商务智能
Business Intelligence

数据挖掘（Data Mining）

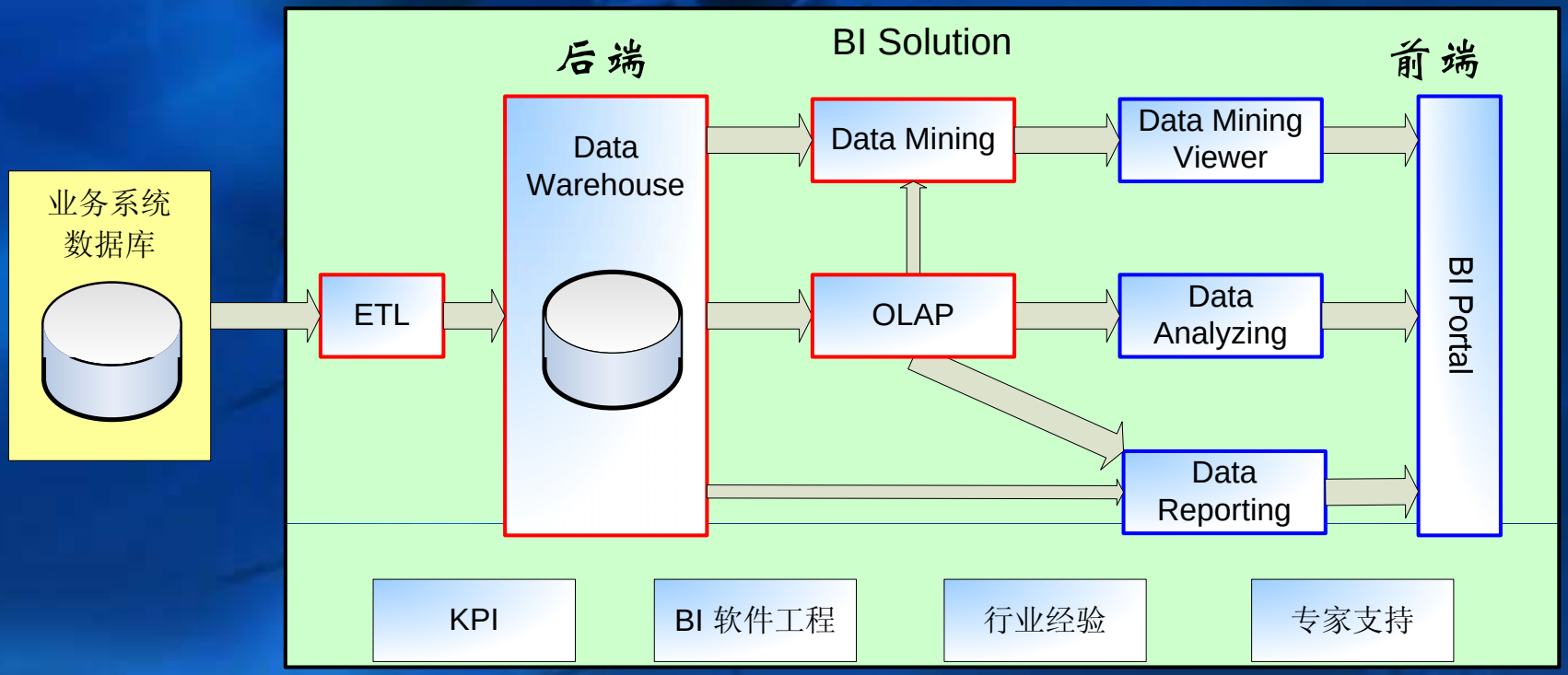
个人简介

- 王如涛
  - 高级**BI**咨询顾问和项目经理
  - 多次在微软**MSDN**和**TechNet**上讲授**BI**课程
  - 曾参与实施了包括大型搜索引擎在内的多个BI项目的实施，涉及的行业有互联网、医药、鞋服、烟草、零售等行业。
  - **MCP**
  - wrtandy@gmail.com

主要内容

- SQL Server 2005数据挖掘概览
- SQL Server 2005数据挖掘具体运用
- 其他工具的整合及其二次开发

BI 解决方案


BI Solution
后端
前端
业务系统
数据库
ETL
Data
Warehouse
Data Mining
Data Mining
Viewer
OLAP
Data
Analyzing
Data
Reporting
BI Portal
KPI
BI 软件工程
行业经验
专家支持

