이분산성(heteroskedascity)의 문제는 상대적으로 작았다고 볼 수 있다. 이 때 사용한 개별 시청자별 예측모형의 추정 알고리즘은 kNN (k-th nearest neighbor) algorithm (based on Euclidean distance)으로 추정하였으며, 예측대상이 되는 사례와 가장 유사한 예측변수 조합을 보이는 기존 사례 중 가장 가까운 3개의 사례(k = 3)를 채택하였다.

<표 4> 예측모형의 드라마 시청률 예측 평가

단위: 분수(minutes)

	방영일	실제시청	K=1	K=2	K=3	평균 (K1,K2,K3)
KBS2 채널						
칼과 꽃-1회†	2013/07/03	6.86	14.73	14.29	13.86	14.29
			(2.15)	(2.08)	(2.02)	(2.08)
아이리스2-1회†	2013/02/13	16.87	12.68	12.20	11.65	12.18
			(0.75)	(0.72)	(0.69)	(0.72)
MBC 채널						
여왕의교실-1회	2013/06/12	7.76	9.38	8.28	7.59	8.42
			(1.21)	(1.07)	(0.98)	(1.09)
남자가사랑할때-1회†	2013/04/03	8.12	12.92	11.99	11.21	12.04
			(1.59)	(1.48)	(1.38)	(1.48)
SBS 채널						
황금의제국-1회	2013/07/01	8.67	11.73	10.37	9.52	10.54
			(1.35)	(1.20)	(1.10)	(1.22)
야왕-1회	2013/01/14	7.42	7.66	5.93	5.25	6.29
			(1.03)	(0.80)	(0.71)	(0.85)

알림: 괄호안의 수치는 예측시청분수를 실제시청분수로 나누어 준 값으로 1일 경우는 정확한 예측이 이루어졌으나 1 미만일 경우는 예측값이 실제 값을 과소 추정한 경우, 1 초과 값인 경우는 예측 값이 실제 값을 과다추정한 경우를 의미. † 표시된 드라마의 경우 예측력이 높지 않다고 판단됨

<표 4>는 개별 시청자별로 예측모형을 구성한 후 드라마 초회분의 시청률을 예측한 결과이다. 선택된 6개의 드라마 KBS2의 '칼과 꽃' '아이리스2' 그리고 MBC의 '남자가 사랑할 때'의 경우 예측 값과 실제 값의 차이가 상당히 높았다. '칼과 꽃'의 경우 실제 값 보다 예측값이 2배 이상이었으며, '남자가 사랑할 때'의 경우 약 1.5배 이상의 과다추정 경향을 보

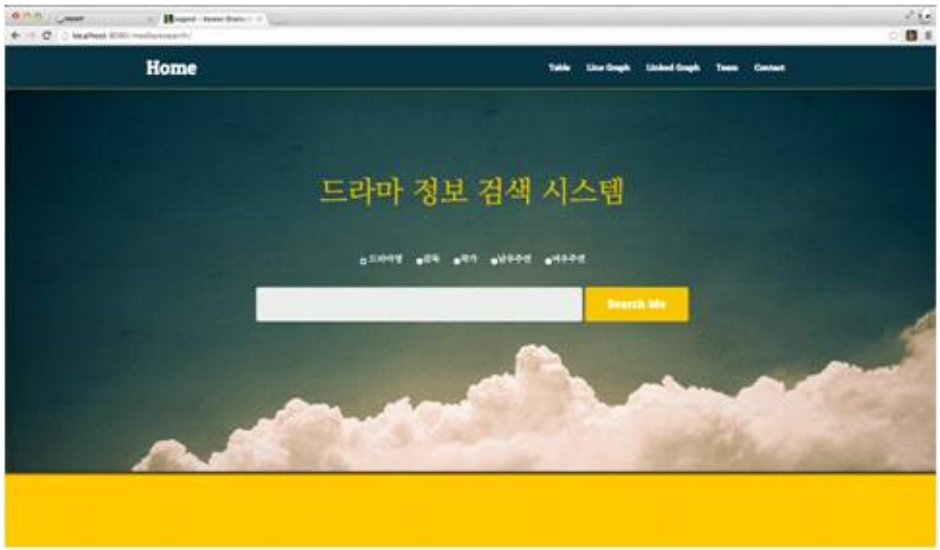
였다. 반면 '아이리스2'의 경우 과소추정 경향을 보였다.
반면 나머지 드라마 3 편은 약간의 과도추정 혹은 과소추정 경향을 확인할 수 있었지만, 전반적으로 나쁘지 않은 예측력을 보였다.

