文章编号:1671-8879(2014)05-0117-06

交通信息基础数据元中文名称短语相似度算法

张绍阳,关胜超,张 恒,李 欣
(长安大学 信息工程学院,陕西 西安 710064)

摘 要:交通信息基础数据元与用户数据项的中文名称短语的对应是数据元建立、标准符合性检测等工作的基础。为了提高名称对应的准确率,提出了一种利用数据元名称组成的特定结构进行数据项名称与数据元名称进行对应的方法,并给出了相似度的计算算法。该算法将用户数据项名称短语的省略情况按照中文语言习惯进行总结,采用数学中干扰修正的思想,分别按照语素和词素对相似度值进行计算,并利用相同语素的个数对相似度进行修正,综合得出词语的相似度。最后利用交通运输部实际工程数据进行了验证。研究结果表明:本算法较文献[1]中算法的"有改善"率提升了 91.20%,"明显改善"率提升了 9.62%;较文献[2]中的"有改善"率提升了 88.40%,"明显改善"率提升了 66.80%。

关键词:交通工程;交通信息数据元;中文短语;相似度算法
中图分类号:U411 **文献标志码**:A

A similarity algorithm for Chinese phrase of basic data element name structured for transportation information

ZHANG Shao-yang, GUAN Sheng-chao, ZHANG Heng, LI Xin
(School of Information Engineering, Chang'an University, Xi'an 710064, Shaanxi, China)

Abstract: The one-to-one correspondence of Chinese name phrase between the data element of transportation information and the data item name from the user is the foundation of data element establishment and standard compliance testing. In order to improve the correspondence accuracy of the names, this paper proposes a correspondence algorithm between data item name and standard data element name taking advantage of the special structure of the data element name, and also the computing method of the similarity. This algorithm, which classifies the phrase of data item name into some different situations on the basis of Chinese language convention, computes the similarity according to the morpheme and lexeme respectively, and corrects the similarity according to the number of the same morpheme. The phrase similarity is calculated synthetically in the end using the interference correction in mathematics. The data verification of the Ministry of Transport Engineering shows that this algorithm raises up the "good improvement" rate and the "evident improvement" rate up to 91.20% and 9.62% compared to the literature[1], and 88.40% and 66.80% compared to the literature[2]. 4 tabs, 2 figs, 10 refs.

Key words: traffic engineering; data element of transportation information; phrase of Chinese

收稿日期:2014-04-20
基金项目:交通运输部西部交通科技项目(2012 364 223 500)
作者简介:张绍阳(1971-),男,山西襄汾人,教授,工学博士,博士后,E-mail:zhsy@chd.edu.cn。

name; similarity algorithm

0 引 言

交通信息数据元标准和标准符合性检测是交通运输信息系统进行共享交换的基础性工作。在建立新的数据元时,首先需要对用户提出的新的数据元名称与已有的数据元名称进行比对,以此来防止出现重复的数据元;在进行数据标准符合性检测时,其第一步也是要将用户的数据项名称与数据元名称进行对应,才能根据标准判断用户数据项的采标情况。

数据元名称与用户数据项(元)名称对应的核心就是中文短语的相似度计算。在中文短语相似度计算方面,Levenshtein 提出一种编辑距离算法,通过替换、插入和删除等操作将 1 个字符串转变为另 1 个字符串,算法将所需要的最小编辑代价作为短语相似度的衡量[1];Li 等在基本编辑距离的基础上,赋予不同的编辑操作不同的权值,再计算相似度[2];刘群等基于《知网》提出一种基于语义分类体系下的词语相似度算法[3];胡俊峰等提出一种基于上下文词汇向量空间模型的词语相似度比对算法[4]。

以上文献都是针对通用的中文短语来计算相似度。由于交通运输行业信息数据元中文名称具有特定的组成结构和习惯用法,因此,本文在总结交通运输信息数据元名称省略习惯的基础上,根据人们理解信息的方式,提出了一种数据项与标准数据元名称对应的相似度算法。该算法极大提高了基于交通信息数据元结构的中文短语对应的准确率[5-8]。本文算法也可用于其他领域中数据元名称和数据项名称的对应。

