# 一种新的本体映射发现方法SME

夏红科[1,2] 郑雪峰[1] 胡 祥[3,4]
(北京科技大学信息工程学院 北京 100083)[1] (北京信息科技大学计算机学院 北京 100192)[2]
(北京邮电大学计算机科学与技术学院 北京 100876)[3]
(华北电力大学计算机科学与技术学院 北京 102206)[4]

**摘 要** 本体映射是基于本体的语义查询与集成的基础。本体映射发现的任务是从源本体和目标本体的相似度中发现本体映射,它是本体映射的关键。将本体映射发现问题看成是集合覆盖问题,提出一种基于集合覆盖的本体映射发现方法SME(SCM-based Mapping Extraction),该方法在训练阶段找到最大程度覆盖训练数据的属性集,在测试阶段利用这些属性集在测试数据上对应属性值的交操作来发现映射。实验证明该方法具有较好的综合性能。

**关键词** 映射发现,集合覆盖,数据依赖球,属性空间
**中图法分类号** TP181 **文献标识码** A

## Novel Approach of Ontology Mapping Extraction SME

XIA Hong-ke[1,2] ZHENG Xue-feng[1] HU Xiang[3,4]
(Information Engineering Institute, University of Science & Tech. Beijing, Beijing 100083, China)[1]
(Computer School, Beijing Info. Science & Tech. University, Beijing 100192, China)[2]
(School of Computer Science & Tech., Beijing University of Posts and Telecommunications, Beijing 100876, China)[3]
(Computer Science & Tech. Dept., North China Electric Power University, Beijing 102206, China)[4]

**Abstract** Ontology mapping is the foundation of semantic query and semantic integration based on ontology. As the crucial point of ontology mapping, the task of mapping extraction is to find whether there exists the ontology mapping among the similarities between source ontology and target ontology. In this paper the problem of mapping extraction was regarded as the problem of set covering, and a novel ontology mapping extraction algorithm based on set covering SME (SCM-based Mapping Extraction) was proposed, by which the property set that covers the training data set in maximum degree was searched during training stage, and the mapping extraction was carried out by means of the conjunction of the properties of property set in testing data set during testing stage. Experimental evaluations show that this method has better comprehensive performances compared with other algorithms.

**Keywords** Mapping extraction, Set covering, Data-dependent ball, Property spaces

## 1 引言

本体映射是信息集成、语义Web和知识管理等领域的一个关键问题。在实现本体映射的各类方法中,基于相似度计算的本体映射是最常用的一种方法,如利用该方法计算源本体和目标本体中两个实体元素之间的相似度,并从中发现本体映射。其中各种相似度计算方法可以分为两大类:单相似度模型与多相似度模型。单相似度模型仅利用一种方法来计算相似度,如Jaccard系数[1]、Cosine方法[2]等都属于单相似度模型;多相似度模型则组合了两种或多种相似度计算方法,如多维权重法[3]、Similarity Flooding[4]、Hybrid方法[5]等都属于多相似度模型。由于各种相似度计算方法分别利用了本体的各种信息,如语言相似性、实例相似性、结构相似性等来计算相似度,因此多相似度模型通常能获得比单相似度模型更好的映射结果。

在计算源本体和目标本体之间概念对的相似度之后,需要进行映射发现。已有的映射发现机制包括使用阈值的选择策略、最大预测值的选择策略[6]、松弛标注(relaxation labeling)的选择策略[5]等。

使用阈值的选择策略首先定义一个阈值,通过判断所有的概念对相似度与阈值之间的关系来确定映射。如果两个概念的相似度大于该阈值,则这两个概念之间存在映射关系,反之则不存在映射关系。阈值法最大的不足在于阈值的选择,由于不同的阈值对最终的映射发现影响很大,因此怎样合理地选择阈值是一个需要深入考虑的问题。