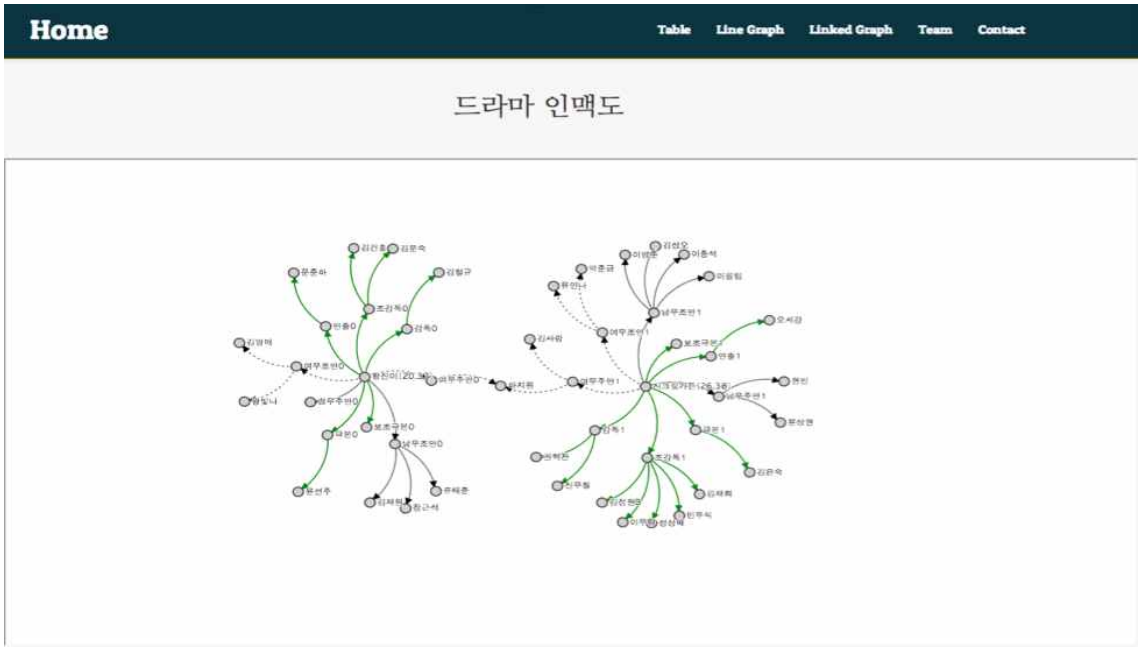
6. 드라마 시청률 예측을 위한 정성적 Linked Data 구축

앞에서 논의한대로 시청률을 예측하는 모형에서 프로그램 자체가 가지고 있는 정성적 속성을 배제한 시청률 예측연구는 한계가 있다. 따라서 배진아(2005)의 연구에서 시도하는 것처럼 드라마라면 그 속성 자체를 고려할 필요가 있다. 제한적이지만 배진아 연구에서는 스타라는 정성적 차원의 속성을 정량적으로 지수화하여 그 영향력을 논의했다.
따라서 기존 연구들에서 제한점 지적되어 온 프로그램 관련 속성들의 정성적 변수들을 보다 안정적으로 정량화 하여 프로그램의 시청률을 예측하는 연구는 매우 중요하다. 특히 시청률 예측 연구가 앞으로 닥칠 방송환경, 광고시장 변화에 부응하기 위해 기존연구 경향에서 벗어나 새로운 연구지평을 열어야 할 시점에 이러한 정성적 변인을 투입해 재조정된 시청률 예측모형을 만드는 것은 아주 중요하다.
이러한 정성변수 후보군으로는 예를 들어 드라마의 경우 연출, 조연출, 작가, 주요 출연자, 보조 출연자, 프로그램 세부 장르(현대극, 사극, 단막극, 일일극, 아침 드라마 등), 주요 시놉시스, 계절 요인, 제작비 규모 등이 있을 수 있다. 그 밖에도 스토리의 구성 차원(플롯의 복잡성 및 인물의 관계도 등), 외주제작 여부나 기타 작가의 영향력, 연출가의 질적 가치, 프로그램 방송사에 대한 채널 신뢰도 등이 다양한 차원의 정성적 변수 후보가 될 수 있다.
정성변수 투입 시 고려할 사항은 가치 내재적 성격의 정성변수를 계량화 하는 것이다. 변수의 속성이 명목적 척도(nominal scale) 특성을 가지고 있을 때는 이항분류(binary coding)를 하면 되는데 이 경우에도 이런 변인이 많을 경우 추정 변수의 수가 늘어나 계량모델 추정 시 연산처리에 큰 부담이 된다. 예를 들어 주요 출연자만을 코딩한다고 해도 5년 넘는 조사기간 중에 방영된 모든 드라마에 나온 탤런트의 수가 수백 명일 것이고 이들의 출연여부에 대한 더미(dummy) 변수도 따라서 수백 개가 된다.
특정 정성변수에 대해 서열적 코딩을 할 때가 더 큰 문제이다. 변수의 서열적 크기 판단에 대한 객관적 신뢰성과 타당도를 유지하기가 쉽지 않다. 프로그램의 질(quality)이 시청률을 결정하는데 중요 변인중 하나라는 연구결과(김은미, 박소라, 김예란, 2008)에 따라 이 변인을 정성변수로 투입할 경우 어떤 기준에 의해 분류할지가 문제가 된다. 그리고 이러한 크

