网络出版时间：2016-06-22 14:28:56
网络出版地址：http://www.cnki.net/kcms/detail/51.1196.TP.20160622.1428.040.html

医疗体检数据预处理方法研究 *

林予松[1,2]，王培培[1,3]，刘 炜[1,2]，李润知[1]，王宗敏[1]

(1. 郑州大学互联网医疗与健康服务河南省协同创新中心，郑州 450052；2. 郑州大学 软件与应用科技学院，郑州 450002；3. 郑州大学 信息工程学院，郑州 450001)

摘 要：原始体检数据存在信息模糊、有噪声、不完整和冗余的问题，无法直接用于疾病的风险评估与预测。由于体检数据在结构和格式等方面的不足，不适合采用传统的数据预处理方法。为了充分挖掘体检数据中有价值的信息，从多角度提出了针对体检数据的预处理方法：通过基于压缩方法的数据归约，降低了体检数据预处理的时间及空间复杂度；通过基于分词和权值的字段匹配算法，完成了体检数据的清洗，解决了体检数据不一致的问题；通过基于线性函数的数据变换，实现了历年体检数据的一致性和连续性。实验结果表明，基于分词和权值的字段匹配算法，相对于传统算法具有更高的准确性。

关键词：体检数据；预处理；字段匹配算法；数据规约；数据清洗；数据变换

中图分类号：TP311

Research on preprocessing methods for medical examination data

Lin Yusong[1,2], Wang Peipei[1,3], Liu Wei[1,2], Li Runzhi[1], Wang Zongmin[1]

(1. Cooperative Innovation Center of Internet Healthcare, Zhengzhou University, Zhengzhou 450052, China; 2. School of Software & Applied Technology, Zhengzhou University, Zhengzhou 45000, China; 3. School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: The original physical examination data has many problems, including ambiguity, noise, incomplete and redundancy information, so it cannot be used for disease risk assessment and prediction directly. Traditional processing methods are not suitable for physical examination data because of its special structure and format. In order to solve these problems and make full use of the valuable information in the data, several methods are proposed in this paper. A compression-based data reduction method is used to reduce the time and space complexity of the data; a field matching algorithm based on segmentation and weights is used to complete the data cleaning and solve the problem of inconsistency; a data transformation method based on linear function is used to get the consistency and continuity of the history data. It is also proved that the filed matching algorithm in this paper is more accurate than the traditional method.

Key Words: Physical examination data; Data preprocessing; Field matching algorithm; Data reduction; Data cleaning; Data transformation

0 引言

目前，大多数的体检中心只对受检者提供当次的体检报告，缺乏对于受检者历史数据的分析。针对体检历史数据进行分析，可预测受检者患上某种慢病的风险和概率，提醒受检者及时发现潜在疾病，为其提供健康指导及疾病治疗措施[1]。由于历史原因，医疗机构现有体检数据库中存在着不少问题，如基本信息缺失、体检项目名称不统一、异常值较多、体检指标参考值范围不同、唯一标识码缺失、重复率高、参检项目不一致等等，这些问题均会影响体检数据慢性病风险评估的准确性，因此，在使用这些体检历史数据之前，还需要对数据进行预处理。

体检数据预处理的现有工作中，国内的朱玲、赵艳燕[2]等提出如何使体检中心采集标准化的体检数据；张茜、朱振昕[3],[4]等利用 Excel 软件中的 Len 函数进行数据剔除和汉字转换，并利用多重填补（Multiple Imputation，MI）的方法解决数据缺失的问题；李晓菲，戴齐[5]主要侧重解决数据的剔除及数据的缺失问

基金项目：河南省重点科技攻关项目（152102210249）

作者简介：林予松（1973-），男，四川西昌人，副主任，副教授，博士，主要研究方向为医疗信息工程（yslin@ha.edu.cn）；王培培（1988-），女，河南睢县人，硕士研究生，主要研究方向为互联网医疗；刘炜（1981 年-），男，河南安阳人，讲师，博士，主要研究方向为网络性能分析、网络安全；李润知（1978 年-），女，河南洛阳人，讲师，博士，主要研究方向为医疗大数据分析；王宗敏+，男（通信作者），河南郑州人，郑州大学互联网医疗与健康服务河南省协同创新中心主任，教授，博士，主要研究方向为互联网新技术，医疗信息工程.