最大预测值的选择策略将所有的概念相似度按降序排

到稿日期:2009-07-28 返修日期:2009-10-19 本文受国家高技术研究发展(863)计划(2006AA01Z231),北京信息科技大学校基金(5026010934)资助。

夏红科(1979—),女,博士生,讲师,主要研究方向为本体及语义网、智能信息处理,E-mail:hongke.xia@gmail.com;郑雪峰(1951—),男,教授,主要研究方向为计算机网络、信息安全;胡 祥(1976—),男,博士生,讲师,主要研究方向为机器学习、智能信息处理。

列，从中选择相似度最大的概念对作为存在映射关系的概念对。该方法本质上与阈值法可以相互替代，Falcon-AO[7]，Rimom[8]都是采用这种方法来发现映射的。

松弛标注法作为图形、图像处理及自然语言处理等领域广泛应用的一种方法，主要思想是一个节点的标注(label)受其邻居的影响，因此可以通过一个节点的邻居标注来判断该节点的标注。在本体映射中，将目标本体的概念看作标注，并用松弛标注法找到最适合源本体中某个概念的标注。GLUE算法[5]采用松弛标注法来发现映射。

本文提出了一种新的本体映射发现方法，该方法将集合覆盖理论引入到映射发现中，从而把映射发现问题看成集合覆盖问题，通过在训练集上找到能最大程度覆盖训练数据的属性集，利用该属性集来判断测试数据是否存在映射。本文第2节介绍相关概念；第3节介绍本体映射发现，提出基于集合覆盖的映射发现算法SME；第4节给出实验结果；最后总结全文并提出进一步的工作。

## 2 集合覆盖问题

集合覆盖问题的数学描述如下[9]：$\{X,F\}$由一个有限集$X$及$X$的一个子集族$F$组成，子集族$F$覆盖了有限集$X$，也就是说$X$中每一元素至少属于$F$中的一个子集，即$X=\bigcup_{S\in F}S$。对于$F$的一个子集$C\subseteq F$，若$C$中的$X$的子集覆盖了$X$，即$X=\bigcup_{S\in C}S$，则称$C$覆盖了$X$。集合覆盖问题就是要找出$F$中覆盖$X$的最小子集$C^*$，使得$|C^*|=\min\{|C||C\subseteq F$且$C$覆盖$X\}$。集合覆盖问题的解是用集合$S$的某些子集来表示的，这些子集彼此之间并无联系，仅需要研究集合$S$中哪些子集会以较高的概率被选作最优解的子集。

集合覆盖问题作为一个NP完全问题，在实际计算中经常使用的近似解法是贪心算法[10]，即每一次在$F$中寻找能最大程度覆盖$X$的集合$S$，将其作为结果集并入到$C^*$中。SCM算法(Set Covering Machine)是近年来求解集合覆盖问题的一种较好的方法。该算法是由Marchand与Shawe-Taylor[11]提出，并在Valiant算法[12]与Haussler算法[13]的基础上改进而来的一种基于贪心算法的新型学习算法。

## 3 基于集合覆盖的本体映射发现

### 3.1 问题描述

本文将集合覆盖引入到本体映射发现问题中，根据SCM算法的思想提出本体映射发现算法SME，这里主要考虑最小相关属性集的交操作。映射发现问题描述如表1所列。在训练阶段，训练数据的形式为(ID，$Sim_1$，$Sim_2$，…，$Sim_n$，Class)，其中ID表示概念对$(ca_i,cb_j)$，$ca_i$和$cb_j$分别代表需要映射的源本体和目标本体中的概念。$Sim_1$，$Sim_2$，…，$Sim_n$表示概念对$(ca_i,cb_j)$的相似度，$Sim_k$表示用第$k$种概念相似度计算策略得到的概念对$(ca_i,cb_j)$的相似度。Class代表映射的有效性，Class$\in\{0,1\}$，1(positive)表示存在映射，0(negative)表示不存在映射。训练阶段找到覆盖属性空间的最小属性集合，测试阶段对于测试数据(ID，$Sim_1$，$Sim_2$，…，$Sim_n$)，利用训练出来的最小相关属性集的交操作来确定测试数据的类别Class(是否存在映射)。

