一种基于内容和协同过滤同构化整合的推荐系统模型

李忠俊 周启海 帅青红
(西南财经大学经济信息工程学院 成都 610074)

摘 要 基于内容的推荐系统和协同过滤系统是最为流行的两种推荐系统,它们都有各自的优点和缺点。提出了一种基于对这两种推荐系统同构化整合的推荐模型,该算法同时拥有协同过滤推荐系统和基于内容推荐系统的优点,并且在一定程度上避免了基于内容或协同过滤的传统推荐系统各自的缺点。实验表明,该同构化整合模型与算法比传统的简单基本推荐模型、基于内容的推荐模型和协同过滤推荐模型提高了推荐的精确率。

关键词 同构化整合,基于内容,协同过滤,推荐系统模型

Recommender System Model Based on Isomorphic Integrated to Content-based and Collaborative Filtering

LI Zhong-jun ZHOU Qi-hai SHUAI Qing-hong
(School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu 610074, China)

Abstract The two recommender systems which are respectively based on content and collaborative filtering methods are most popular. Both types of filtering methods have advantages and disadvantages. This paper proposed a new isomorphic integrated model and algorithm which have the merits of the traditional recommender systems based on above two methods, and avoid the shortages of them to some extent. The experimental results show that the presented isomorphic integrated model and algorithm can improve the performance of the traditional recommender systems in predictive accuracy.

Keywords Isomorphic integrated, Content-based, Collaborative filtering, Recommender system model

1 引言

互联网规模和应用面的迅速增长逐渐产生了越来越严重的信息过载(information overload)问题。过量信息同时呈现,使得用户无法轻易从中获得对自己有用的部分(例如在中文谷歌中以"推荐系统"作为关键词进行搜索,即可获得超过1000万条查询结果)。现在很多网络应用,例如网址导航、搜索引擎、门户网站、专业数据库索引,本质上都是帮助用户过滤信息的工具或手段。然而,这些工具几乎都是只满足了主流的信息获取需求,没有个性化的考虑,依然无法很好地解决信息过载问题。推荐系统(recommender system)作为一种信息过滤的重要手段,是当前解决信息过载问题最有效的方法。因此,由于巨大的应用需求(Amazon. com, CDNOW, eBay, Levis, Moviefinder. com, Reel. com, Barnesandnoble 等商业网站均在其系统中部署了推荐功能模块[1]),推荐系统得到了极其广泛的关注。国内外许多学者研究推荐系统,美国计算机协会(ACM)还多次把推荐系统作为研究讨论的主题[2],大量的国内外期刊业纷纷将推荐系统作为研究专题[3],推荐系统的研究和应用呈现出一派"欣欣向荣"的景象。

目前,对推荐系统的分类没有统一的标准,很多学者都从不同的角度对推荐系统进行了不同的划分[4,5],但主流的趋势是将推荐系统分为以下几种:1)基于内容推荐[6];2)协同过滤推荐[7];3)基于知识推荐[8];4)基于数据挖掘推荐[9-11];5)整合推荐[8]。由于整合推荐是以前面4种推荐方法为基础的,因此一般认为,推荐系统的基本方法主要是基于内容推荐、协同过滤推荐、基于知识推荐和基于数据挖掘推荐4种。知识推荐不能自我学习,因此难以获得足够的知识进行模型的构建[8,14]。同时,基于数据挖掘的推荐存在着关联规则抽取难、耗时、产品名同义性问题、个性化程度低等问题[12],因此一般认为,协同过滤推荐和基于内容推荐是推荐系统的两类最基本的推荐算法。

以不同理论为基础构建起来的协同过滤推荐系统和基于内容推荐系统,分别有各自的优点和缺点。协同过滤推荐系统的优势主要体现在能够处理像音乐和视频等复杂的非结构化商品、自适应性好、推荐效果随着用户数量的增加而提高、不需要专业知识、容易发现用户新的兴趣、能获得用户充分的隐式反馈等方面;而基于内容推荐系统则有推荐结论直观且容易解释、不需要用户的历史交易数据、自适应性好、推荐效果随着交易数据的增加而提高、不存在新对象问题和数据稀疏问题、有成熟的分类技术的支撑、能获得用户充分的隐式反馈等优点。与此同时,协同过滤推荐系统存在着数据稀疏问题(sparsity problem)、不易测量(poor scalability)、新用户和新对象的引入问题、推荐质量依赖于大量的历史数据、"灰色绵羊"等问题;而基于内容推荐系统则有受商品属性提取方法