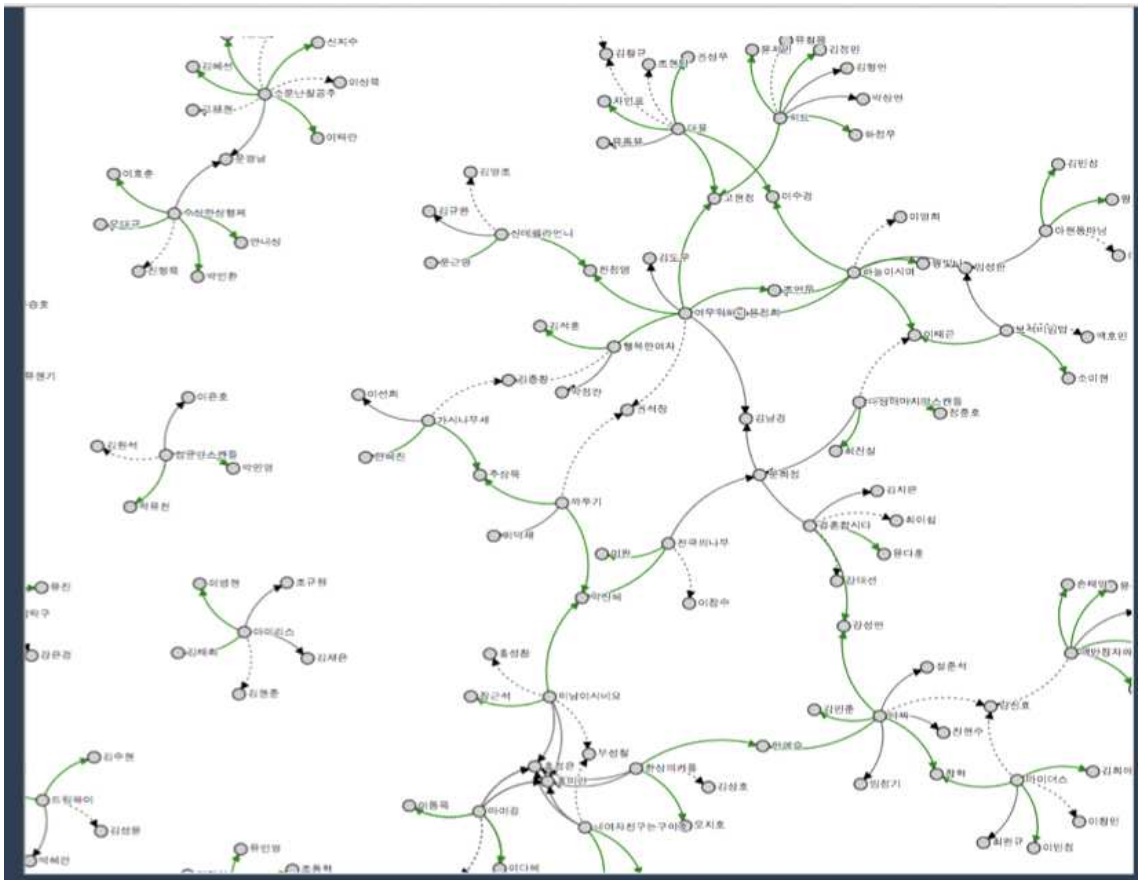
기의 유목(category)을 몇 단계로 할지도 문제가 되는데, 예를 들어 일반적으로 쓰이는 5점 척도로 만들 경우 의사결정나무 모형을 사용하면 {(2**k-1) - 1}개만큼의 평가대상이 추가되게 되어 계산과정에 부담이 커진다. 따라서 정성변수 투입 시 이러한 점을 고려해 변수의 수를 적절하게 조절할 필요가 있다.