表1 本体映射发现问题的描述

| | ID | $Sim_1$ | $Sim_2$ | … | $Sim_n$ | Class |
|---|---|---|---|---|---|---|
| Training | $(ca_1,cb_1)$ | 0.75 | 0.64 | … | 0.9 | 1(Positive) |
| | $(ca_1,cb_2)$ | 0.11 | 0.02 | … | 0 | 0(Negative) |
| | … | … | … | … | | |
| Testing | $(ca_7,cb_5)$ | 0.4 | 0.38 | … | 0.6 | ? |
| | … | … | … | … | … | … |

### 3.2 基于集合覆盖的映射发现算法SME

**定义1** 假定输入空间$S$是$R^n$的一个子集，$x$表示输入空间的一个$n$维向量，$x_i\in R^n$；$y$表示输入数据所属的类别，$y_i\in\{1,0\}$。定义$P$为肯定的训练样例集合，$\{P\}=\{x_i,x_i\in S,\text{iff } y_i=1\}$，定义$N$为否定的训练样例集合，$N=\{x_i,x_i\in S,\text{iff } y_i=0\}$。

从定义1可知，$P$和$N$是$S$的一个划分，$P\cup N=S$，$P\cap N=\emptyset$，其中输入向量$x$的格式为($Sim_1$，$Sim_2$，…，$Sim_n$)。

**定义2** 对于每个具有类别$y_i$($y_i\in\{1,0\}$)的训练数据$x_i$以及给定的半径$\rho$，定义$h_{i,\rho}$为以$x_i$为圆心、$\rho$为半径的数据依赖球；定义$H$为$h_{i,\rho}$(简记为$h_i$)组成的属性空间，$H=\{h_i(x)\}_{i=1}^{|H|}$，其中数据依赖球$h_{i,\rho}$为$H$中的属性。

对于任一训练数据集$S$，可确定出相应的$h_{i,\rho}$及$H$。$h_{i,\rho}$确定后，任意一个训练样例$x_j$与$h_{i,\rho}$之间的关系可以表示成

$$h_{i,\rho}(x_j)\triangleq\begin{cases}y_i & \text{if } d(x_i,x_j)\leqslant\rho\\ \bar{y}_i & \text{if } d(x_i,x_j)>\rho\end{cases}$$

式中，$d(x_i,x_j)$表示$x_i$与$x_j$之间的距离，$x_i$与$x_j$均为$n$维向量，考虑到可分性，可通过映射函数$\phi(x)$将$d(x_i,x_j)$映射到高维空间，$d(x_i,x_j)=\|\phi(x_i)-\phi(x_j)\|$。根据RKHS理论[14]，对于任何确定的核函数$k(x_i,x_j)$，都存在一个隐含的映射函数$\phi(x)$满足$k(x_i,x_j)=\phi(x_i)\cdot\phi(x_j)$，因此$d(x_i,x_j)$可转换成$d(x_i,x_j)=\|\phi(x_i)-\phi(x_j)\|=\sqrt{k(x_i,x_i)+k(x_j,x_j)-2k(x_i,x_j)}$。

如图1所示，如果数据$x_j$到$x_i$之间的距离小于或等于$\rho$，则$x_j$位于球$h_{i,\rho}$内；否则$x_j$位于球$h_{i,\rho}$外。由此可知，一旦确定了$h_{i,\rho}$，$S$中所有训练数据$x_j$($x_j\in S$)与$h_{i,\rho}$之间的关系就只有两种：$x_j$位于球$h_{i,\rho}$内或者是$x_j$位于球$h_{i,\rho}$外。因此，对于每一个数据依赖球$h_{i,\rho}$，在$P$和$N$上分别能形成与$h_{i,\rho}$相关的球内和球外两部分数据集合。