到稿日期:2009-05-03 返修日期:2009-06-19 本文受西南财经大学科研基金资助项目(08YJ18)资助。

李忠俊 讲师,主要研究方向为电子商务、网络营销、计算机应用,E-mail:webxml@tom. com;周启海 教授,博(硕)士生导师,主要研究方向为计算几何、算法研究与应用、财经计算、同构化信息处理等;帅青红 副教授,主要研究方向为电子商务、电子支付、计算机应用。

的限制(例如不易对音乐和视频信息进行提取)、新用户问题、分类训练需要大量的数据做支撑、不易测量等方面的缺陷。

尽管基于内容的推荐方法和协同过滤方法在诸多方面得到了广泛应用,但由于一些固有特点导致了一系列(如时效性差、准确度不高等)问题[13]。针对上述情况,本文综合基于内容和协同过滤两种算法,提出了一种新的整合推荐算法,以期获得较高的匹配准确度、较好的时效性和用户满意度。

2 基于内容和协同过滤的整合推荐算法

鉴于协同过滤推荐和基于内容推荐两种算法均有各自的优点和缺点,许多学者从不同角度提出了基于该两种方法的混合模型[15,16],他们声称其模型能够在充分利用协同推荐模型优点的同时还不会失去基于内容推荐算法的优点,并表明他们的模型能够比两种算法各自的表现更好。遵循类似的思路,本文提出一种新的基于内容推荐和协同过滤推荐的同构化整合推荐模型与算法。

2.1 模型的系统结构

结合基于内容推荐算法和协同过滤推荐算法,本文提出的算法在现有平台上整合与实施,其模型及结构如图 1 所示。

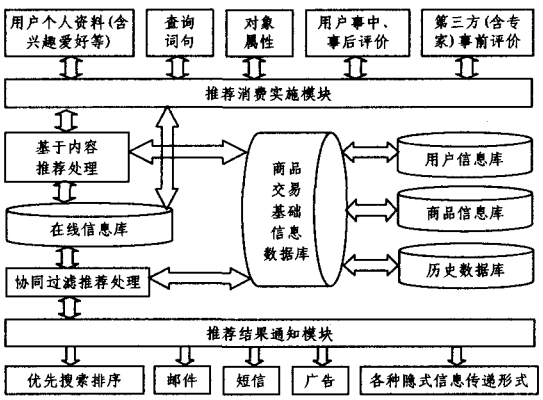

图 1 基于内容和协同过滤的同构化整合推荐系统结构

本模型与其他推荐模型的不同在于:1)构建了完善的模型输入模块。该模块包含可以根据需要向模型中一次或多次输入用户个人资料、查询(检索)时需要的关键词和句子、商品对象的各项属性值、用户使用推荐系统时和使用推荐系统后的评价,以及来自第三方对商品对象的评价等参数。2)构建了基于内容和协同过滤的整合推荐实施模块。该模块以回归模型为基本依据,以用户信息库、商品信息库、历史信息库和在线信息库为基础,分别将基于内容推荐的算法和协同过滤推荐算法应用于一般预测值的计算及其随机误差项的计算。3)构建了较完备的推荐结果输出模块。该模块依据系统的不同表现形式,将商品信息库中的资源"主动"推荐给用户。推荐方法包括搜索排序优先、邮件以及各种隐式信息传递等。

2.2 算法的主要步骤及核心

本算法的主要过程和核心是,先利用回归算法提取商品的内容属性,再对用户进行基于内容的协同过滤,最后同构化整合二者结论,并进行加权求和与有序化。

在详细阐述模型之前,有必要分析一下模型的输入数据。推荐系统典型的输入数据是给每个商品及其属性的评价。例如,表 1 是一个代表用户对电影优劣评价的 $m\times n$ 用户电影评价矩阵,该表的主要功能是在基于已经评价元素的基础上对空缺值进行预测。在评价矩阵中,用户 1 已经对电影 1、电影 2、电影 4、电影 5 做了评价,那么用户 1 会对电影 3、电影 6 做何评价呢? 同样地,用户 2 对电影 3,以及用户 3 对电影 4 和电影 5 该做何评价呢?