数据挖掘的基本知识

1、数据挖掘是怎样一个过程呢？

从海量数据中，提取隐含在其中的、人们事先不知道的但又可能有用的信息和知识的过程。

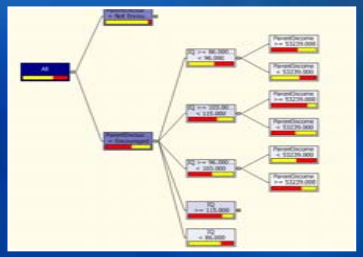
2、数据挖掘的数据源是什么呢？

数据仓库、数据库或其他数据源。

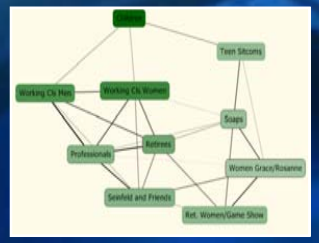
3、数据挖掘特性？

数据挖掘的反复特性。

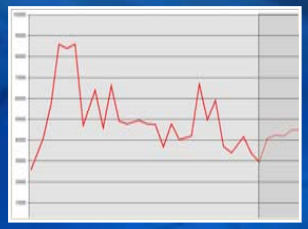
SQL Server 2005算法集合



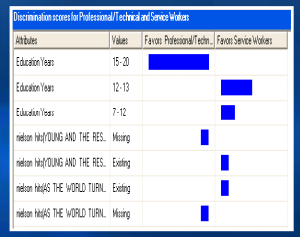
决策树



聚类



时间序列

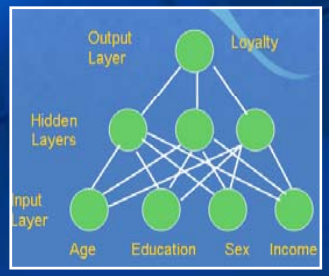


Naïve 贝叶斯

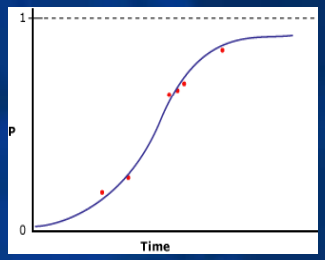
SQL Server 2000 中已提供了

序列聚类

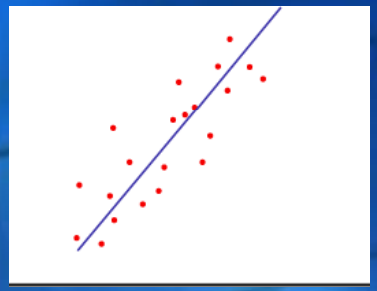
关联



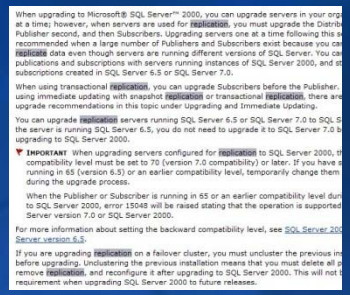
神经网络



逻辑回归



线性回归



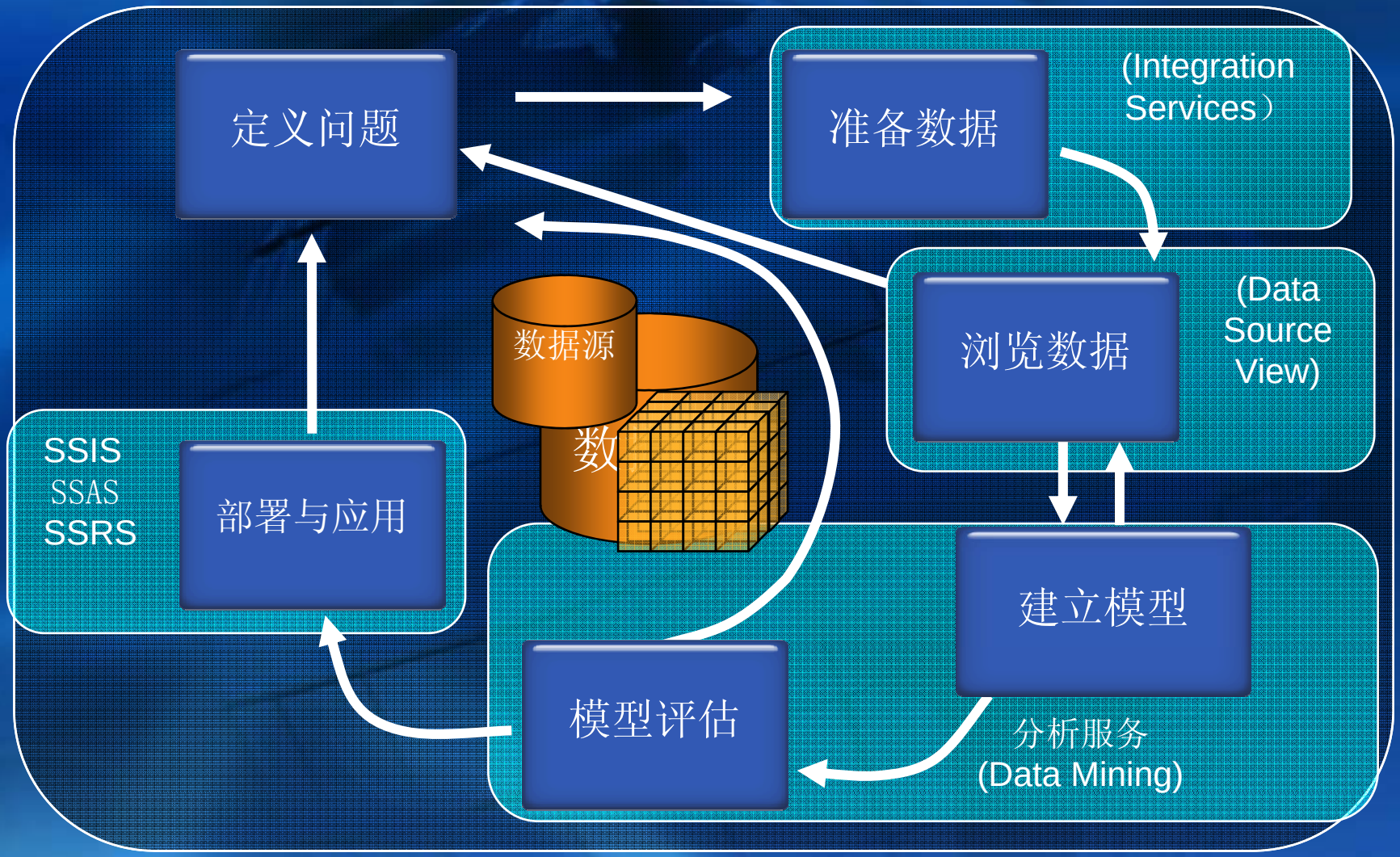
文本挖掘

多维数据分析和数据挖掘的区别

	OLAP	Data Mining
技术核心	维	算法
基本分析操作	钻取、切片和切块、以及旋转、Drill Through等	调整参数、算法优化、预测、Drill Through等
侧重点	侧重决策支持	侧重找到有价值的未知
研究人员	从事数据库的人员	从事数据库、自人工智能、统计工作的人员
过程	演绎推理	总结归纳

基于数据仓库的联机分析处理技术与数据挖掘技术的融合和互补，将是商务智能技术的发展方向。

SQL Server 2005数据挖掘处理流程



DMX简介

1、定义

DMX----Data Mining Extensions

是一种语言，数据挖掘语言

数据挖掘扩展插件，是对SQL的扩展

语言	全　　称	应用场景
SQL	结构化查询语言	关系型数据库
MDX	多维表达式	多维数据库
DMX	数据挖掘扩展插件	数据挖掘

DMX简介

2、功能

创建和处理数据挖掘模型。

创建新数据挖掘模型的结构

为挖掘模型定型

浏览、管理和预测

DMX简介

3、基本框架

数据定义语言 DDL

数据操作语言 DML

函数

运算符

语法元素

- 标示符
- 数据类型
- 运算符
- 内容类型
- 函数

语法元素

1、标示符

常规标示符

Unicode 标准 2.0 定义的字母,
下划线（_）
数字

分隔标示符:'[ ]'

- 保留关键字作为对象名或对象名的一部分时
- 不是限定标识符的字符时
- 分隔标示符容量:分隔标识符可以包含与常规标识符相同的字符数（1 到 100 个，不包括分隔符本身）。

语法元素

2、数据类型

Text：　例如：姓名：张三

Long：　例如：年龄：23

Date：　例如：日期：2006:10:11

Boolean：例如：true/false

Double：　例如：单价：2.15

语法元素

3、运算符

- 算术运算符 +， -， *， /
- 比较运算符 >, <, >=, <=, <>, =
- 逻辑运算符 AND, OR, NOT
- 一元运算符 +，-
- 注释符号 //，--，/*…*/

定义功能

定义功能一：Create structure

```
create mining structure MovieSurvey
(
    SurveyTakenID text key,
    Movies table
    (
        SurveyTakenID text discrete,
        Movie text key,
        MoviePre text discrete

    )
)
```

定义功能

定义功能二：Create model

```
create mining model Movie
(
SurveyTakenID text key,
Movies table
    (
        SurveyTakenID text discrete,
        Movie text key,
        MoviePre text discrete predict
    )
)using
    Microsoft_association_rules(Minimum_support=20,minimum_pr
    obability=0.05)with drillthrough;
```

自动创建Movie_Structure

定义功能

定义功能三：alter structure add model

```
alter mining structure MovieSurvey
add mining model [MovieSurvey]
(
    SurveyTakenID,
    Movies
    (
        SurveyTakenID,
        Movie,
        MoviePre predict
    )
)usingMicrosoft_association_rules(Minimum_support=20,minimum
    _probability=0.05)with drillthrough;
```

定义功能

定义功能四：select … into …

```
select * into Movie
using Microsoft_Association_rules
(
    Minimum_Support=20,
    Minimum_Probability=0.005
) with drillthrough
from [MovieSurvey];
```

定义功能

其他定义功能：drop、import、export…

```
drop mining model movie
drop mining structure [movie_structure];
Export mining structure MovieSurvey model MovieSurvey to
    'E:\\MovieSurvey.abf'
export mining model [MovieSurvey ] to 'E:\\ MovieSurvey.abf ' with
    dependencies;
import from 'E:\\ MovieSurvey.abf';
```