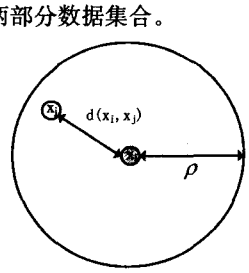


图1 数据依赖球$h_{i,\rho}$

**定义3** 对于给定的训练集$S$中每一个数据依赖球$h_{i,\rho}$，定义$Q_i$为球$h_{i,\rho}$覆盖到的否定样例集，$Q_i=\{x_j,x_j\in N,\text{iff } d(i,j)\leqslant\rho\}$；定义$E_i$为球$h_{i,\rho}$未覆盖到的肯定样例集，$E_i=\{x_j,x_j\in P,\text{iff } d(i,j)>\rho\}$。

$Q_i$，$E_i$与$P$，$N$的关系如表2所列。

表2 $h_{i,\rho}$对应的$Q_i$，$E_i$

| | training examples S | |
|---|---|---|
| | $x_i\in P$ | $x_i\in N$ |

| | | | | |
|---|---|---|---|---|
| covered by $h_i(x)$ | / | | $x_i \in P$ | $Q_i=\{x_j, x_j \in N, \text{iff } d_{ij}>\rho_i\}$ |
| | | | $x_i \in N$ | $Q_i=\{x_j, x_j \in N, \text{iff } d_{ij}\leqslant\rho_i\}$ |
| not covered by $h_i(x)$ | $x_i \in P$ | $E_i=\{x_j, x_j \in P, \text{ iff } d_{ij}>\rho_i\}$ | | |
| | $x_i \in N$ | $E_i=\{x_j, x_j \in P, \text{iff } d_{ij}\leqslant\rho_i\}$ | | |

对于某一给定的 $h_{i,\rho}$，$Q_i$ 表示 $h_{i,\rho}$ 在 $N$ 上正确覆盖的训练样例集合，反映了 $h_{i,\rho}$ 在 $N$ 上的正确覆盖程度；$E_i$ 表示 $h_{i,\rho}$ 在 $P$ 上未覆盖到的训练样例集，反映了 $h_{i,\rho}$ 在 $P$ 上的错误覆盖程度。表 2 显示了当 $h_{i,\rho}$ 为正/负数据依赖球时计算 $Q_i$，$E_i$ 的方法。当 $h_{i,\rho}$ 覆盖到 $S$ 中更多的训练样例时，由 $h_{i,\rho}$ 所确定的 $|Q_i|$ 越大，$|E_i|$ 越小。因此，通过与 $h_{i,\rho}$ 相关的 $|Q_i|$，$|E_i|$，可以判断出 $h_{i,\rho}$ 在训练集 $S$ 上的覆盖程度。这里用有效值 $U_i$ 表示 $h_{i,\rho}$ 在 $S$ 上的覆盖有效性。

**定义 4** 对于 $H$ 中的每个属性 $h_{i,\rho}$ 及惩罚因子 $p$，定义 $h_{i,\rho}$ 的覆盖有效性 $U_i$ 为 $U_i=|Q_i|-p\cdot|E_i|$。

因此，寻找能最大覆盖 $S$ 的数据依赖球 $h_k$ 的过程可以表示为 $k=\underset{i}{\text{argmax}}U_i=\underset{i}{\text{argmax}}(|Q_i|-p\cdot|E_i|)$。在训练阶段，每次找到能最大限度覆盖 $S$ 的属性依赖球 $h_k(h_k\in H)$，最终找到能最大限度覆盖 $S$ 的数据依赖球的集合 $R(R\subset H)$。在测试阶段判断测试数据针对 $R$ 中每个数据依赖球 $h_k$ 的所属类别，可通过各类别的交运算得到该测试数据的最终类别。