题；骆芝英[6]提出体检数据存在的不规范问题，但只是从体检机构方面进行改进，并未针对历年体检数据进行预处理。国外的研究有采用 K-means 聚类算法降低数据预处理的时间及空间复杂度[7],[8]，且偏重西文的数据预处理。由于我国国情及中西文字本身的差异性，使得同一个匹配算法，并不能完全适用于所有语言的匹配问题，国外数据预处理技术不能完全应用于中文数据的预处理，且国内外体检形式不一样[9],[10]，获得的体检数据在结构和格式上亦存在差异。可见，现有的预处理方案不能有效解决体检数据预处理的问题，特别是针对中文的体检数据。

对此，本文从多个角度解决体检数据的预处理问题，主要工作包括：针对信息冗余的问题，提出基于压缩方法的数据归约；针对体检项目不统一、异常值较多的问题，提出基于分词和权值的字段匹配算法的数据清洗；针对唯一标识码缺失的问题，提出基于线性函数的数据变换等方法来完成健康体检数据的预处理，使最终结果能够直接用于慢性病风险评估以及个性化体检套餐设计。

1 体检数据预处理的步骤

数据预处理主要包括数据清洗、数据集成、数据变换及数据规约四个基本功能，实际的数据预处理过程可以使用全部或部分功能，而且某项预处理功能可以多次进行。

体检数据预处理方法的实施步骤如下：

步骤 1：减小数据量。采用一定的方法剔除无效数据。

步骤 2：排序。在关系表中以某一属性列为关键字进行排序，并使排序后的相似重复记录总是位于相邻的位置。

步骤 3：计算匹配度。通过算法计算相邻记录的匹配度。

步骤 4：阈值选择。采用一定的方法，选取阈值的最优解。

步骤 5：规范数据。利用步骤 4 中求取的阈值，清洗体检数据属性列；区分同一字段重名的体检者。

步骤 6：唯一标识码。利用某种方法将体检者的历年体检数据用唯一标识码表示。

通过以上步骤最终完成体检数据的预处理。

本文提出基于压缩方法的数据归约用于步骤 1；基于分词和权值的字段匹配算法的数据清洗用于步骤 3、4、5；基于线性函数的数据变换用于步骤 6。

2 基于压缩方法的数据归约

数据归约技术可以得到数据集的归约表示，在保持原数据的完整性的基础上，大大减少待处理的数据量，并能产生和规约前相同的分析结果。本文采用基于压缩方法的数据规约降低体检数据预处理的复杂度。

本文在体检数据的预处理的过程中用到数据压缩和维归约的思想。数据压缩是剔除数据库表中与体检结果无关的表，提取出体检者的基本信息表和体检数据表；维归约是将数据库表中与数据挖掘任务不相关的属性记录和冗余信息舍去，包括属性规约和记录规约。数据规约的目的是缩短数据清洗时间，降低体检数据预处理的复杂度。

3 基于分词和权值的数据清洗

数据清洗的作用是清洗噪声数据、无关数据、处理遗漏数据、填补空缺值、识别删除孤立点等。数据清洗的内容主要包括属性清洗和相似重复记录清洗。其中，“相似重复记录”是指一个现实实体在数据集合中用多条不完全相同的记录来表示，由于他们在格式、拼写上的差异，导致数据库管理系统不能正确识别。相似重复记录清洗主要包含两个步骤：检测重复记录和消除重复记录[11]，这两个问题都是数据清洗领域研究的重点，而检测相似重复记录则是数据清洗的核心。针对如何检测重复记录，本文在现有的字段匹配技术——简单的基于字符的字段匹配算法的基础上，提出基于分词和权值的字段匹配算法来完成体检单位数据的清洗，以解决体检数据不一致的问题。