主要内容

- SQL Server 2005数据挖掘概览
- SQL Server 2005数据挖掘具体运用
- 其他工具的整合及其二次开发

互联网与数据挖掘

- 搜索关键字之间的关联性
- 网站栏目之间的关联性
- 网站的点击流序列
- 访客是否会访问某个栏目
- 网上商店的关联销售

……

使用的算法主要是

关联规则算法

顺序分析和聚类分析算法

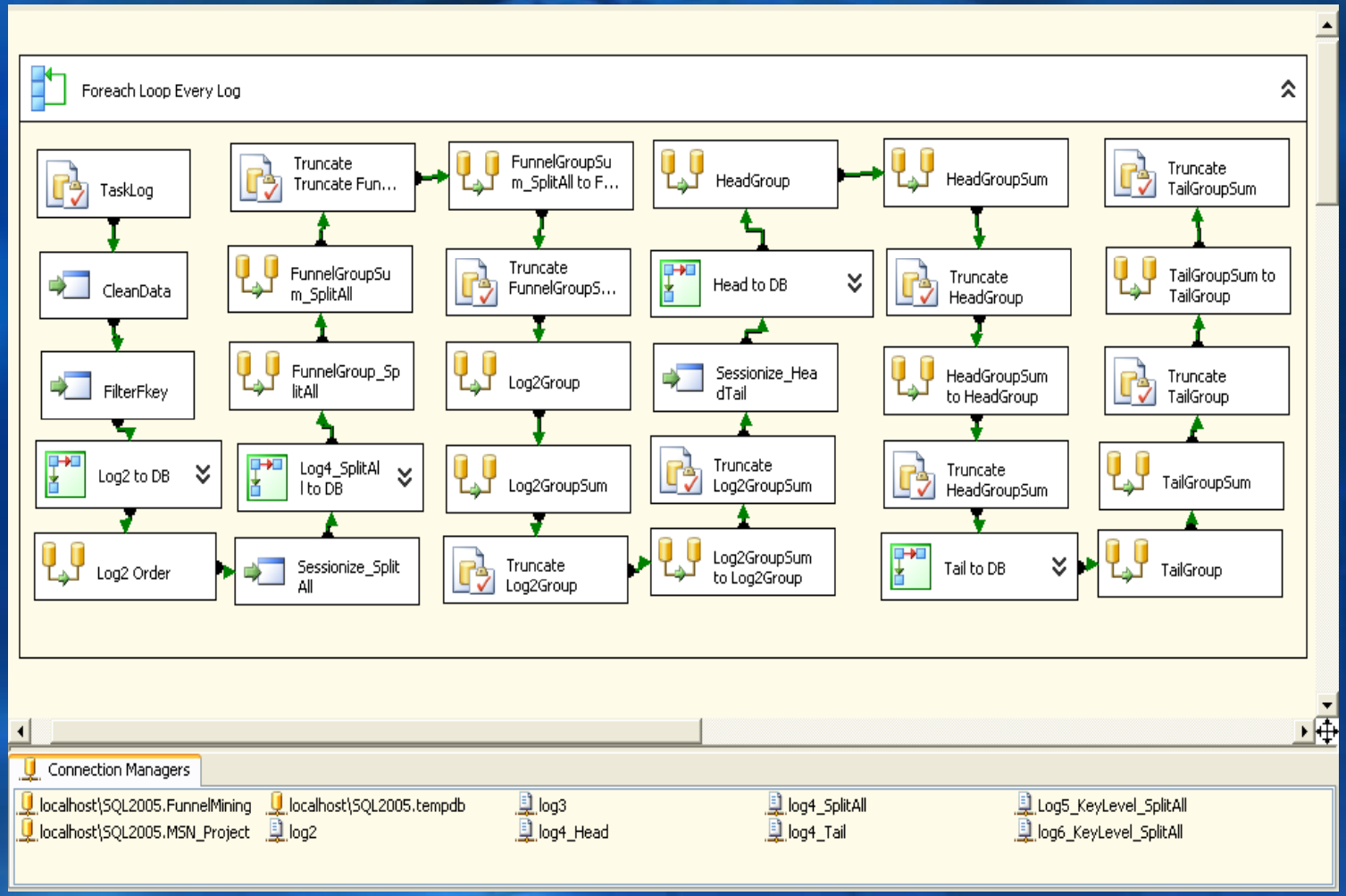
决策树算法

数据源

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
162.105.146.11 - - [28/Nov/2005:04:02:10 +0800] "GET http://it.520vc.com/news1/cssd/index_35.html HTTP/1.1" 200 61263 "-" "P.
162.105.146.11 - - [28/Nov/2005:04:02:19 +0800] "GET http://www.it.com.cn/f/diy/0510/24/189826.htm HTTP/1.1" 200 63598 "-" "P
162.105.146.49 - - [28/Nov/2005:04:02:26 +0800] "GET http://pic.it.com.cn/f/desktop/0412/21/041215_dt_app_wz_10.jpg HTTP/1.1"
162.105.146.49 - - [28/Nov/2005:04:02:26 +0800] "GET http://it.com.cn/f/market/0410/12/34095b.jpg HTTP/1.1" 200 29175 "-" "Mo
162.105.146.49 - - [28/Nov/2005:04:02:26 +0800] "GET http://www1.it.com.cn/f/games/053/10/0310ship_game10.jpg HTTP/1.1" 200 3
162.105.146.49 - - [28/Nov/2005:04:02:29 +0800] "GET http://www.it.com.cn/f/games/053/30/0330-3.jpg HTTP/1.1" 200 40763 "-" "
162.105.146.11 - - [28/Nov/2005:04:02:34 +0800] "GET http://www.it.com.cn/f/projector/059/4/168054_36_pre.htm HTTP/1.1" 200 3
162.105.146.49 - - [28/Nov/2005:04:02:35 +0800] "GET http://pic.it.com.cn/f/desktop/0412/21/041215_dt_app_wz_11.jpg HTTP/1.1"
162.105.146.49 - - [28/Nov/2005:04:02:40 +0800] "GET http://www.it.com.cn/f/games/053/30/0330_ES_YANJING_EPS1_01.jpg HTTP/1.1
162.105.146.49 - - [28/Nov/2005:04:02:44 +0800] "GET http://it.com.cn/f/market/0410/12/34095c.jpg HTTP/1.1" 200 28078 "-" "Mo
162.105.146.49 - - [28/Nov/2005:04:02:44 +0800] "GET http://www1.it.com.cn/f/games/053/10/0310ship_game11.jpg HTTP/1.1" 200 4
162.105.146.49 - - [28/Nov/2005:04:02:48 +0800] "GET http://pic.it.com.cn/f/desktop/0412/21/041215_dt_app_wz_12.jpg HTTP/1.1"
162.105.146.49 - - [28/Nov/2005:04:02:50 +0800] "GET http://it.com.cn/f/market/0410/12/34095d.jpg HTTP/1.1" 200 33813 "-" "Mo
162.105.146.49 - - [28/Nov/2005:04:02:51 +0800] "GET http://www1.it.com.cn/f/games/053/10/0310ship_game12.jpg HTTP/1.1" 200 2
162.105.146.49 - - [28/Nov/2005:04:02:57 +0800] "GET http://www.it.com.cn/f/games/053/30/0330_ES_YANJING_EPS1_hot1.jpg HTTP/1
162.105.146.11 - - [28/Nov/2005:04:02:59 +0800] "GET http://prod.it.com.cn/dealerhtm/6334/index.html HTTP/1.1" 200 20476 "-"
162.105.146.49 - - [28/Nov/2005:04:03:00 +0800] "GET http://pic.it.com.cn/f/desktop/0412/21/041215_dt_app_wz_13.jpg HTTP/1.1"
222.201.88.52 - - [28/Nov/2005:04:03:00 +0800] "GET http://www.it.com.cn/demo/images/logo_14060.gif HTTP/1.1" 304 138 "http:/
162.105.146.11 - - [28/Nov/2005:04:03:21 +0800] "GET http://www.it.com.cn/f/mobile/0510/19/187859_6.htm HTTP/1.1" 200 70703 "