例如图 2 中，“+”代表 class=1 的训练样例，“−”代表 class=0 的样例。图中在训练阶段找到的依赖球由 3 个属性 $h_1$，$h_2$，$h_3$ 组成，测试阶段对于测试数据 $x$（图中带? 的位置）所属类别的判断是通过判断 $x$ 对于这 3 个数据依赖球所属类别的交集来进行的，记为 $f(x)=\wedge_{i\in R}h_i(x)$。整个算法如图 3 所示。

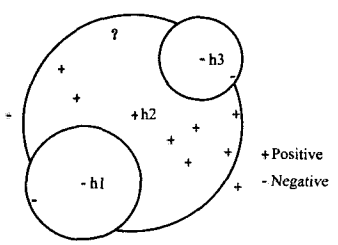


图 2 通过 SME 算法发现映射的过程

SME($S, p, s, T, C$)
Input: A training set $S$, a penalty value $p$, a stopping point $s$, a testing data set $T$
Output: the category set $C$ of $T$
// Training Stage, Preprocessing
(1) $R=\emptyset, H=\emptyset$,
(2) InitializationPN($S, P, N$)
//Constructing $H$
(3) Foreach $x_i$ in $S$
Foreach $x_j$ in $P$
(3.1) $\rho=d(x_i, x_j)$
(3.2) $H=H\cup\{h_{i,\rho}\}$
(4) Foreach $h_i$ in $H$
InitializationQE($h_i, Q_i, E_i$)
// Constructing $R$
(5) Do
//Finding $h_k$ with most useful value
(5.1) $k=\underset{i}{\text{argmax}}\{|Q_i|-p\cdot|E_i|\}$
(5.2) $R\leftarrow R\{h_k\}$
//Updating $P$ and $N$
(5.3) $N\leftarrow N-Q_k, P\leftarrow P-E_k$
//Updating $Q_i, E_i$ of all $h_i$ in $H$
(5.4) Foreach $h_i$ in $H$
$Q_i\leftarrow Q_i-Q_k, E_i\leftarrow E_i-E_k$
While $N=\emptyset$ or $|R|\geqslant s$
//Testing Stage, Judging class of testing set
(6) Foreach $x_j$ in $T$
(6.1) $C_j=$True
(6.2) Foreach $h_i$ in $H$
$C_j=C_j\wedge h_i(x_j)$

图 3 SME 算法描述

## 4 实验结果

### 4.1 实验数据

本文采用 OAEI(Ontology Alignment Evaluation Initiative)2007 竞赛[15] 所提供的 Benchmark 数据集进行实验。Benchmark 数据集包含了 50 个本体，共分为 3 组(R1XX, R2XX, R3XX)。其中 R1XX 中的 101 本体作为参考本体。每组映射都是以该本体作为目标本体，因此在 Benchmark 数据集中，全部 50 个本体作为源本体，101 本体作为目标本体。本文将 R1XX 以及 R2XX 的本体作为训练数据，R3XX(包括 301, 302, 303, 304)作为测试数据。参照文献[16]，采用(平均)查准率 MP、(平均)查全率 MR 以及 F-Measure 作为评价的主要准则，它们的定义为

平均查全率

$$MR=\sum_{k=1\cdots n}R_kN_k/\sum_{k=1\cdots n}N_k$$

平均查准率

$$MP=\sum_{k=1\cdots n}R_kN_k/\sum_{k=1\cdots n}(R_kN_k/P_k)$$

F-Measure

$$F=2/(1/MR+1/MP)$$

式中，$R_k$，$P_k$，$N_k$ 分别为第 $k$ 组映射中查全率、查准率和参考映射结果中映射的数量。

### 4.2 实验结果

在 R3XX 上的实验结果如表 3 所列。

表 3 SME 算法在 R3XX 上的实验结果

| Datasets | R3XX | | | |
|---|---|---|---|---|
| | 301 | 302 | 303 | 304 |
| Precision | 0.91 | 0.79 | 0.68 | 0.91 |
| Recall | 0.82 | 0.69 | 0.86 | 0.97 |
| MP | 0.83 | | | |
| MR | 0.85 | | | |
| F-Measure | 0.84 | | | |