表 1 用户电影评价矩阵

	电影 1	电影 2	电影 3	电影 4	电影 5	电影 6	电影 7	电影 8
用户 1	5	1	?	4	2	?	?	
用户 2	2	4	?	1	5	2	?	
用户 3	4	2	3	?	?	5	5	
用户 4			3					
用户 5								

本整合推荐模型的算法由 6 个主要步骤组成。

首先,需要有一份存有所有商品(或对象)评价值(可能含有空缺值)的表格,初始表格中的评价值主要由用户直接或间接给出。例如,www. movielens. org 就是一个电影推荐网站,网站访问者只要注册并提供对 15 部以上的电影进行评价,那么他(她)就可以获得网站的推荐服务。像电影类别(例如动作、戏剧、爱情等)、导演、主要演员、制片商等关键信息都有助于对用户进行推荐。因此,用户在获得推荐服务之前应当尽可能向网站提供此类信息。

其次,一旦用户提交了推荐模型所需要的关键信息,那么对用户的电影推荐就可以应用式(1)来计算:

$$R_{i,j}=\alpha_{i,0}+\alpha_{i,1}X_{i,j,1}+\alpha_{i,2}X_{i,j,2}+\cdots+\alpha_{i,m}X_{i,j,m}+\varepsilon_{i,j} \quad (1)$$

其中,$R_{i,j}$是用户 i 对商品 j 的总体评价值,$X_{i,j,k}$是用户 i 对商品 j 的第 k 个属性的评价值,m 是需要评价的商品的属性总个数。

在回归模型(1)中,需要测量的各个参数 $\alpha_{i,x}(x\in[0,m])$ 决定了各个属性的数值对用户的重要性。用户过去的评价值都应用到模型(1)中,一旦用户 i 的模型中各个参数被估计出来,那么他(她)的空缺值就可以通过模型计算出来。例如,在表 1 中,用户 1 的回归模型通过他(她)已经看过的电影来构造,模型构造出来后,用户 1 对电影 3 的期望值就可以通过电影 3 和模型中的各个参数 $\alpha_{i,x}(x\in[0,m])$ 预测出来。

第三步,基于回归模型(1),计算适合所有用户和商品及其属性的预测值($\hat{R}_{i,j}$)。该预测值既适合实际已经发生的评价结果检验,也适合用户实际未发生或丢失的空缺值的预测。

第四步,构建一个预测误差的数据矩阵。该误差矩阵的值由实际发生的数据与预测数据之间的差值计算出来,即

$$\varepsilon_{i,j}=R_{i,j}-\hat{R}_{i,j} \quad (2)$$

式(2)中的 $\varepsilon_{i,j}$实际上就是模型(1)中的随机误差项。由于对于需要预测的商品没有实际发生评价值 $R_{i,j}$,因此该商品的随机误差项就无法计算。这样,通过实际的评价值和回归模型计算值之差构建的预测误差数据矩阵的形式也就和表 1 类似,与表 1 相对应的空白处也是没有数据的。

第五步,将协作过滤技术用于第四步构建的预测误差数据矩阵,我们在此采用基于邻居的协同推荐算法。为了计算用户的空缺值,采用式(3)来计算用户的随机误差项 $\varepsilon_{i,j}$:

$$R_{a,j}=\lambda\sum_{i=1}^{n}w_{a,i}(\varepsilon_{i,j}-\overline{\varepsilon_a})+\overline{\varepsilon_a} \quad (3)$$

其中,$R_{a,j}$是用户 a 对商品 j 的评价,n 是协同过滤数据矩阵中已经对商品 j 评价过的用户的数量,$W_{a,i}$是用户 i 和目标用户 a 之间的相似性,λ 是一个总和值为 1 的正太化因素。

在表 1 的电影评价例子中,假设对用户 1 而言,电影 3 是

需要通过预测来判断是否值得推荐的。在基于邻居的算法中，需要获得用户 2、用户 3 等其他用户对电影 3 评价的加权平均数。另外，权重($W_{a,i}$)的大小取决于用户 1 在电影评价方面与其他用户的相似性，包括皮尔森(Pearson)相关系数法、斯皮尔曼(Spearman)秩相关系数法和向量相似度等诸多方法均可以用来测定该值。在协同过滤的算法方面还有很多其他重要的文献，本文就不探讨了。