162.105.146.49 - - [28/Nov/2005:04:03:25 +0800] "GET http://it.com.cn/f/market/0410/12/34095e.jpg HTTP/1.1" 200 34578 "-" "Mo
162.105.146.49 - - [28/Nov/2005:04:03:25 +0800] "GET http://www1.it.com.cn/f/games/053/10/0310ship_game13.jpg HTTP/1.1" 200 2
162.105.146.49 - - [28/Nov/2005:04:03:25 +0800] "GET http://www.it.com.cn/f/games/053/30/0330_PC_yanjing_rock_01.jpg HTTP/1.1
162.105.146.49 - - [28/Nov/2005:04:03:28 +0800] "GET http://pic.it.com.cn/f/desktop/0412/21/041215_dt_app_wz_14.jpg HTTP/1.1"
61.157.198.172 - - [28/Nov/2005:04:03:30 +0800] "GET http://www.it.com.cn/f/edu/058/19/pp.jpg HTTP/1.1" 304 139 "http://www.d
162.105.146.11 - - [28/Nov/2005:04:03:38 +0800] "GET http://www.it.com.cn/f/edu/0510/24/189837.htm HTTP/1.1" 200 57303 "-" "P
162.105.146.11 - - [28/Nov/2005:04:03:39 +0800] "GET http://www.it.com.cn/f/edu/0510/24/189839.htm HTTP/1.1" 200 58329 "-" "P
162.105.146.11 - - [28/Nov/2005:04:03:40 +0800] "GET http://www.it.com.cn/f/mobile/0510/16/185859_3.htm HTTP/1.1" 200 69376 "
162.105.146.49 - - [28/Nov/2005:04:03:41 +0800] "GET http://pic.it.com.cn/f/desktop/0412/21/041215_dt_app_wz_15.jpg HTTP/1.1"
202.119.215.67 - - [28/Nov/2005:04:03:44 +0800] "GET http://www.it.com.cn/demo/images/it.gif HTTP/1.1" 200 10945 "http://www.
162.105.146.11 - - [28/Nov/2005:04:03:53 +0800] "GET http://www.it.com.cn/f/market/057/2/138265.htm HTTP/1.1" 200 53557 "-" "
162.105.146.49 - - [28/Nov/2005:04:03:58 +0800] "GET http://www.it.com.cn/f/games/053/30/0330_PC_yanjing_rock_02.jpg HTTP/1.1
162.105.146.49 - - [28/Nov/2005:04:03:58 +0800] "GET http://pic.it.com.cn/f/desktop/0412/21/041215_dt_app_wz_16.jpg HTTP/1.1"
162.105.146.49 - - [28/Nov/2005:04:03:58 +0800] "GET http://it.com.cn/f/market/0410/12/34095f.jpg HTTP/1.1" 200 25084 "-" "Mo
162.105.146.49 - - [28/Nov/2005:04:03:59 +0800] "GET http://www1.it.com.cn/f/games/053/10/0310ship_game14.jpg HTTP/1.1" 200 3
162.105.146.11 - - [28/Nov/2005:04:04:07 +0800] "GET http://www.it.com.cn/f/games/057/29/151034.htm HTTP/1.1" 200 57259 "-" "
162.105.146.11 - - [28/Nov/2005:04:04:30 +0800] "GET http://www.it.com.cn/f/games/057/29/151070.htm HTTP/1.1" 200 60640 "-" "
162.105.146.49 - - [28/Nov/2005:04:04:39 +0800] "GET http://pic.it.com.cn/f/desktop/0412/21/041215_dt_app_wz_17.jpg HTTP/1.1"
162.105.146.49 - - [28/Nov/2005:04:04:39 +0800] "GET http://www.it.com.cn/f/games/053/30/0330_PC_yanjing_rock_03.jpg HTTP/1.1
162.105.146.49 - - [28/Nov/2005:04:04:39 +0800] "GET http://it.com.cn/f/market/0410/12/341371012-1.jpg HTTP/1.1" 200 34697 "-
162.105.146.11 - - [28/Nov/2005:04:04:48 +0800] "GET http://www.it.com.cn/f/games/057/29/151042.htm HTTP/1.1" 200 60586 "-" "
162.105.146.49 - - [28/Nov/2005:04:05:01 +0800] "GET http://pic.it.com.cn/f/desktop/0412/21/041215_dt_app_wz_18.jpg HTTP/1.1"