1 交通信息基础数据元名称特性

1.1 交通信息基础数据元中文名称组成结构

根据文献[9]的规定,数据元中文名称的最小构成包括以下 3 个部分。

(1) 对象词:现实世界中的想法、抽象概念或事物的集合,有清楚的边界和含义,并且特性和其行为遵循同样的规则而能够加以标识。

(2) 特性词:对象类的所有个体所共有的某种性质。

(3) 表示词:值域、数据类型的组合,必要时也包括度量单位或字符集。

特性词可以进行限定,这些限定所用的词称为“限定词”[10]。表 1 中是几个数据元名称分解的例子。

表 1 交通运输信息基础数据元名称组成分解

Tab. 1 Structure of the name of basic data element for transportation information

序号	数据元名称	对象词	特性词	表示词
1	信息资源发布日期	信息资源	发布	日期
2	车辆下次技术等级评定日期	车辆	[下次]技术等级评定	日期
3	高速公路路况评定等级	高速公路	路况评定	等级
4	船舶必备检验文件代码	船舶	[必备]检测文件	代码

注:[]中的内容为限定词。

1.2 用户数据项名称定义习惯及分类

完整的用户数据项名称短语也应具有与数据元名称类似的组成,但由于语言习惯不同,或者用户假定读者具有先验知识,在用户数据项名称中,往往省略一些组成部分,本文根据多个交通运输信息系统的数据字典的用户习惯,将常见的省略情况总结如下页表 2 所示。表 2 中以“车辆下次技术等级评定日期”数据元名称为完整的名称,给出了各种省略情况下的用户数据项名称。

2 算法设计

2.1 算法思想

本文算法设计的目标是使表 2 中数据项名称的各种习惯用法都能与标准的数据元名称进行对应。基本编辑距离算法只考虑了 2 个中文短语中的相似字数及其顺序,即对语素的处理,而未能考虑词素对相似度判断的影响。在基于数据元名称结构中,词素和相同语素的个数对中文短语的相似度判断具有较大的影响。例如,在进行数据项名称“车辆技术等级评定”与目标数据元名称“车辆下次技术等级评定日期”对应时,如果只考虑对单个语素的处理时,需要进行 4 步插入操作,编辑距离较大,导致相似度较小;如果采用文献[2]中的“重心后移”加权处理方法,会导致返回的结果中以“评定”结尾的数据元名称为主,很难检索到目标数据元名称。

但是,按照人们理解信息的方式,更倾向于使用

整体来获得信息。从这个角度，相同的词语即使是顺序不同，也能够为相似度的判断提供更好的参考，人类大脑具有自动进行组合排序的能力。另外，2 个词语中相同的语素对相似度的贡献也很大。在上述例子中，2 个短语中的相同词素和语素如“车”、“等级”、“评定”、“日期”等为相似度的计算提供了信息，大多数短语的省略情况都具有类似的规律。基于以上思路，本文采用数学中的干扰修正思想，首先进行基于语素的相似度计算，然后采用相同语素和词素个数对相似度进行修正。

表 2　常见的用户数据项名称省略情况

Tab. 2　Catalog of name parts omitting of user data item

分类	数据元名称	用户习惯用法	省略内容	限定词	定义描述
情况 1	车辆下次技术等级评定日期	车辆技术评定日期	特性词不全	限定词缺失	对象词一致，特性词、表示词规范
情况 2		车辆技术等级评定	表示词不全		
情况 3		车辆技术等级评定日期	特性词、表示词均完整		
情况 4		车辆技术下次评定日期	特性词不全	限定词位置颠倒	
情况 5		下次车辆技术等级评定	表示词不全		
情况 6		汽车技术评级	表示词缺失	限定词缺失	对象词不一致，特性词、表示词不规范
情况 7		汽车技术评级日期	特性词、表示词均完整		
情况 8		汽车技术下次评级	表示词缺失	限定词位置颠倒	
情况 9		汽车技术下次评级日期	特性词、表示词均完整		

2.2　基于语素的相似度计算

语素即单个的中文字符，本文采用编辑距离算法对基于语素的短语相似度进行计算。编辑距离算法通过替换、插入、删除操作，将用户数据项名称转成目标数据元名称，最少的编辑次数即为二者的编辑距离。编辑距离越小，二者的相似度越大。其基本思想如下

