大数据背景下江苏有线电视收视率预测

内容提要

关键词

2009   AC

–   CSM

2010   3

500

2014   7   1

–   –

3

4

2. 电视收视率分析中 ETL 的实现方式。ETL
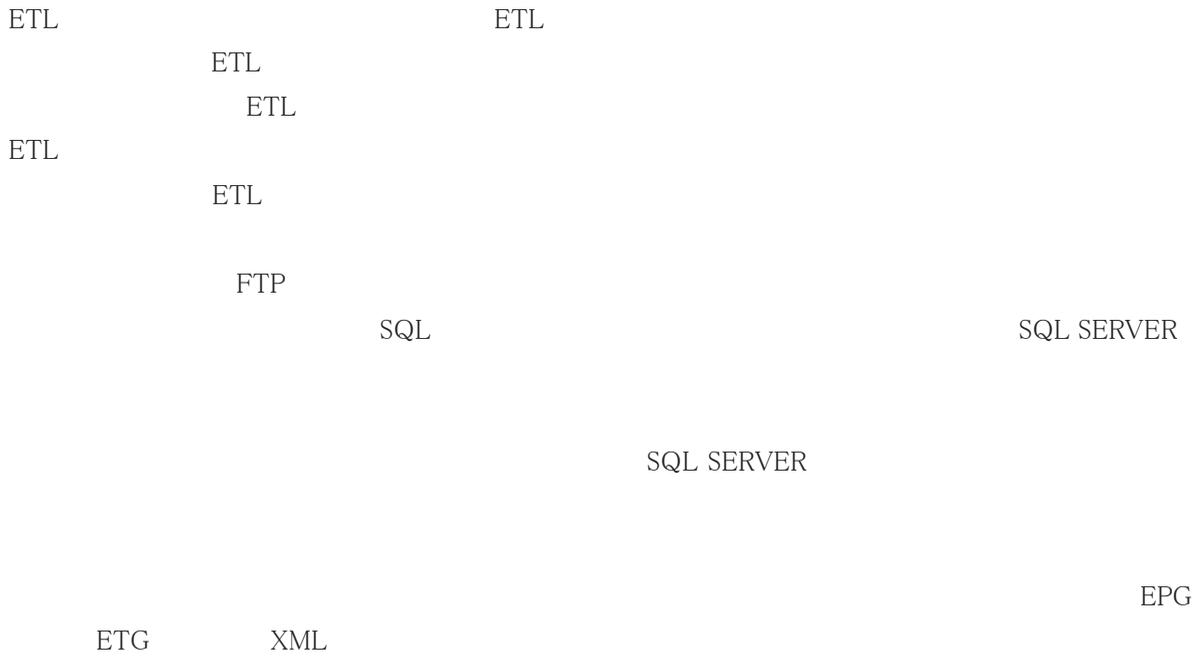
ETL ETL

ETL

ETL

ETL

ETL

FTP

SQL SQL SERVER

SQL SERVER

EPG

ETG XML

3. 电视收视率数据处理分析流程。

1

2 ①

②

EPG　　　　ETG

XML ③

3 ①

② ③

SQL ④

⑤ Excel

⑥ Word word

⑦

ETL 2013　3　9

2014　6　8 20:00　23:59

60　　　1

－　－
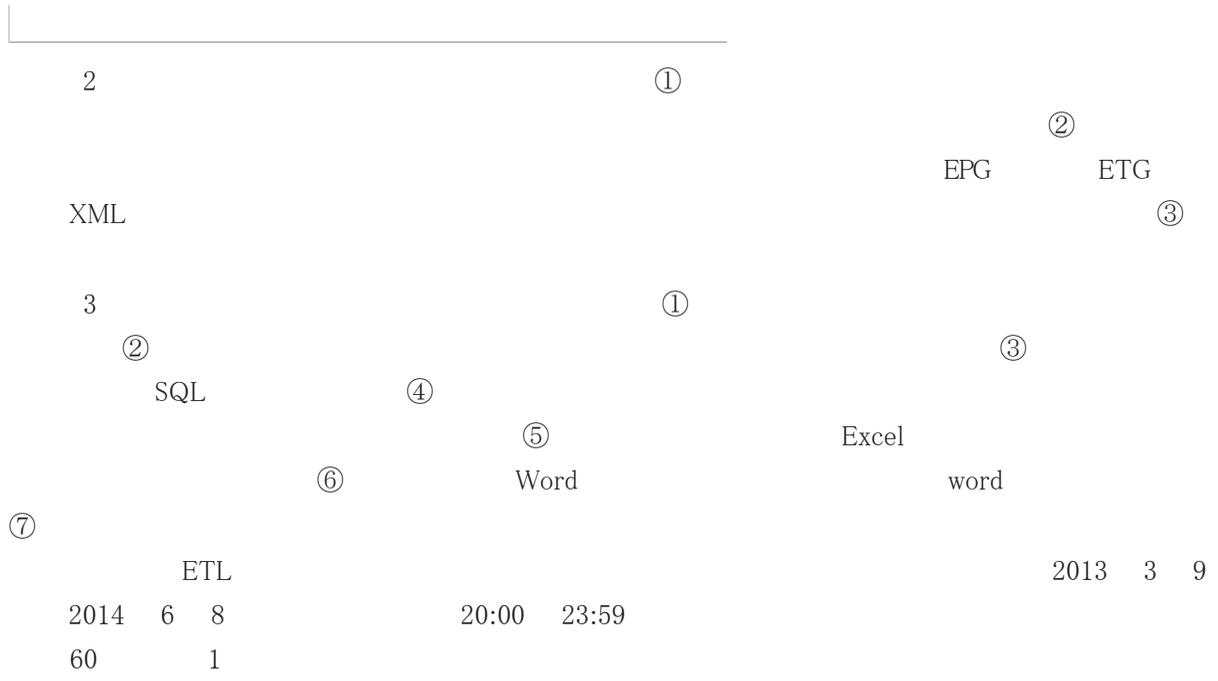
ARIMA　　　　　　　　　　　　　　ARMA　　　　　20

70　　　　　　George Box　　　　　　　Gwilym Jenkins　　　　　　　　　　　　　Box-Jenkins

B-J　　　　　　　　　　　　　　　　　　　　　　　　　　　ARMA

ARIMA

ARMA

1. ARIMA 模型

ARIMA(p,d,q)　　　　　　p　　　　　　AR(p)　　　　d　　　　　　　　　q　　　　　　MA　q

ARIMA(p,d,q)　　　ARMA(p,q)　　　　　　　3　　　　　AR(p)　MA(q)　ARMA(p,q)

AR(p)

$= + _1 _{-1}++ _2 _{-2}+...+ + $ MA(q)

$= + _1 _$

$_1+ _2 _{-2}+...+$ ARMA(p,q)

$= + _1 _{-1}++ _2 _{-2}+...+ + + _1 _{-1}+ _2 _{-2}+...+$

p,q

ARIMA