162.105.146.49 - - [28/Nov/2005:04:05:37 +0800] "GET http://pic.it.com.cn/f/desktop/0412/21/041215_dt_app_wz_19.jpg HTTP/1.1"
```

数据清洗


Foreach Loop Every Log
TaskLog
CleanData
FilterFkey
Log2 to DB
Log2 Order
Truncate
Truncate Fun...
FunnelGroupSu
m_SplitAll
FunnelGroup_Sp
litAll
Log4_SplitAl
l to DB
Sessionize_Split
All
FunnelGroupSu
m_SplitAll to F...
Truncate
FunnelGroupS...
Log2Group
Log2GroupSum
Truncate
Log2Group
HeadGroup
Head to DB
Sessionize_Hea
dTail
Truncate
Log2GroupSum
Log2GroupSum
to Log2Group
HeadGroupSum
Truncate
HeadGroup
HeadGroupSum
to HeadGroup
Truncate
HeadGroupSum
Tail to DB
Truncate
TailGroupSum
TailGroupSum to
TailGroup
Truncate
TailGroup
TailGroupSum
TailGroup
Connection Managers
localhost\SQL2005.FunnelMining
localhost\SQL2005.MSN_Project
localhost\SQL2005.tempdb
log2
log3
log4_Head
log4_SplitAll
log4_Tail
Log5_KeyLevel_SplitAll
log6_KeyLevel_SplitAll

数据仓库

UserID	Country	Others
420001	USA	……
420002	USA	……
420003	China	……
420004	China	……
420005	China	……
420006	Europe	……

UserID	Keywords
420001	holywood
420001	hollywood
420002	sport express
420002	tina turner
420002	ike
420003	cbcnew
420004	footlocker
420004	overstock
420004	google
420005	goet
420005	montgomeryadvertiser
420006	msnbc
420006	macontelegraph
420006	montgomeryadvertiser

Case表

Nested表

数据挖掘

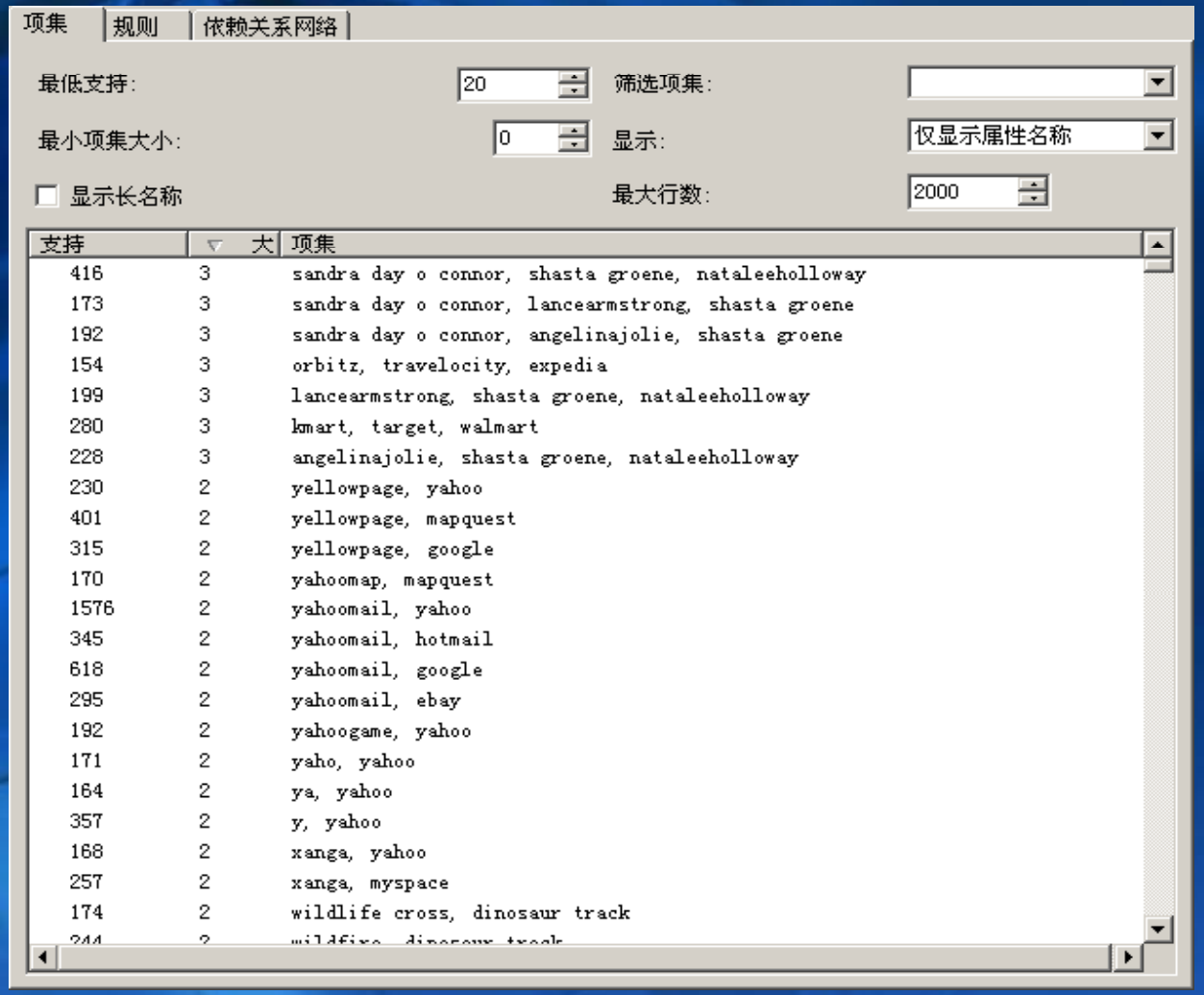
UserID	Country	Others	Keywords
420001	USA	……	holywood
			hollywood
420002	USA	……	sport express
			tina turner
			ike
420003	China	……	cbcnew
420004	China	……	footlocker
			overstock
			google
420005	China	……	goet
			montgomeryadvertiser
420006	Europe	……	msnbc
			macontelegraph
			montgomeryadvertiser

DEMO

互联网搜索关键字关联性分析

ItemSet

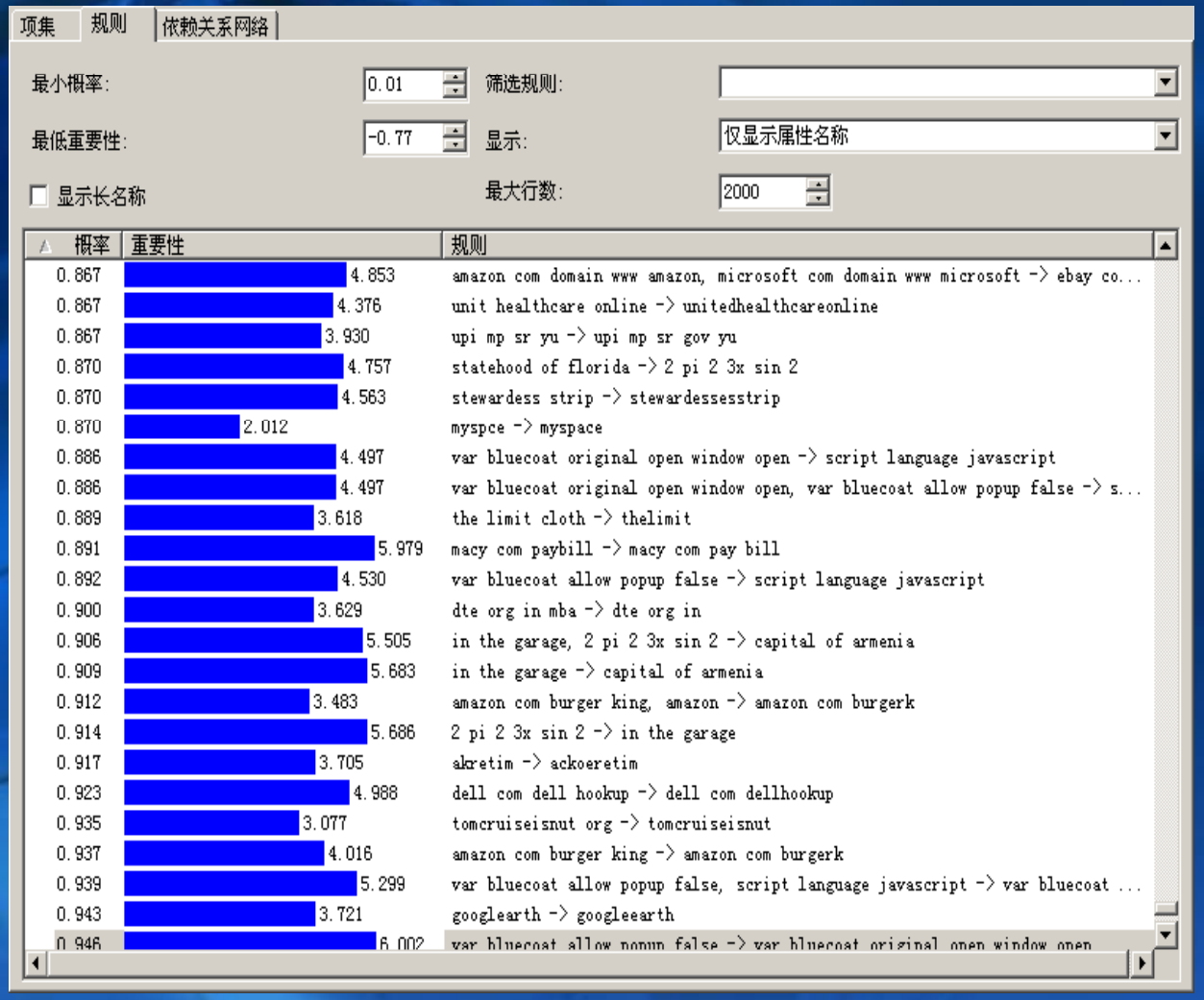
- Support
- ItemSet Size
- ItemSet



支持	大	项集
416	3	sandra day o connor, shasta groene, nataleeholloway
173	3	sandra day o connor, lancearmstrong, shasta groene
192	3	sandra day o connor, angelinajolie, shasta groene
154	3	orbitz, travelocity, expedia
199	3	lancearmstrong, shasta groene, nataleeholloway
280	3	kmart, target, walmart
228	3	angelinajolie, shasta groene, nataleeholloway
230	2	yellowpage, yahoo
401	2	yellowpage, mapquest
315	2	yellowpage, google
170	2	yahoomap, mapquest
1576	2	yahoomail, yahoo
345	2	yahoomail, hotmail
618	2	yahoomail, google
295	2	yahoomail, ebay
192	2	yahoogame, yahoo
171	2	yaho, yahoo
164	2	ya, yahoo
357	2	y, yahoo
168	2	xanga, yahoo
257	2	xanga, myspace
174	2	wildlife cross, dinosaur track
244	2	wildfire, dinosaur track

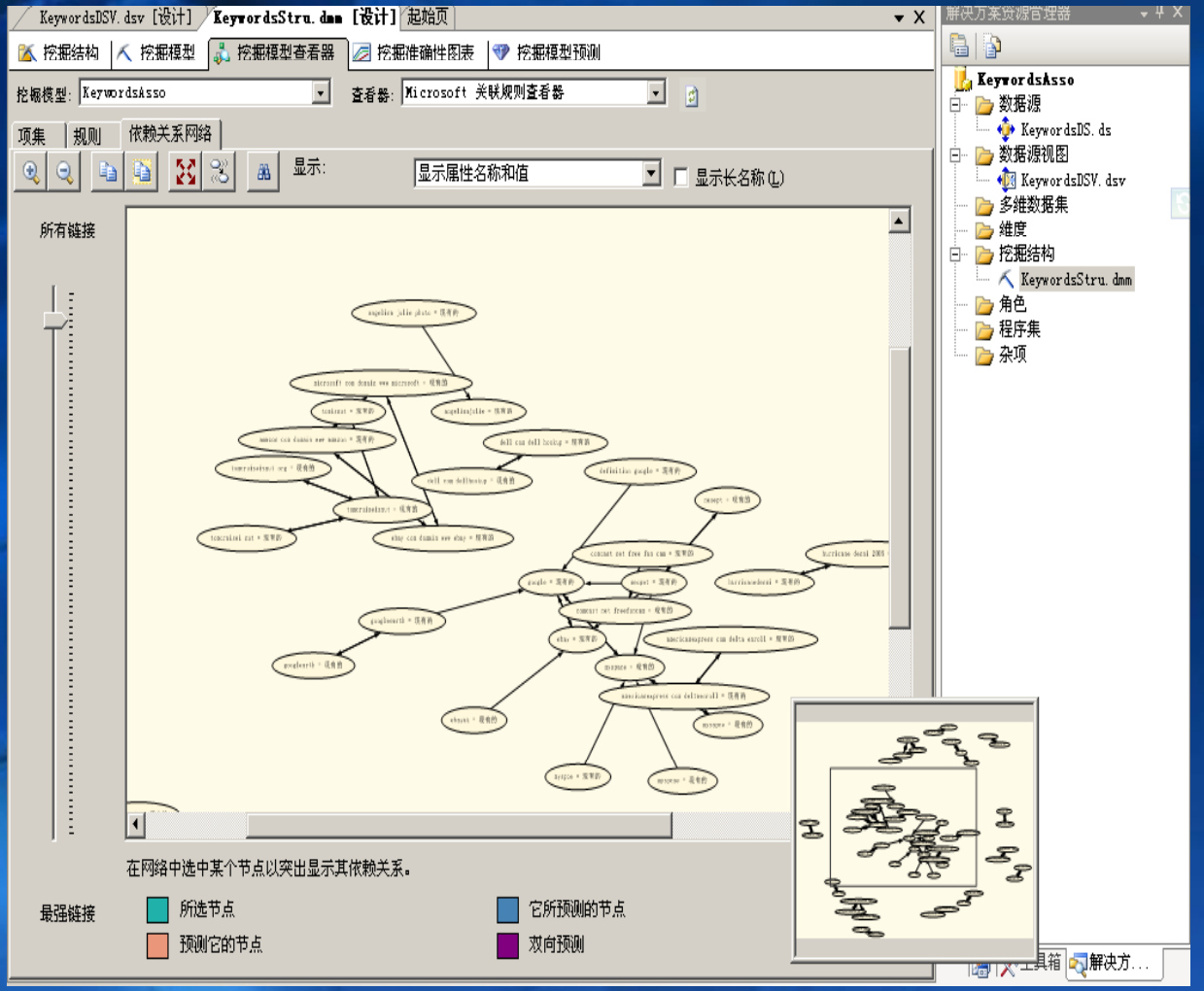
Rule

- Probability
- Importance
- Rule



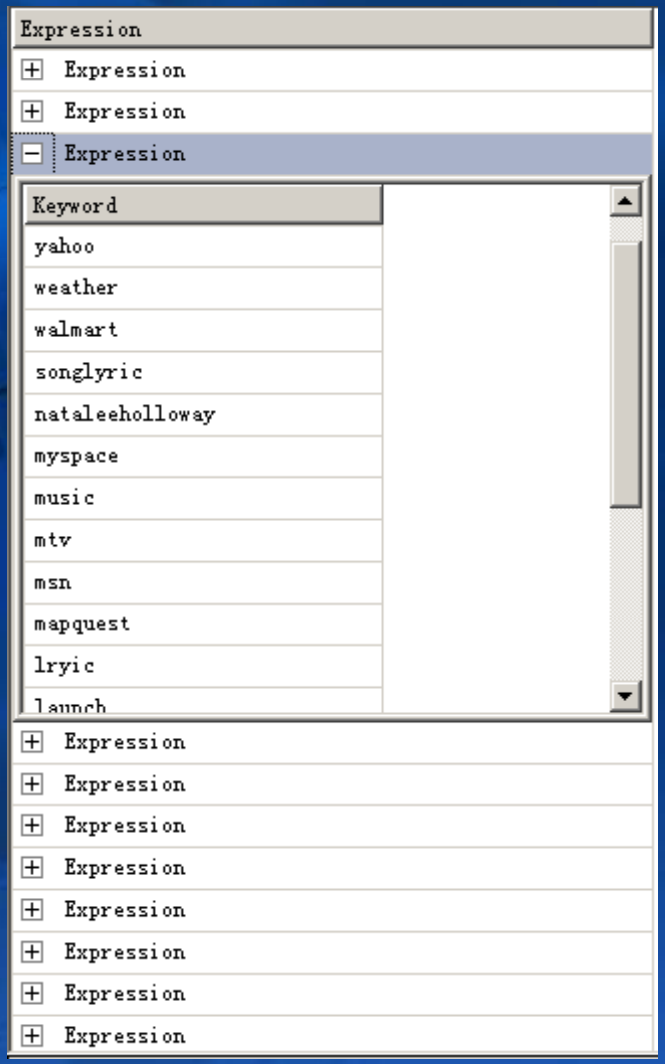
Dependence NetWork

- 滑块
- 节点的颜色
- 有无连线
- 连线的箭头指向



预测查询