第六步，把第三步和第五步的计算结果求和，即第三步中利用基于内容推荐方法得出的$\hat{R}_{i,j}$和第五步基于协同过滤推荐方法得出的 $\varepsilon_{i,j}$，通过对式(2)进行变化并计算，得 $R_{i,j}=\hat{R}_{i,j}+\varepsilon_{i,j}$，即用户 i 对商品 j 的预测评价值。

3 实验结果与分析

为了充分验证整合推荐模型的有效性，本文既采用学术上通用的推荐系统评测数据集 MovieLens 进行检验，也将该系统用于现实中的某电子商务网站进行实际的检验。

3.1 基于 MovieLens 数据集的实验分析

为测试基于内容和协同过滤的整合推荐算法，所用的实验数据是 MovieLens[17]。在 MovieLens 规模不同的 3 个数据集中，用户对自己看过的电影进行评分，且每个用户至少对 20 部以上的电影进行过评分，分值为 1～5。评分越高，意味着用户对该部电影越赞同。

考虑到实验环境中计算机的运算能力，本文选择 MovieLens 的中等规模数据集(其中包含 6040 个独立用户对 3900 部电影做的约 100 万次评分)。在 MovieLens 的中等规模数据集中包含比较完整的用户信息(字段主要为年龄、性别、职业和邮政编号)、电影信息(字段主要为标题和电影所属类别)和用户对电影的评价信息。

随机地从数据集中取 100 个用户对电影评价的记录进行分析，检验本文提出的整合推荐算法能否得到较好的支持。这 100 个用户总共对 2318 部电影进行了 12976 次评价。为了对算法做出验证，本文将每个用户的 5%的电影评价数据作为有效性验证样本。为了更好地比较本文提出的推荐算法，作者选取了另外 3 种推荐算法作为参照。第一种算法是最基本的推荐算法，通过计算每部电影的用户评价算术平均数来获得。第二种算法是基于内容的推荐算法，这种算法以电影类别作为内容。这些类别包括动作、探险、卡通、儿童、喜剧、犯罪等 18 种[17]，并将这些类别分别作为一个待估计的变量。同时一部电影可以同时属于一个以上的类别。这 18 个种类变量的回归参数通过实际发生的用户评价值计算而得。然后，利用得到的回归函数计算验证样本中的评价值。第三种算法是以皮尔森相关系数法作为邻居算法的协同过滤算法。

表 2 显示的是基于分析样本的 4 种推荐模型的计算结果。每个模型均通过两个指标来评价，即接受者操作特性曲线灵敏度值(ROC)和绝对平均误差(MAE)。ROC 测量的是推荐算法的区分能力，值越大表示算法的效果越好。这个值通过计算接受者操作特性曲线下面的面积而得，其结果为 0 到 1 之间的实数。绝对平均误差 $MAE=\sum_{i=1}^{n}|A_i-\hat{A}_i|/n$(其中 A_i 为用户对电影 i 的实际评价值，$\hat{A}_i$ 为模型的估计值)用于测量模型的统计准确性。值越低，模型越准确。

表 2 各种推荐模型预测的准确性比较

模型类别	接受者操作特性曲线灵敏度值(ROC)	绝对平均误差(MAE)
基本模型	0.7448	0.2217
基于内容推荐模型	0.7765	0.2109
协同过滤推荐模型	0.8158	0.1935
整合推荐模型	0.8378	0.1821

与预料的情况类似，简单的平均法在接受者操作特性曲线灵敏度值和绝对平均误差两个指标上的表现都最差。另外，协同推荐算法的效果比基于内容推荐算法的效果更好。在实验数据集中，内容变量只有电影名称和电影类别两个。在其他场景中，随着更多重要内容变量的引入，其推荐效果可能会得到较大的改进和提升。同时，还可以得出的结论是，整合推荐算法在两个评价指标上都获得了最好的表现。

3.2 电子商务网站对推荐系统的应用分析

另外，本算法已在某电子商务系统中应用，并取得良好的效果。下面通过外部用户问卷调查和内部系统运营监测两部分实证并分析算法的效果。网站向用户以电子邮件/书面信函形式发送调查问卷 400 份，收回问卷 353 份(26 份为无效问卷)，有效问卷为 327 份，其中 91%的用户认为接收到的销售信息比过去更有效，85%的用户认为整合推荐算法的推荐结果比弹出式广告等形式更容易接受，75%的用户认为推荐能使他们更快地搜索到所需商品，58%的用户认为系统自动生成的推荐商品目录符合他们的需求。