图 4 给出了 SME 算法在平均查准率(MP)、平均查全率(MR)以及 F-Measure 方面与其它算法的比较结果。水平坐标表示实验的各算法，垂直坐标表示平均查准率、平均查全率以及 F-Measure 的值。

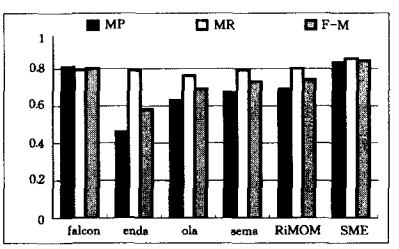


图 4 SME 与其他算法的比较结果

从图 4 可以看出，在平均查全率及 F-Measure 上，本文提

出的方法比起其他算法都有所提高，MP 平均提高 21.4%（从 2.4%～44.6%），MR 平均提高 7.5%（从 5.9%～10.6%），F-measure 平均提高 15.7%（从 4.8%～31%）。整体说来，SME 算法具有较好的综合性能。

**结束语** 本文提出一种新的本体映射发现方法 SME。作为一种实例压缩学习方法，SME 从训练数据集中获得能最大程度覆盖训练数据的数据依赖球的集合（压缩集），通过压缩集中数据依赖球在测试集上的交操作来判断测试数据的类别（是否存在本体映射），其泛化误差边界依赖于压缩集的规模[11]。怎样降低泛化误差，提高映射发现的准确性，是今后进一步的研究方向。

## 参 考 文 献


[1] Doan A, Madhavan J, Domingos P, et al. Learning to map between ontologies on the semantic Web[C]//Proc. of 11th World Wide Web Conference. 2002:662-673

[2] Lacher M S, Groh G. Facilitating the exchange explicit knowledge through ontology mappings[C]//Proc. of 14th Intl. Florida Artificial Intelligence Research Society Conf. 2001:305-309

[3] Ehrig M, Sure Y. Ontology mapping - an integrated approach [M]. Bussler C, Davies J, Fensel D, et al. , eds. Lecture Notes in Computer Science. Vol. 30653, Springer-Verlag, 2004:76-91

[4] Melnik S, Molina-Garcia H, Ralm E. Similarity flooding: a versatile graph matching algorithm[C]//Proc. of the 18th International Conference on Data Engineering (ICDE 2002). San Jose, California: IEEE Press, 2002:117-128

[5] Madhavan B J P, Rahm E. Generic schema matching with cupid [C]//Proc. of VLDB. 2001:49-58

[6] Noy N F, Musen M A. PROMPT: Algorithm and tool for automated ontology merging and alignment[C]//Proceedings of the 2000 National Conference on Artificial Intelligence. Austin, Texas, 2000:450-455

[7] Hu Wei, Cheng Gong, Zheng Dongdong, et al. The results of falcon-ao in the oaei2006 compaingn[C]//Proc. of the ISWC 2006 Workshop on Ontology Matching. Athens, GA, USA, November 2006

[8] Tang Jie, Li Juanzi, Liang Bangyong, et al. Using Bayesian decision for ontology mapping[J]. Web Semantics: Science, Service and Agents on the World-Wide Web, 2006, 4(12)

[9] 权光日. 集合覆盖问题的启发函数算法[J]. 软件学报, 1998, 9(2)

[10] Chvátal V. A greedy heuristic for the set covering problem[J]. Mathematics of Operations Research, 1979, 4:233-235

[11] Marchand M, Shawe-Taylor J. The Set Covering Machine[J]. Journal of Machine Learning Research, 2002, 3:723-746

[12] Valiant L G. A theory of the learnable [J]. Communications of the Association of Computing Machinery, 1984, 27(11): 1134-1142

[13] Haussler D. Quantifying inductive bias : AI learning algorithms and Valiant's learing framework [J]. Artificial Intelligence, 1998, 36:177-221