ARIMA　　　　　　　　ARIMA　　　　　　　　　　　p,d,q　　　　　　　　　　　　　d

d　　　　　　　　　　　　　　　d　p　q

ACF

PACF　　　　　　　　　　　1

表 1　ARIMA

	AR(p)模型	MA(q)模型	ARMA(p,q)模型
ACF	拖尾,指數衰减	Q 期後截尾	拖尾,指數衰减
PACF	P 期後截尾	拖尾,指數衰减	拖尾,指數衰减

AIC(Akaike Information Criterion)　SC(Schwarz Criterion)　　HQ(Hannan-Quinn Criterion)

AIC　SC　HQ

ARIMA

ARIMA

LM

Breusch　Godfrey　1978

ARIMA　　　　　　　　　　　　　　　　　　　ARIMA

2. ARIMA 模型的应用

1

2013　3　9　　　2014　6　8

－　　－

20:40 21:10 21:40 21:50 22:00 23:00 23:20 7

ARIMA   EViews 7.2   R 3.1.1   21:10

2

ARIMA(p,d,q)

①   ②

③

图2

2

3

ARMA

图3

k

0   k   2   3

3

15

0

4

0

图4

— —

图4 一阶差分后ADF检验
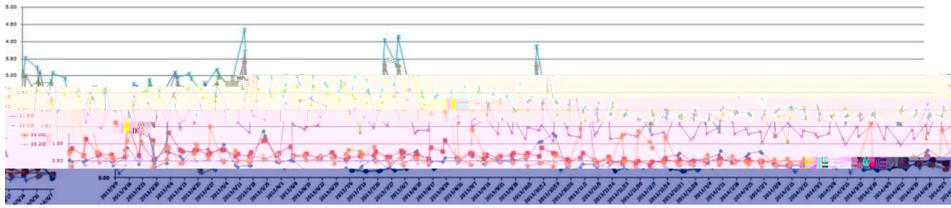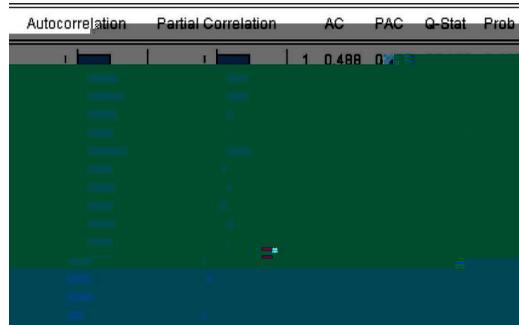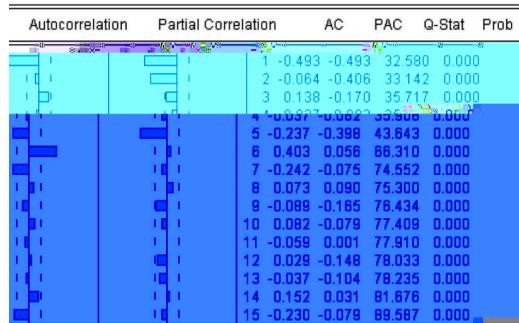
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-10.06764	0.0000
Test critical values:	1% level	-4.032498	
	5% level	-3.445877	
	10% level	-3.147878	