따라서 여기서는 정성변수의 수를 연출자, 작가, 주요배우(주연 및 비중 높음 조연)등으로 한정했고 각 정성적 요인들의 관련성을 네트워크(Network)하여 클러스터링(Clustering)을 구축하여 평가하였다. 정성변수 Data는 3년간 방영되었던 드라마 368편의 관계 데이터 수 32,384개를 SQL 데이터로 구성하였고 구체적 기술요인(descriptors)은 드라마 명, 감독, 연출, 남/여주연, 남/여조연으로 이루어졌다. 구축한 프로그램에서는 일반 데이터 테이블, 시청률 그래프, 인맥 관계도 네트워크 등을 검색 할 수 있다. 아래 그림은 이에 대한 예제이다.





[주연배우 하지원 인맥 네트워크]



[시청률 15%이상 인맥 네트워크 클러스터링]

정성변수 Data 중 정확한 정보가 없거나 문제가 있는 프로그램들은 제거하는 과정을 거쳐

최종 정리대상 프로그램과 제작요소 인물들을 확정하였다. 제작요소 정리기준 인물들의 경우도 대표적인 주연인 1순위와 2순위, 준주연이라고 볼 수 있는 3순위와 4순위가 프로그램에 기여하는 대상이라는 점으로 고려해 4순위까지만 정리하였다. 그리고 각 인물들의 과거 평균 시청률과 참여율, 인물 간의 과거 친밀도를 알고리즘 항목으로 정의해 수치화 하였다. 예를 들어, 참여 참가율의 경우 연출가는 과거 제작한 드라마의 참가 가중치를 주연출 1.2, 보조연출 0.8로 하여 제작 편수에 따른 가중치 평균을 구하였고, 주연배우의 경우 과거 출연한 작품의 참가 가중치를 주연배우 1, 조연배우 0.5로 하여 출연한 드라마의 평균 시청률에 곱하여 가중치를 계산하였다. 인맥관계 네트워크를 구성할 때는 참여 인물 간 친밀도를 연출가를 중심으로 과거 같이 제작한 드라마의 과거 시청률을 기준으로 판단하였다. 단, 인물 간 처음 제작한 경우는 1로 하였다. 인물 간 친밀도 가중치 상수는 아래와 같다.

과거 시청률	친밀도
10% 미만	0.8
10% 이상 15%미만	1
15%이상 20%미만	1.2
20%이상 25%미만	1.5
25%이상 30%미만	1.8
30%이상	2

예를 들어 일일 드라마 지성이면 감천의 연출자, 작가, 출연 주요출연자들의 가중치는 다음과 같다.

	연출	작가	주요인물			
인 물	김명욱	김현희	박세영	유건	이해인	박재경
시청률	24.38	16.08	0	9.81	0	30.35
참여률	1.2	1	1	0.5	0.5	1
친밀도	1	1	1	1	1	2