```
SELECT  Predict([KeywordsAsso].[Fact])
From  [KeywordsAsso]
PREDICTION JOIN
  SHAPE {
  OPENQUERY([KeywordsDS],
    'SELECT   [userID]  FROM [dbo].[User]
    ORDER BY  [userID]')}
  APPEND
  ({OPENQUERY([KeywordsDS],
    'SELECT  [Keyword], [userID]
FROM [dbo].[Fact]
    ORDER BY  [userID]')}
    RELATE  [userID] TO [userID])
    AS  [Fact] AS t
ON
  [KeywordsAsso].[Fact].[Keyword] =
t.[Fact].[Keyword]
```



预测查询

```
SELECT
  Predict([KeywordsAsso].[Fact],10)
From
  [KeywordsAsso]
NATURAL PREDICTION JOIN
(SELECT (SELECT 'bestbuy' AS [Keyword])
AS [Fact]) AS t
```

Keyword
circuitcity
walmart
compusa
radioshack
circutcity
staple
officedepot
officemax
samclub
sony

DEMO

搜索关键字关联关系优化处理一

处理方法

使用Include_Statistics

Keyword	$SUPPORT	$PROBABILITY	$ADJUSTEDPROBABILITY
circuitcity	2468	0.1863799283154122	0.90066080161131812
walmart	9628	0.11148839841539332	0.78023243959031052
compusa	704	0.038671948688926615	0.88548651432491554
radioshack	877	0.027541973212601396	0.85394439012273793
circutcity	436	0.026976042256178081	0.891381427502506
staple	1497	0.024712318430484815	0.80912247291167827
officedepot	1518	0.023580456517638182	0.80438211142821436
officemax	1147	0.023580456517638182	0.825495478182903
samclub	1171	0.013393699302018487	0.7791744740111396
sony	653	0.010375400867608	0.806153388579432

DEMO

搜索关键字关联关系优化处理二

处理方法

- 使用AdjustedProbability而不使用Probability
- AdjustedProbability = PredProb * (1 - MargProb) ^ SomeConstant

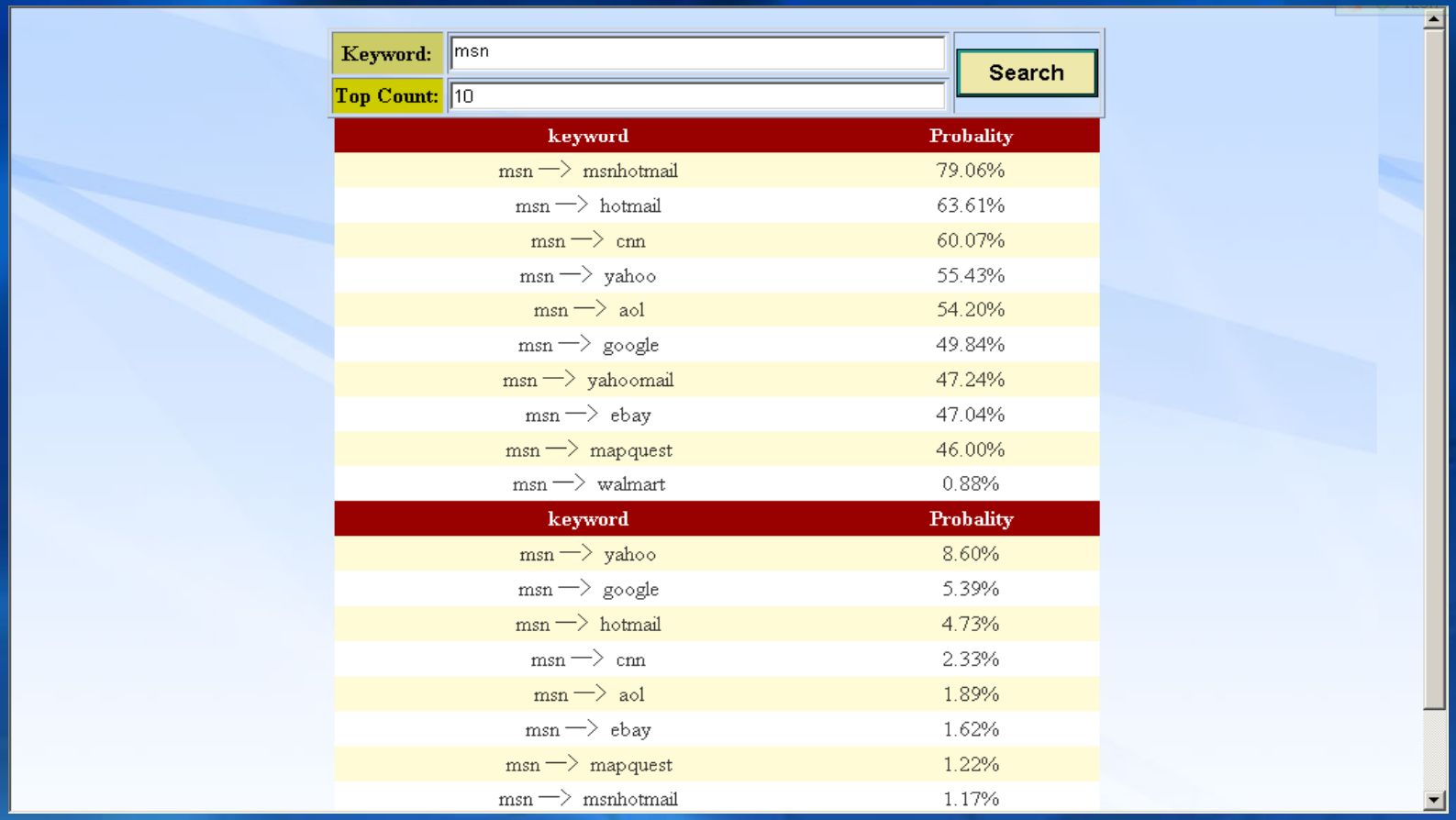
Keyword	$SUPPORT	$PROBABILITY	$ADJUSTEDPROBABILITY
circuitcity	2468	0.1863799283154122	0.90066080161131812
circutcity	436	0.026976042256178081	0.891381427502506