图 2 给出的是 2009 年 7 月开始分 5 批对电子商务网站用户的在线调查结果。该图形显示的是每一个星期作为一批，并分别以基本模型、基于内容推荐模型、协同过滤推荐和整合推荐作为后台推荐算法，将一个星期的用户综合搜索时间求简单的均值计算得出的指数。由此可见，基于混合算法并以此为依据进行商品推荐，与单独基于内容和协同过滤算法相比更能综合提高用户有效信息获取的效率，进而能进一步提高用户对网站的满意度，并最终增加销售额。

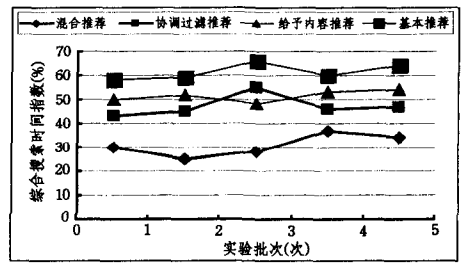

图 2 几种推荐算法的应用效果在线调查

以某工商档案图像应用系统为例，大量存在诸如图 2 所示的模板和图 1 所示的相似图像，我们采用 COT 方法来对其进行压缩。

结束语 本文提出了一种基于内容和协同过滤的同构化整合推荐模型。算法有效融合并改进了基于内容和协同过滤两种推荐算法，实现了对这两种算法的动态衔接。经 MovieLens 中随机抽样数据集的检验，在接受者操作特性曲线灵敏度值和绝对平均误差两个评价指标方面，整合推荐模型较单一的基于内容推荐算法和单一的协同过滤算法都表现出了更好的预测准确性。另外，在实际的某电子商务网站的应用中，其综合搜索时间与对比的其他模型相比，混合模型能大大降低有效获取信息的时间。

与此同时，需要看到的是，混合模型需要两种模型作为基础。因此，在模型应用时，必须同时以满足两种模型的应用条件为前提。例如，由于一般中小型电子商务购物网站很难获得用户的偏好信息，也不易得到用户对商品的评价，这就不能很好地应用本混合模型。这也是本模型以后需要继续完善和改进的地方。

参 考 文 献

[1] Ben Schafer J, Konstan J, Riedl J. Recommender systems in e-commerce[C]//Proc. of the 1st ACM Conf. on Electronic Commerce. New York: ACM press, 1999: 158-166

[2] Riedl J, Dourish P. Introduction to the Special Section on Recommender Systems[J]. ACM Transactions on Computer-Human Interaction, 2005, 12: 371-373

[3] Ricci F, Werthner H. Introduction to the special issue: Recommender systems[J]. International Journal of Electronic Commerce, 2006, 1(2): 5-9

[4] Balabanovic M, Shoham Y. Content-based, collaborative recommendation[J]. Communications of the ACM, 1997, 40(3): 66-72

[5] Terveen L, Hill W. Beyond recommender systems: Helping people help each other[C]//Carroll J, ed. HCI in The New Millennium. New York: Addison-Wesley Publishing Co., 2001: 1-21

[6] Han Peng, Xie Bo, Yang Fan. A Scalable P 2 P Recommender System Based on Distributed Collaborative Filtering[J]. Expert Systems with Applications, 2004, 27 (2): 203-210

[7] Min Sung-hwan, Han Ingoo. Detection of the Customer Time-variant Pattern for Improving Recommender Systems[J]. Expert Systems with Applications, 2005, 28 (2): 189-199

[8] 许海玲，吴潇，李晓东，等. 互联网推荐系统比较[J]. 软件学报，2009，20(2)：350-362

[9] Lazcorreta E, Botella F, Fernández-Caballero A. Towards personalized recommendation by two-step modified Apriori data mining algorithm[J]. Expert Systems with Applications, 2008, 35(3): 1422-1429

[10] Cao Yu-kun, Li Yun-feng. An Intelligent Fuzzy2Based Recommendation System for Consumer Electronic Products[J]. Expert Systems with Applications, 2007, 33 (1): 230-240

[11] Liu Duen-ren, Lai Chin-hui, Huang Chiu-wen. Document Recommendation for Knowledge Sharing in Personal Folder Environments[J]. Journal of Systems and Software, 2008, 81(8): 1377-1388

[12] 余力，刘鲁. 电子商务个性化推荐研究[J]. 计算机集成制造系统，2004，10(10)：1306-1313

[13] Yu Li, Liu Lu, LI Xue-feng. A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in E-Commerce[J]. Expert Systems with Applications, 2005, 28 (1): 67-77