설명하면 연출 김명욱의 과거 시청률 평균은 24.38%로 주연출로 3개의 드라마를 제작하였

다. 참여율은 3개 작품 모두 주연출로 1.2x3의 평균 1.2이고, 배우 박세영, 이해인은 처음 작품이므로 알고리즘에서 제외하였다. 유건은 과거 작품 1편에 대한 시청률 9.81%로 조연 배우로 참여하여 참여율 0.5이다. 배우 박재정의 경우, 연출 김명욱과 과거 제작한 프로그램에 출연을 하였고 그 프로그램의 시청률 30.35%이었으며 친밀도는 2가 된다.
이렇게 구축된 정성적 데이터는 정량적 데이터를 사용한 드라마 시청률 예측 모형에 대한 보완 정보로 유용하게 사용할 수 있을 것이다. 모든 미래 예측이 그러하듯이 정량 모델에만 의존하는 예측은 부정확 할 수 있다. 특히 프로그램 제작에 있어 인적 요소가 중요한 드라마의 시청률 예측은 체계적 정성정보 없이는 정확한 예측에 문제가 있을 수 있다. 예를 들어 kNN 알고리즘을 사용해서 계산된 시청률 예측치 중에서(<표 4>의 k=1, k=2, k=3 중)에서 어떤 것을 선택할까에 중요한 판단 잣대가 되고, 또 정량적으로 예측된 시청률을 정성 정보에 의해 인위적으로 가감할 수도 있고 이 때 이러한 결정의 근거가 된다[2].

7. 연구의 제한점 및 의의

개별시청자별 예측모형의 가장 큰 장점은 예측시청률 계산의 유연성이 매우 높다는 점이다. 개별시청자별로 예측 값을 계산한다는 점에서 무엇보다 타겟 집단의 예측시청률을 매우 세분화된 형태로 계산할 수 있다. 기존 시청률 모형에서는 프로그램별로 시청률을 예측하기 때문에 타겟 집단이 바뀔 때마다 예측모형을 별도로 설정하고 프로그램 예측 값을 설정하게 된다. 그러나 개별시청자별 예측모형은 개개인 샘플의 프로그램 시청률을 예측하기 때문에, 기존과 같은 가구시청률과 개인 시청률은 물론 타겟 집단의 예측시청률을 손쉽게 도출할 수 있는 장점이 있다.
두 번째의 장점은 중복률 계산이 가능하다는 것이다. 이 글에서는 보고되지 않았으나 두 개의 프로그램에 대한 예측시청률을 계산할 경우, 중복률 또한 예측가능하다. 기존 프로그램 단위의 예측모형으로는 프로그램의 예상 시청률을 계산할 수는 있지만 채널 내(혹은 채널 간) 중복률의 예측은 불가능하다.

개별시청자 데이터를 이용한 정량적 예측모형에서 개선될 사항은 다음과 같다. 첫째, 개별시청자의 시청을 예측할 경우 현재의 모형에서는 계절적 요인과 예견된 프로그램 경쟁상황만을 이용하였다. 예측대상이 되는 개별시청자가 어떤 특성의 프로그램을 보았는지, 그리고

2) 정성적 Linked data만 가지고 드라마 시청률을 예측하는 방안에 대한 시뮬레이션 연구가 현재 진행 중임.

또 예측 대상이 되는 프로그램이 어떤 특성을 갖는지에 대한 변수화가 가능하다면 프로그램 예측력은 더 높아질 수 있다. 일례로 앞에서 살펴본 총 6개의 드라마의 경우도 드라마의 내용적 특성을 반영할 수 있는 변수가 있었다면 예측력이 더 높아 졌을 것이다.

둘째, 분산컴퓨팅 환경에서 예측이 더 효율적이고 빨라졌다는 점에서 kNN이 아닌 다른 기계학습 알고리즘을 이용하는 것도 적극 고려할만 하다. 상대적으로 kNN 알고리즘은 속도가 빠르기는 하지만, 정확도가 느린 단점이 있다. 이번 연구에서는 예측모형의 속도라는 점에서 일반적 PC와 분산컴퓨팅을 비교하기 위해 일반적 PC에서도 상대적으로 속도가 빠른 kNN을 사용하였지만, 차후 연구에서는 다른 대안적 기계학습 알고리즘도 적극 고려할 수 있다.

셋째, 이번 연구에서는 시청흐름(audience flow)에 대해 고려하지 못하고 있다. 시청흐름의 효과가 다채널 상황에서 점점 줄고 있다고는 하지만 여전히 프로그램 편성의 주요원칙으로 존재하며 그 효과를 인정받고 있다. 차후 모형에서는 이를 변수 화할 필요가 있다.