[14] Haussler D. Convolution kernels on discrete structures [R]. USCS-CRL-99-10. Computer Science Department, University of California in Santa Cruz, July 1999

[15] http://oaei.ontologymatching.org/2007/

[16] Rodriguez M A, Egenhofer M J. Determing semantic similarity among entity classes from different ontologies[J]. IEEE Trans on Knowledge and Data Engineering, 2003, 15(2):442-456


（上接第 225 页）

用率的值一下跃至 0.18 左右，显然这时的加分政策利用率值是让人非常满意的。让我们来解读这组数据。“集居少数民族”的加分政策享受人数的高增长没有带动加分政策录取利用人数的高增长，体现在图 7 中的 2002 年到 2006 年的加分政策录取利用率的值一直维持在低位。显然，招生部门发现了这一问题，故而在 2007 年开始调整了该项加分政策的加分规则：由原先的“考生报考重点本科院校享受 10 分的加分、报考一般本科院校享受 20 分的加分”调整为“考生报考所有院校都享受 20 分的加分”。很显然，该调整方案取得了很好的效果，加分政策的录取利用率重新回到高位值。

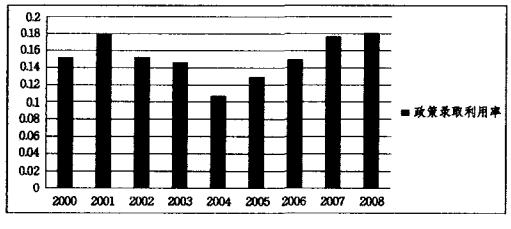


图 7 “集居少数民族”加分政策录取利用率（截图）

通过以上的分析，利用加分政策录取利用率来评估一项加分政策是否合理的有效性是显而易见的。

### 3.4 评估方案小结

加分政策享受率是用来评估高考加分政策的整体规划合理性的，而加分政策录取利用率则是用来评估单个加分政策的合理性的有效手段。

**结束语** 本文的研究意义在于将数据仓库与 OLAP 技术应用于高考招生工作的加分政策。通过建立数据仓库和多维数据集，为高考加分政策分析提供准确的信息，更为后续的数据挖掘等分析提供准确、更具有针对性的数据。本文同时提供一套评估高考加分政策的合理性的方案，可为决策人员制定和调整高考加分政策提供依据。下一步的工作重点除完善数据仓库和 OLAP 模型，还将引入数据挖掘技术来分析高考加分政策的潜在规律等。

## 参 考 文 献


[1] 向莉娟，孟立军. 教育公平视阈下我国高考加分政策探析[J]. 内蒙古师范大学学报：教育科学版，2008，21(06)：1-3

[2] 万永生. 高考加分政策的现状及思考[J]. 教学与管理，2009，36(10)：75-77

[3] Inmon W H. 数据仓库[M]. 北京：机械工业出版社，2003：20-24

[4] 朱德利. SQL Server 2005 数据挖掘与商业智能完全解决方案[M]. 北京：电子工业出版社，2007：77-80

[5] Vincent R. Building a Data Warehouse: With Examples in SQL Server[M]. New York: Springer, 2008:71-72

[6] 陈文伟，黄金才. 数据仓库与数据挖掘[M]. 北京：人民邮电出版社，2004：54-55

[7] 吴远红. ETL 执行过程的优化研究[J]. 计算机科学，2007，34(1)：81-83

[8] Ralph K, Margy R. The Data Warehousr Toolkit: the Complete Guide to Dimensional Modeling[M]. Wiley, 2002:89-95

[9] Ralph K, Joe C. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleanin[M]. Wiley, 2004:29-48

[10] 陈京民. 数据仓库与数据挖掘技术[M]. 北京：电工工业出版社，2007：93-100

[11] 李红良. 智能决策支持系统的发展现状及应用展望[J]. 重庆工学院学报：自然科学版，2009，23(10)：140-144