图5 ADF

图6 ARIMA(4,1,1)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.003583	0.000566	-6.331231	0.0000
AR(1)	0.117415	0.092506	1.269269	0.2068
AR(2)	0.045323	0.075712	0.598623	0.5505
AR(3)	0.182864	0.082994	2.203348	0.0295
AR(4)	0.001468	0.089607	0.016380	0.9870
MA(1)	-0.999578	0.037504	-26.65	

图6 ARIMA(4,1,1)

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	0.475492	Prob. F(2,119)	0.6228
Obs*R-squared	0.996541	Prob. Chi-Square(2)	0.6076

图7 ARIMA(4,1,1)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.0...			

图8 ARIMA(0,1,1)

F R2

图8 ARIMA(0,1,1)   c β1

图9 ARIMA(0,1,1)
F
R2

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	0.157151	Prob. F(2,127)	0.8547
Obs*R-squared	0.315001	Prob. Chi-Square(2)	0.8543

图9 ARIMA(0,1,1)

7   ARIMA
ARIMA(p,d,q)   ARIMA

表2   ARIMA

表2
ARIMA(3,1,7)   ARIMA(3,1,
5)   ARIMA(4,1,1)   ARIMA(0,1,1)

模型	Adjusted R^2	AIC	SC	残差序列相關性
.RIMA(3,1,7)	0.6089	−1.1341	−0.8890	相關
.RIMA(3,1,5)	0.5596	−1.0297	−0.8291	相關
.RIMA(4,1,1)	0.4378	−0.8321	−0.6977	不相關
.RIMA(0,1,1)	0.4403	−0.8180	−0.7742	不相關

R^2 0   1   AIC
SC

表3   ARIMA

表3   ARI-
MA(3,1,7) R^2

20:40	21:40	21:50	22:00	23:00	23:20
RIMA(1,0,1)	ARIMA(1,1,1)	ARIMA(0,1,2)	ARIMA(1,1,3)	ARIMA(1,0,0)	ARIMA(2,0,0)

AIC   SC   ARIMA(3,1,5)   ARIMA(4,1,1)   ARIMA(0,1,1)   ARIMA(3,1,7)
ARIMA(3,1,5)   ARIMA(4,1,1)   ARIMA(0,1,1)

R^2 AIC   SC   ARIMA(0,1,1)   Prob.   ARIMA(4,
1,1)   ARIMA(0,1,1)   =−0.003258+
−0.929927 _{−1}   IMA(1,1)

20:40   21:40   21:50   22:00   23:00   23:20   6   ARIMA   3

8 ARIMA

4

7   ARIMA

表4   ARIMA

	20:40	21:10	21:40	21:50	22:00	23:00	23:20
	ARIMA(1,0,1)	ARIMA(0,1,1)	ARIMA(1,1,1)	ARIMA(0,1,2)	ARIMA(1,1,3)	ARIMA(1,0,0)	ARIMA(2,0,0)
1	0.4616080	0.5100765	1.883294	1.322597	1.766510	0.6484898	0.4789755
2	0.5062416	0.5100765	1.861985	1.361489	1.880563	0.7048125	0.4850459
3	0.5223789	0.5100765	1.852263	1.361489	1.825471	0.7193314	0.5054597
4	0.5282133	0.5100765	1.847827	1.361489	1.878681	0.7230741	0.5103129

ARIMA   7   4

4

−   −

ETL

ARIMA

ARIMA

1.   2010   177-178

2.   Ruey S. Tsay   （ 3 ）   2012
55-70

3.   :   2012   48-93

4.   （ 5 ）   2012   786-793

5.   EViews   （ ）   2013   147-160

6.   ARMA   2009   5 S0
9-11

7.   30   2011   7
34-38

8.   RBF   2013   9   1428-1431

9.   2007   3   251-252

A Cable Television (CATV) Audience Rating Forecast
in Jiangsu Province in the Context of Big Data

Xing Yabin   Shi Ziguo

Abstract: Audience rating is a key indicator in broadcasting industry, but in recent years it has met with incredulity and attracted much attention due to the great deviation as a result of the effect of human factors on part of samples. The present research, based on the audience data collected through CATV set top boxes (STB) in an area of Jiangsu Province, got a full sample of audience rating data by using ETL data mining techniques. Selecting several representative time points, it applied ARIMA Model in time sequence analysis to data analysis, producing ARIMA models at each point of time and predicting the audience rating.

Keywords: full sample; Television audience rating; ETL instrument; ARIMA Model