设用户数据项名称为 $S[m]$，目标数据元名称为 $T[n]$，S 与 T 的编辑距离为 $M[m][n]$。$\boldsymbol{M}[i][j]=M[S_1,\cdots,S_i][T_1,\cdots,T_j]$，$0\leqslant i\leqslant m$，$0\leqslant j\leqslant n$，表示从 $S_1,\cdots,S_j$ 到 $T_1,\cdots,T_j$ 的编辑距离，那么 $(m+1)(n+1)$ 阶矩阵 $\boldsymbol{M}[i][j]$ 的值可通过式(1)计算得到

$$\boldsymbol{M}[i][j]\begin{cases} j & i=0 \\ i & j=0 \\ \min\begin{cases} \boldsymbol{M}[i-1][j]+1 \\ \min\begin{cases} \boldsymbol{M}[i][j-1]+1 \\ \boldsymbol{M}[i-1][j-1]+\text{cost} \end{cases} \end{cases} \end{cases} \tag{1}$$

式中：$\min[i][j]$ 为从 $S_1,\cdots,S_j$ 到 $T_1,\cdots,T_j$ 的最小编辑距离；$\text{cost}=(S(i-1)=T(j-1))\ ?\ 0:1$，为用户数据项名称 S 中第 i 个字符 S_i 转换到标准库中目标数据元名称 T 中第 j 个字符 T_j 所需要的操作次数，如果 $S_i=T_j$，则不需要任何操作，此时 $\text{cost}=0$；否则，需要替换操作，此时 $\text{cost}=1$。

定义距离调节因子 L。设 L_1 为用户数据项名称 S 的字符串长度，L_2 为标准库中目标数据元名称 T 的字符串长度。当 $L_1>L_2$ 时，$L=L_1$；当 $L_1<L_2$ 时，$L=L_2$。可以得到基于语素的相似度值 A，其计算公式为

$$A=(L-M[L_1][L_2])/L \tag{2}$$

2.3　相同语素个数对相似度的修正

相同语素个数对相似度的修正是通过计算用户数据项名称 S 与目标数据元名称 T 的相同语素个数，利用相同语素占 S 中的比例与相同语素的位置和系数(S 中在 T 中无相同的语素所在顺序号和的倒数)的乘积构造出相似度修正值。具体设计如下。

定义相同语素个数修正算法相似度值为 B，判断用户数据项名称 S 中各字符与目标数据元名称 T 中有无相同字符，若有 $S[i]$ 处字符与 $T[j]$ 字符相同，记录此时的 i 值，将所有相同语素的 i 值相加，记为 E，同时记录相同语素的个数为 n。将用户数据项名称 $S[i]$ 中的所有 i 值相加，记为 W。最终得到相同语素个数修正相似度值 B 的表达式

$$B=(n/L_1)(E/W) \tag{3}$$

B 值反映了相同语素个数对相似度的贡献，并同时考虑文献[2]中讨论的数据元名称理解时表示词的重要地位。算法实现流程见下页图 1。

2.4　基于词素的相似度修正

基于词素的相似度修正是通过计算用户数据项名称 S 与目标数据元名称 T 的相同词素所处位置，利用位置积倒数(T 中无相同的语素所在的顺序号乘积的倒数)构造出词素相似度修正值。算法思路如下。