넷째, 이 연구의 예측모형은 가구내의 다른 구성원들의 시청이 특정 개별시청자의 시청활동에 어떠한 영향을 주는가에 대해 적극적으로 고려하지 못하였다. 다채널환경에서 가정내 시청환경의 효과가 높다고 보기는 어렵지만, 가구구성원의 수가 많은 경우 개별시청자의 시청활동에 영향을 미치는 중요한 예측변수 중 하나라고 예측할 수 있다.

다섯째, 표본의 대표성을 보장하기 위해 사후 유층화(post-stratification) 가중치를 사용하여 예측력을 살펴보아야 한다. 현재는 개별시청자가 어떤 프로그램을 몇 분 정도나 시청했는가에 대해서만 살펴보았으나, 이는 산업계에서 유통되고 있는 방식의 시청률 자료와는 거리가 멀다. 향후의 개선모형은 가중치를 적용 보다 실재에 맞는 방식의 시청률 계산을 시도해야 할 것이다.

정성변수 데이터를 이용한 Linked Data를 사용하면 인적 요소가 중요한 영향을 끼치는 드라마 제작에서 안정적인 시청률 예측 값을 제시할 수 있다. 과거 작품 활동의 참가한 비중에 따른 고려와 제작진과 배우들의 작품에 대한 영향도 및 친밀도를 계산하여 시각적으로 나타나는 연결 그래프를 제안하였다. 이는 인물간의 시너지(Synergy)가 시청률에 영향을 끼치고 있다는 점을 고려한 것이다. 그러나 이러한 Linked Data는 과거 기록이 많지 않은 경우, 그 활용도가 낮아진다. 그리고 과거 기록보다 소위 말하는 전혀 예상치 못한 연출자, 출연자의 조합으로 만들어 내는 소위 말하는 '대박 드라마'의 경우 이러한 정성적 정보의 유용성이 제한된다.

여기서 사용한 정성적 데이터를 개선할 경우 고려할 수 있는 변인들은 다음과 같다.

첫째, 시나리오 주제에 대한 트랜드 분석이다. 기존의 데이터를 분석한 결과시나리오 주제

가 작품이 방영되는 시점의 사회의 관심 이슈일 경우, 시청률이 증가하는 현상이 나타났다. 둘째, 작품의 주요 시청 대상의 연령대에 대한 분석이다. 시청자의 연령에 따른 본 방송 시청률의 영향은 여러 연구에서 볼 수 있다. 이것은 연령대별 방송의 청취 스타일이 틀린 점이 시청률에 영향을 끼치기 때문이다. 이에 작품의 주제에 따른 연령대별 관심도에 대한 고려가 필요하다.

<부록 1>

데이터	속성	설명	예제	타입
가구	ID	가구번호	1401077	String
	File Date	날짜	20130701	String
	Area Zone	패널가구의 구별 정보	Jongro-gu	String
	Size	패널가구의 주택평수 정보	20~49 / From 100 Up	String
	Building Form	패널가구의 주택형태 정보	Detached house	String
	Ownership	패널가구의 저가/전세 정보	Own	String
	Income	패널가구의 가구소득 정보	120~199 / From 500 Up	String
	Member Number	패널가구의 가족 수 정보	1~2 / From 5 Up	String
	TV Number	패널가구의 TV대수 정보	1 / From 2 Up	String
	Child under 13	패널가구의 13세 이하 아이 유무	Exist	String
	Cable Form	패널가구의 케이블 상품 정보	Popular(pay)	String
	Sky Life HD	패널가구의 스카이라이프 HD 여부	Non-exist	String
	IPTV	패널가구의 IPTV 여부	Exist	String
	Cable TV	패널가구의 케이블 가입 여부	Narrowband Cable	String
	Satellite	패널가구의 위성수신 여부	Non-exist	String
	Sky Life Reception	패널가구의 스카이라이프 수신여부	Exist	String
	Sky Life Form	패널가구의 스카이라이프 상품 정보	Fundamental(pay)	String
	IP License	패널가구의 IPTV 사업자 정보	KT QOOK LIVE	String
	Market	해당 가구가 속해 있는 마켓	Seoul	String
	Weight	해당 가구의 가중치 정보	8385.8	Float
패널	ID	패널이 속한 가구번호	1401077	String
	File Date	날짜	20130701	String
	Name	구성원 구별자	aa	String
	Sex	패널의 성별 정보	male	String
	Age Scope	패널의 연령대 정보(5세 단위)	10~14 / From 65 Up	String