compusa	704	0.038671948688926615	0.88548651432491554
radioshack	877	0.027541973212601396	0.85394439012273793
officemax	1147	0.023580456517638182	0.825495478182903
staple	1497	0.024712318430484815	0.80912247291167827
sony	653	0.0103754008677608	0.806153388579432
officedepot	1518	0.023580456517638182	0.80438211142821436
walmart	9628	0.11148839841539332	0.78023243959031052
samclub	1171	0.013393699302018487	0.7791744740111396

DEMO

搜索关键字关联关系优化处理三

处理方法

把挖掘结果部署在应用程序当中



Keyword: msn

Top Count: 10

Search

keyword	Probality
msn —> msnhotmail	79.06%
msn —> hotmail	63.61%
msn —> cnn	60.07%
msn —> yahoo	55.43%
msn —> aol	54.20%
msn —> google	49.84%
msn —> yahoomail	47.24%
msn —> ebay	47.04%
msn —> mapquest	46.00%
msn —> walmart	0.88%

keyword	Probality
msn —> yahoo	8.60%
msn —> google	5.39%
msn —> hotmail	4.73%
msn —> cnn	2.33%
msn —> aol	1.89%
msn —> ebay	1.62%
msn —> mapquest	1.22%
msn —> msnhotmail	1.17%

主要内容

- SQL Server 2005数据挖掘概览
- SQL Server 2005数据挖掘具体运用
- 其他工具的整合及其二次开发

和IS整合

- Analysis Services Processing Task
- Data Mining Model Training
- Data Mining Query Task
- Data Mining Query

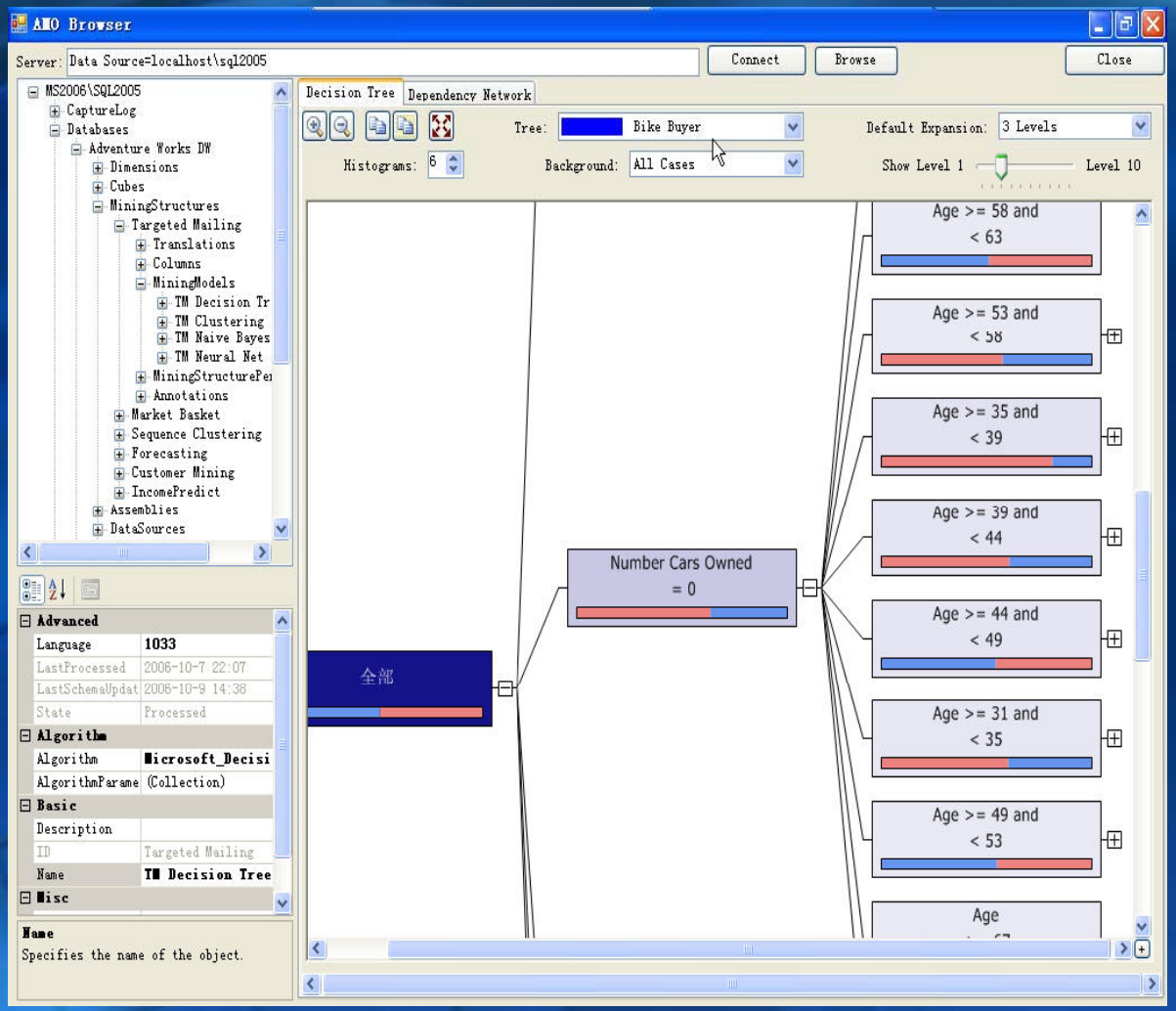
多维数据库架构信息界挖掘模型的浏览

功能：

挖掘模型的浏览

分析服务器的架构信息浏览

分析服务器中个对象属性的浏览



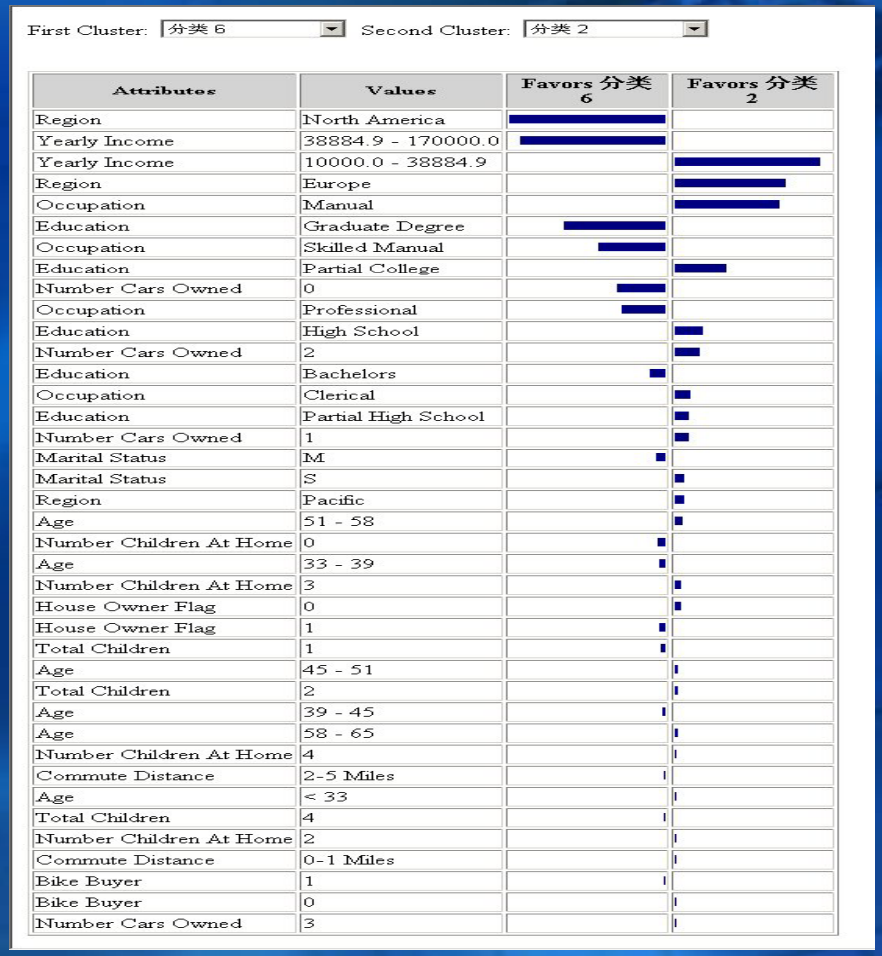
利用DataMiningHTMLViewers

三种算法：决策树、聚类、**Naïve Bayes**

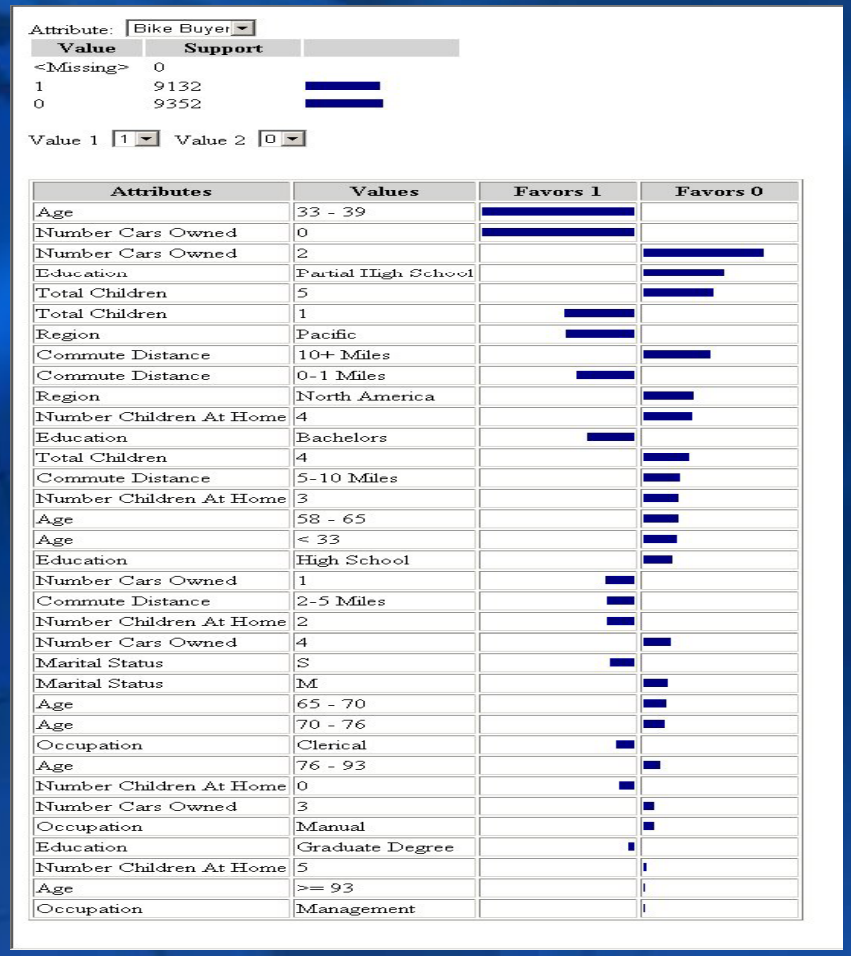
DataMiningHTMLViewers

只有图表

利用DataMiningHTMLViewers


First Cluster: 分类 6 Second Cluster: 分类 2

Attributes	Values	Favors 分类 6	Favors 分类 2
Region	North America		
Yearly Income	38884.9 - 170000.0		
Yearly Income	10000.0 - 38884.9		
Region	Europe		
Occupation	Manual		
Education	Graduate Degree		
Occupation	Skilled Manual		
Education	Partial College		