定义相同词素修正算法相似度值 C，通过逐一字符比较，计算出用户数据项名称 $S[i]$ 与目标数据

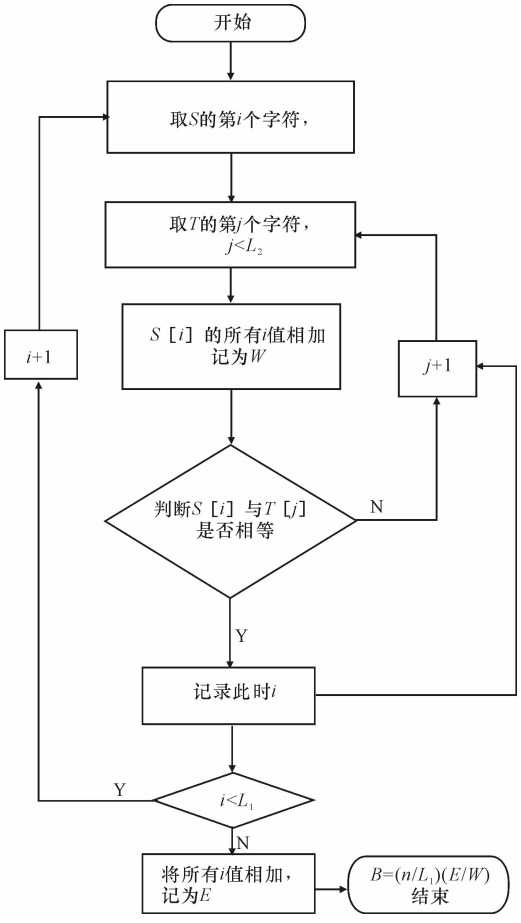

图 1 相同语素个数修正算法实现流程

Fig. 1 Flow for the correction algorithm of the number of the same morpheme

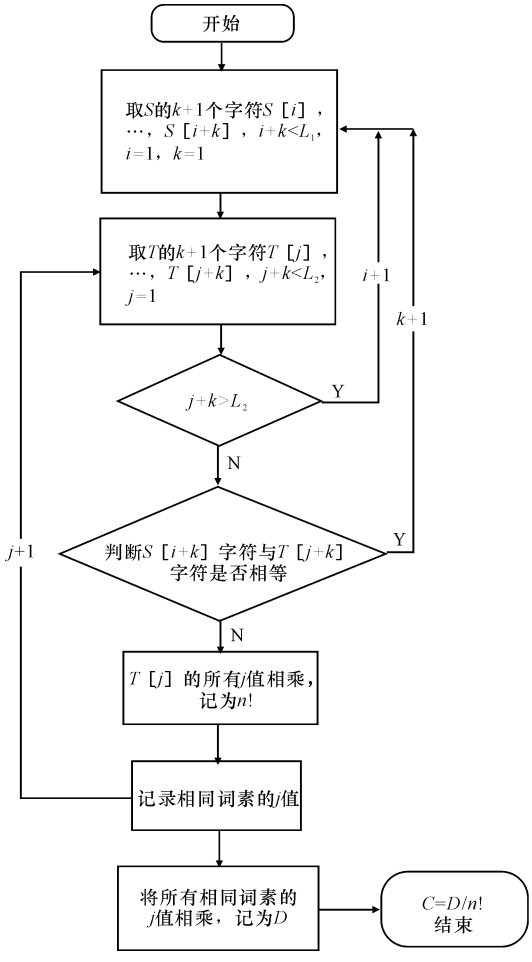

图 2 相同词素修正算法实现流程

Fig. 2 Flow for the correction algorithm of the same lexeme

元名称 $T[j]$的相同词素(词素必须由相邻的大于等于 2 个语素构成)所处位置 j 值,将计算出的 j 值相乘,记为 D。将目标数据元名称 $T[j]$的所有 j 值相乘,记为 $n!$。最终得到相同词素修正算法相似度值 C,计算式为

$$C=D/n! \tag{4}$$

相同词素反映了 2 个名称相似更多的可能性,但由于 B 值对相同语素已经进行了修正,因此此处 C 值使用位置积倒数进行计算,减少重复计算带来的影响。算法实现流程图如图 2 所示。

2.5 综合相似度计算算法

定义检测人员输入的用户数据项名称 S 与标准库中目标数据元名称 T 的相似度为 sim,sim 的表达式为

$$N_{\text{sim}}=\alpha A+\beta B+\gamma C,\ 0\leqslant \text{sim}\leqslant 1 \tag{5}$$

式中:N_{sim}为 sim 的计算参数符号;α、β、γ 为相关参数,且 $\alpha+\beta+\gamma=1$,它们的值由试验确定。

经过大量试验验证,当 $\alpha=0.6$,$\beta=0.3$,$\gamma=0.1$ 时,本相似度算法准确率最高。

2.6 算法计算量和复杂度分析

本文提出的算法在设计时为了提高准确度,较文献[1-2]中算法的计算量略高,但是算法的时间复杂度相同,仍为 $O(mn)$。并且由于标准库中所有的数据元名称和用户数据项名称的长度都为有限的常数(一般小于 50 个字符长度),所以对于现在的计算机性能来说,算法的计算量是可以接受的。

3 实例对比验证及分析

3.1 算法实例测试

为了进行算法对比,在“交通运输信息数据元的建立和维护工具”的开发中,分别应用文献[1]中的基本编辑距离算法(简称“算法 1”)、文献[2]中提出的数据名称与数据元名称对应方法(简称“算法 2”)与本文提出的中文短语相似度算法(简称“算法 3”),供用户进行重复数据元检索。文献[3-4]的相似度计算需要行业词典和语料库,限于条件,暂时无