[14] 黎星星，黄小琴，朱庆生. 电子商务推荐系统研究[J]. 计算机工程与科学，2004，26(5)：7-10

[15] Basu C, Hirsh H, Cohen W. et al. Recommendation as classification: Using social and content-based information in recommendation[C]//Proceedings of the 1998 Workshop on Recommender Systems. 1998: 43-52

[16] Ansari A, Essegaier S, Kohli R. Internet Recommendation Systems[J]. Journal of Marketing Research, 2000, 37: 363-375

[17] http://www.grouplens.org/node/73

[18] 周建硕. 供应链信息共享与利用的相关模式[J]. 重庆工学院学报：自然科学版，2009，23(2)：118-121

(上接第 118 页)

[12] Franconi E, Palma A L, Leone N, et al. Census data repair: a challenging application of disjunctive logic programming[C]// Proceedings of the Artificial Intelligence on Logic for Programming (LPAR). 2001

[13] Bravo L A B L. Logic programs for consistently querying data integration systems[C]//Proceedings of the International Joint Conference on Artificial Intelligence. 2003

[14] Cali A, Lembo D, Rosati R. On the decidability and complexity of query answering over inconsistent and incomplete databases [C]//Proceedings of the Symposium on Principles of Database Systems (PODS). 2003

[15] Cali A, Lembo D, Rosati R. Query rewriting and answering under constraints in data integration systems[C]//Proceedings of the International Joint Conference on Artificial Intelligence. 2003

[16] Chomicki J A M. Minimal-change integrity maintenance using tuple deletions[J]. Inform. Comput., 2005, 197(1/2): 90-121

[17] Greco G, Greco S, Zumpano E. A logical framework for querying and repairing inconsistent databases[J]. IEEE Trans. Knowl. Data Engin., 2003, 15(6): 1389-1408

[18] Wijsen J. Database repairing using updates[J]. ACM Trans. Datab. Syst., 2005, 30(3): 722-768

[19] Fan Wenfei, Jia Xibei. Anastasios Kementsietsidis, Conditional Functional Dependencies for Capturing Data Inconsistencies[J]. ACM Transactions on Database Systems (TODS), 2008, 33(2)

[20] Fan W. Dependencies Revisited for Improving Data Quality[C]//ACM Symposium on Principles of Database Systems (PODS) (invited). 2008

[21] Fan Wenfei. A Revival of Integrity Constraints for Data Cleaning[C]// The 34th International Conference on Very Large Data Bases (VLDB). tutorial, 2008

[22] Fan W. Extending Constraints with Conditions for Data Cleaning[C]// IEEE 8th Int'l Conf. on Computer and Information Technology (invited). 2008

[23] Bohannon P, et al. Conditional Functional Dependencies for Data Cleaning[C]// The 23rd International Conference on Database Engineering (ICDE) (the best paper award). 2007

[24] Fan Wenfei, Hu Yanli, Liu Jie, et al. Propagating Functional Dependencies with Conditions[C]// The 34th International Conference on Very Large Data Bases (VLDB). 2008

[25] 叶舟，王东. 基于规则引擎的数据清洗[J]. 计算机工程，2006，32 (23)：52-54

[26] 郭志懋，俞荣华，田增平，周傲英. 一个可扩展的数据清洗系统[J]. 计算机工程，2003，29(3)：95-96

[27] 谈子敬，施伯乐. 函数依赖和规范化在关系和 XML 间的传播[J]. 软件学报，2005，16(4)：533-539

[28] Fan Wenfei, Hu Yanli, Liu Jie, et al. Propagating Functional Dependencies with Conditions[C]// VLDB. Auckland, New Zealand, 2008

[29] Golab L, Korn F, Srivastava D, et al. On Generating Near-Optimal Tableaux for Conditional Functional Dependencies[C]// The 34th International Conference on Very Large Data Bases (VLDB). 2008

[30] Fan Wenfei, Jia Xibei. Semandaq: A Data Quality System Based on Conditional Functional Dependencies[C]//The 34th International Conference on Very Large Data Bases (VLDB), demo, 2008

[31] Bravo L, et al. Increasing the Expressivity of Conditional Functional Dependencies without Extra Complexity[C]// The 24th International Conference on Database Engineering (ICDE). 2008