	Schooling	패널의 학력수준 정보	College graduate	String
	Occupation	패널의 직업 정보	Professional / Management	String
	Marriage	패널의 결혼 여부	Married	String
	Age	패널의 실제 나이	24 / From 70 Up	String
	Weight	패널의 가중치	10448.4	Float
시청	ID	시청정보가 속한 가구번호	1401077	String
	File Date	시청한 날짜	2013-07-01	String
	Channel	채널코드	SBS ESPN	String
	TV ID	TV 번호	1	Int
	Platform	해당하는 채널에 시청되는 플랫폼	Skylife	String
	Name	시청한 멤버	aa	String
	Time Start	시청 시작 시간	140800	String
	Time End	시청 종료 시간	182559	String
	Time Total	총 시청 시간	5 hour 12 min 59 sec	String
프로그램	Local Code	지역코드	whole country	String
	Channel	채널코드	MBC	String
	End Time	프로그램 종료시간	2010-07-01 20:17:46	Int
	Start Time	프로그램 시작시간	2010-07-01 20:53:11	Int
	Class Code	시급	C	String
	Program Description	프로그램명	일일연속극(황금물고기)	String
	Class Code	광고비	12630	String
	Type Top	대분류	Drama & Movie	String
	Type Middle	중분류	Drama	String
	Type Bottom	소분류	Daily soap opera	String
	Plan Time	프로그램 계획시간	20:15:00	Int

8. 참고문헌

강남준·김은미 (2010). 다중 미디어 이용의 측정과 개념화: 오디언스를 향한 새로운 시선. 『언론정보연구』, 47권 2호, 5~39.

박원기·김수영 (2003). 시청률 예측에 관한 연구: 회귀모형과 데이터마이닝모형의 예측력 비교를 중심으로. 『광고연구』, 봄호.

박노성·송석현 (2005). 시청률 예측에 관한 연구 (연구보고서 2005). 한국방송광고공사.

배진아 (2005). 드라마 시청률 영향 요인 분석: 드라마 속성 및 수용자 요인을 중심으로. 『한국방송학보』 19권 2호, 270~309.

이혜갑 (1994). 텔레비전 프로그램 시청률 예측 가능성 연구. 『방송광고연구총서 : 매체관련편』. 서울: 한국방송광고공사, 368-378.

정진욱 (2003). 한국의 TV 시청 유형 분석 및 시청자 유형별 시청률 예측 모형.『연세경제연구』, 10권 1호, 37~52.

Danaher, P. J., Dagger, T. S., & Smith, M. S. (2011). Forecasting television ratings. International Journal of Forecasting, 1-26.

Danaher, P., & Mawhinney, D. (2001). Optimizing television program schedules using choice modeling. Journal of Marketing, 38, 298-312

Gensch. D., & Shaman. P. (1980). Models of competitive television ratings. Journal of Marketing Research, 17, 307~315.

Holmes, A., (2012). Hadoop in Practice, Manning: NY

Meyer, D., & Hyndman, R. J. (2005). Rating forecasts for television program. Unpublished working paper in Mornash University.

Napoli. (2001). The Unpredicable Audience : An Exploratory Analysis of Forecasting Error for New Prime-Time Network Television Programs, Journal of Advertising, 30(2), 53-60.

Nikopoulos, K., Goodwin, P., Patelis, a., & Assimakpoulos, V. (2007). Forecasting with cue information: A comparison of multiple regression with alternative forecasting approaches. European Journal of Operational Research, 180, 354-368.

Rust, R. & M. Alpert. (1984). An audience flow model of television viewing choice. Journal of Marketing Research, 17(4), 6-13.

Rust, R. Kamakura, W., & M. Alpert. (1992). Viewer preference segmentation and viewing choice models for network television. Journal of Advertising, 21(1), 1-17

Tavakioli, M. & Cave, M. (1996). Modeling Television Viewing Patterns. Journal of Advertising(winter), 71-86.

Weber, R. (2000). Prognosemodelle zur vorhersage der fernsehnutzung. Neuronale Netze, Tree-modlle und klassische statistik im Vergleich. Reihe Medienskripten. 34. Munchen: Reinhard Fischer.

Yao, C. Y., & Kim, H. G. (2002). An analysis of prediction error for new prim-time television programmes: competitive study between USA and Korea. International Journal of Advertising, 21, 525-546.