Number Cars Owned	0		
Occupation	Professional		
Education	High School		
Number Cars Owned	2		
Education	Bachelors		
Occupation	Clerical		
Education	Partial High School		
Number Cars Owned	1		
Marital Status	M		
Marital Status	S		
Region	Pacific		
Age	51 - 58		
Number Children At Home	0		
Age	33 - 39		
Number Children At Home	3		
House Owner Flag	0		
House Owner Flag	1		
Total Children	1		
Age	45 - 51		
Total Children	2		
Age	39 - 45		
Age	58 - 65		
Number Children At Home	4		
Commute Distance	2-5 Miles		
Age	< 33		
Total Children	4		
Number Children At Home	2		
Commute Distance	0-1 Miles		
Bike Buyer	1		
Bike Buyer	0		
Number Cars Owned	3		

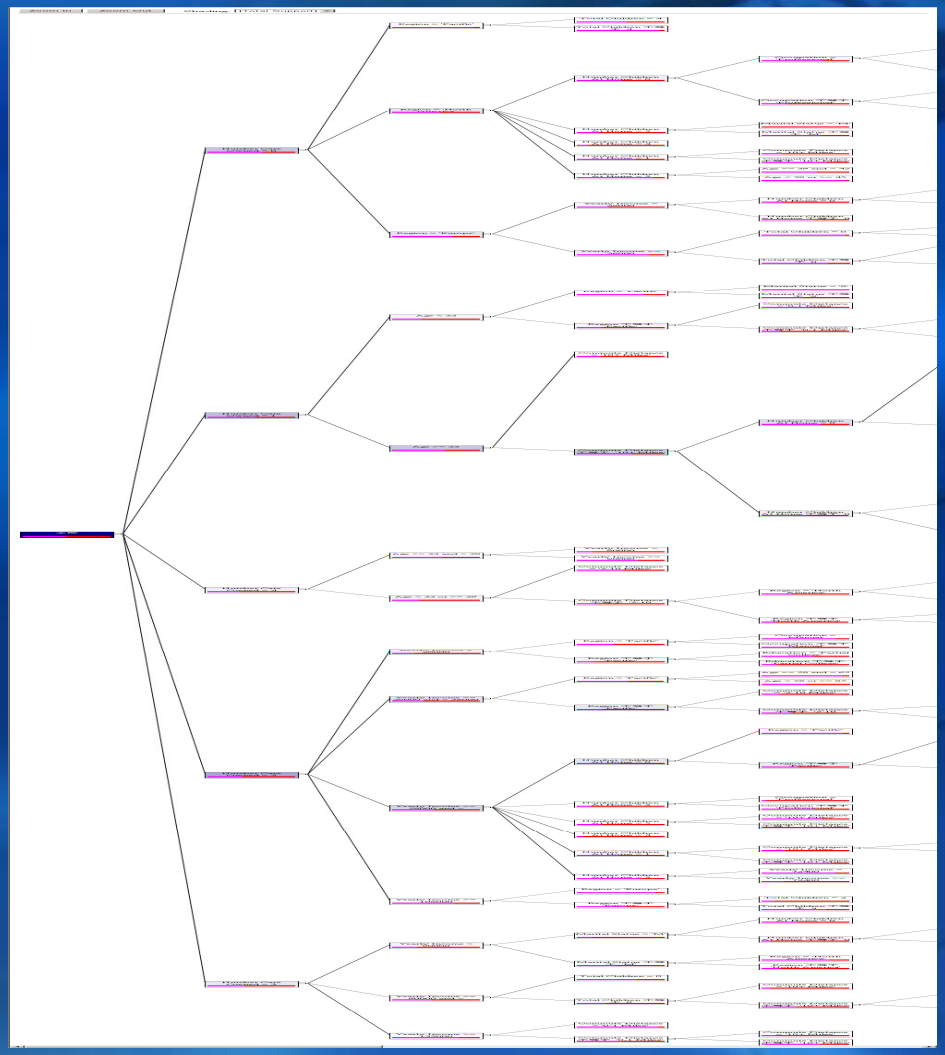

Attribute: Bike Buyer

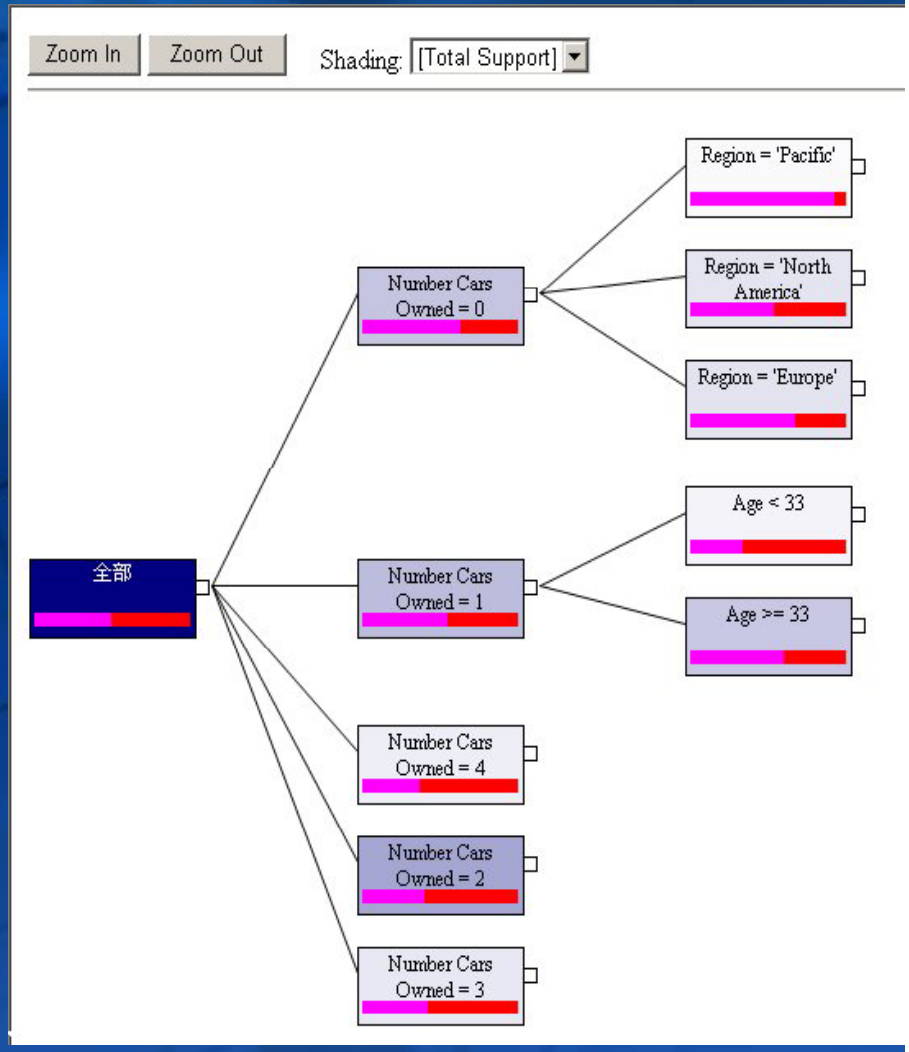
Value	Support
<Missing>	0
1	9132
0	9352

Value 1 1 Value 2 0

Attributes	Values	Favors 1	Favors 0
Age	33 - 39		
Number Cars Owned	0		
Number Cars Owned	2		
Education	Partial HIgh School		
Total Children	5		
Total Children	1		
Region	Pacific		
Commute Distance	10+ Miles		
Commute Distance	0-1 Miles		
Region	North America		
Number Children At Home	4		
Education	Bachelors		
Total Children	4		
Commute Distance	5-10 Miles		
Number Children At Home	3		
Age	58 - 65		
Age	< 33		
Education	High School		
Number Cars Owned	1		
Commute Distance	2-5 Miles		
Number Children At Home	2		
Number Cars Owned	4		
Marital Status	S		
Marital Status	M		
Age	65 - 70		
Age	70 - 76		
Occupation	Clerical		
Age	76 - 93		
Number Children At Home	0		
Number Cars Owned	3		
Occupation	Manual		
Education	Graduate Degree		
Number Children At Home	5		
Age	>= 93		
Occupation	Management		

利用DataMiningHTMLViewers






DMAddin
DEMO


Thank You!