法进行对比。

现将表 2 中用户数据项省略的 9 种情况，分别应用 3 种算法检索测试，测试结果见表 3。其中排序是人工判断的目标数据元名称在检索结果列表中的顺序。表中“0”表示在检索结果列表的前 100 个结果中没有出现目标数据元名称。

表 3　3 种算法对表 2 中的检索实例对比

Tab. 3　Comparison of the three algorithms retrieving the examples in table 2

类别	用户数据项名称测试实例	算法 1		算法 2		算法 3	
		排序	相似度	排序	相似度	排序	相似度
情况 1	车辆技术评定日期	12	0.50	0	0	2	0.90
情况 2	车辆技术等级评定	8	0.50	0	0	2	0.90
情况 3	车辆技术等级评定日期	2	0.80	3	0.87	2	0.90
情况 4	车辆技术下次评定日期	2	0.60	4	0.75	1	0.90
情况 5	车辆技术等级下次评定	9	0.40	49	0.53	1	0.90
情况 6	汽车技术评级	0	0	0	0	2	0.84
情况 7	汽车技术评级日期	0	0	0	0	2	0.86
情况 8	汽车技术下次评级	0	0	0	0	1	0.86
情况 9	汽车技术下次评级日期	48	0.30	5	0.72	1	0.87

通过表 3 的统计结果可以看出，算法 3 在表 2 中列举出来的几种情况下，所得到的结果在排序和相似度上都要优于算法 1 和算法 2，并且具有更高的相似度和相对靠前的排序。

3.2　试点工程应用及分析对比

标准符合性检测需要将用户数据项与标准数据元的中文名称进行对应。本文从交通运输部已开展的标准符合性检测试点工程中，随机抽取用户数据字典中定义的数据项名称 250 条，按照表 2 的分类方法进行分类，然后分别应用 3 种算法进行标准对应，将算法 3 分别与算法 1、算法 2 的对应效果进行对比，其结果统计如表 4 所示。

表 4　3 种算法在实际工程中应用效果对比

Tab. 4　Comparison of the three algorithms in the practical engineering

类别	数据项个数/个	算法 3 与算法 1 对应效果				算法 3 与算法 2 对应效果			
		明显改善	有改善	相似	变差	明显改善	有改善	相似	变差
情况 1	41	1	40	1	0	27	40	1	0
情况 2	49	2	45	4	0	48	48	1	0
情况 3	17	0	16	1	0	4	12	5	0
情况 4	10	1	10	0	0	5	10	0	0
情况 5	12	2	12	0	0	10	12	0	0
情况 6	42	8	39	3	0	39	42	0	0
情况 7	49	3	37	9	3	13	33	11	5
情况 8	13	4	13	0	0	13	13	0	0
情况 9	17	3	16	1	0	8	14	3	0

对 250 条用户数据项分别使用 3 种算法进行检索，在人为可接受的检索结果排序中和相似度值的范围内，分别对算法 3 与算法 1、算法 3 与算法 2 进行效果对比。当检索到目标数据元名称排在检索结果前 10 位，且利用算法 3 获得的排序与相似度值都高于另外一种算法或者排序相同但相似度值高时，视为算法 3 较另一种算法效果“有改善”；当利用算法 3 获得的排序与另外一种算法相同但相似度值低或者排序低相似度值高时，视为算法 3 较另一种算法效果“相似”；当利用算法 3 获得的排序低于另外一种算法并且相似度值也低时，视为算法 3 较另一种算法改善效果“变差”；其中有一种对比效果值要在此着重提出，即应用算法 3 检索到的目标数据元名称排在前 10 位，而应用另一种算法却在前 100 位未检索到目标数据元名称，视为算法 3 较另一种算法效果“明显改善”。

在表 4“算法 3 与算法 1 对应效果”中可以看出，算法 3 与算法 1 相比，在每种类别情况下“有改善”效果的比率都很高，在情况 1、2、3、4、5、6、8、9 中高达 90％以上，在 9 种情况中改善效果“相似”和

"变差"的比率都很低,尤其是"变差"率除了情况 7 外都为 0。特别值得提出的是,"明显改善"率在 8 种情况下都有出现。

在表 4"算法 3 与算法 2 对应效果"中可以看出,算法 3 与算法 2 相比,在各种情况下"有改善"效果的比率都很高,在情况 4、5、6、8 下高达 100%,情况 1、2 下达到 97%以上,在 9 种情况中改善效果"相似"和"变差"的比率都很低,尤其是"变差"率除了情况 7 外都为 0。而"明显改善"率更是在 6 种情况下达到 50%以上。

4 结 语

(1)在交通信息数据元建立和维护过程中,检索人员输入的数据项的情况会多种多样,应用本算法较 Levenshtein 的编辑距离算法效果"有改善"率提升了 91.20%,"明显改善"率提升了 9.62%;较文献[2]的数据名称与数据元名称对应方法"有改善"率提升了 88.40%,"明显改善"率提升了 66.80%。因此使得在数据元建立过程中,能够尽量避免重复。

(2)本算法在对数据项名称各构成部分模糊的情况下,检索的准确度还需要进一步提高。在以后的算法优化中,可以加入语义分析来弥补其不足。

参考文献:

References:

[1] Ristad E S, Yianilos P N. Learning string-edit distance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(5): 522-532

[2] Li N N, Zhang S, Zhao G. A corresponding method of traffic information data item name with data element name based on improved edit distance[C]//IEEE. Proceedings of Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Chongqing: IEEE, 2012: 2537-2540.

[3] 刘 群,李素建. 基于《知网》的词汇语义相似度计算[J]. 中文计算语言学,2002,7(2):59-76.
LIU Qun, LI Su-jian. Caculation of semantic similarity based on "How Net"[J]. Chinese Computational Linguistics, 2002, 7(2): 59-76. (in Chinese)

[4] 胡俊峰,俞士汶. 唐宋诗中词汇语义相似度的统计分析及应用[J]. 中文信息学报,2002,16(4):39-44.
HU Jun-feng, YU Shi-wen. Word meaning statistical analysis and in Chinese ancient poetry and its application[J]. Journal of Chinese Information, 2002, 16(4): 39-44. (in Chinese)

[5] 朱礼军,陶 兰,刘 慧. 领域本体中的概念相似度计算[J]. 华南理工大学学报:自然科学版,2004,32(增 1):147-150.
ZHU Li-jun, TAO Lan, LIU Hui. Caculation of the concept similarity in domain ontolagy[J]. Journal of South China University of Technology: Natural Science Edition, 2004, 32(S1): 147-150. (in Chinese)

[6] 黄 果,周竹荣. 基于领域本体的概念语义相似度计算研究[J]. 计算机工程与设计,2007,28(10):2460-2463.
HUANG Guo, ZHOU Zhu-rong. Research on domain ontology-based concept semantic similarity computation[J]. Computer Engineering and Design, 2007, 28(10): 2460-2463. (in Chinese)

[7] 张承立,陈剑波,齐开悦. 基于语义网的语义相似度算法改进[J]. 计算机工程与应用,2006,42(17):165-166,179.
ZHANG Cheng-li, CHEN Jian-bo, QI Kai-yue. Improvements on the arithmetic of semantic similarity based on the semantic wed[J]. Computer Engineering and Applications, 2006, 42(17): 165-166, 179. (in Chinese)

[8] 薛晔伟,沈钧毅,张 云. 一种编辑距离算法及其在网页搜索中的应用[J]. 西安交通大学学报,2009,42(12):1450-1454.
XUE Ye-wei, SHEN Jun-yi, ZHANG Yun. Modified edit distance algorithm and its application in wed search[J]. Journal of Xi'an Jiaotong University, 2009, 42(12): 1450-1454. (in Chinese)

[9] GB/T 18391.1—2009,信息技术-元数据注册系统(MDR)第 1 部分:框架[S].
GB/T 1839.1—2009, Information technology-metadata registries (MDR), part 1: Framework[J]. (in Chinese)

[10] 文必龙,付 玥. 数据集成中数据项与数据元匹配算法[J]. 计算机系统应用,2012,21(3):240-243,231.
WEN Bi-long, FU Yue. Matching algorithm between data items and data element during data integration [J]. Computer Systems & Applications, 2012, 21(3): 240-243, 231. (in Chinese)