3.1 传统的字段匹配算法

传统的基于字符的字段匹配算法主要用于获取两个英文字符串的匹配度，其核心思想是：首先，将每个字符串的单词抽取出来排序；其次，用一个字符串中的每一个单词到另一个字符串中进行搜寻匹配，并将匹配的单词个数记录下来[12]。由于英文与中文在字段匹配上有以下差别：英语中的缩写、简称、姓名、地址等与汉语都不一样；汉语中没有语法形态变化；汉语中词与词之间没有自然界限，即不像英文使用空格将单词分开。因此，传统的字段匹配算法无法满足体检数据预处理中对中文字段匹配的要求。针对该问题，本文提出一个基于分词和权值的字段匹配算法，提高匹配中文字段的准确度。

3.2 改进的字段匹配算法

改进算法解决的主要问题是在中文环境下获取两个字段的匹配度。首先描述相关定义。

3.2.1 匹配度

匹配度又名相似度，即指两个字段值可转换的程度，匹配度越高，表明两个字段越相似。以下引入分词匹配度和权值匹配度。

1)分词匹配度

分词匹配度：两个分词串中匹配的分词个数除以他们所有分词总数的平均值。如果两个分词相同，就认为他们匹配。如公式(1)所示。

$$ppd = \frac{2K}{|A| + |B|} \tag{1}$$

其中，K 表示待匹配的两个分词串中相同词个数，$|A|$表示字段 1 中分词个数，$|B|$表示字段 2 中分词个数。

2)权值匹配度

权值匹配度计算的数学模型可用集合来表示：A 为待匹配词集合，B 为匹配词集合，C 为两词中相同字的集合，匹配度即为 C 占 A 和 B 全部的比重，则 A 与 B 的匹配度如公式(2)所示：

$$ppd = \frac{2s_C}{s_A + s_B} \tag{2}$$

其中 C 占 A 的比重为 x，C 占 B 的比重为 y，换言之，x 为 C 在 A 中权值之和，y 为 C 在 B 中权值之和，其中 x，y 如公式

(3)所示：

$$x = \frac{s_C}{s_A},\ y = \frac{s_C}{s_B} \tag{3}$$

由公式(2)和公式(3)可得匹配度如公式(4)所示：

$$\frac{1}{ppd} = \frac{1}{2x} + \frac{1}{2y} \tag{4}$$

其中 x，y 不为 0。而当 x 或 y 为 0 时，没有交集，相似度为 0。

至此，求取字在字段中的权值，方法如下：

假设某一字段有 N 个字组成，各个字的权值 $\omega(k)$由字 k 在词中的位置确定，如公式(5)所示，首字为 1，以此类推。

$$\omega(\mathrm{k}) = \frac{K}{1 + 2 + \ldots + N} \tag{5}$$

其中，K 表示该字在词汇中所处的位置，词汇的权值如公式(6)所示：

$$\omega(\mathrm{N}) = \sum_{k=1}^{N} \omega(\mathrm{k}) = 1 \tag{6}$$

例如：利用权值求取“郑大”与“郑州大学”的权值匹配度：$x=1$，$y=2/5$，$ppd=57.14\%$；利用中文分词求“郑|大|”和“郑州大学|”之间的分词匹配度：$K=0$，$|A|=2$，$|B|=1$，$ppd=0.00\%$，“|”在这里定义为切分符号。

3.2.2 基于分词和权值的字段匹配算法

基于分词和权值的字段匹配算法通过计算两个字段之间的匹配度来检测相似重复记录，首先通过分词匹配度判断，再利用权值匹配度判断，利用两个匹配度来提高检测重复记录的准确性。

同样以某三甲医院近五年约 80 万人次的体检数据为例，该算法运行流程如图 1 所示：

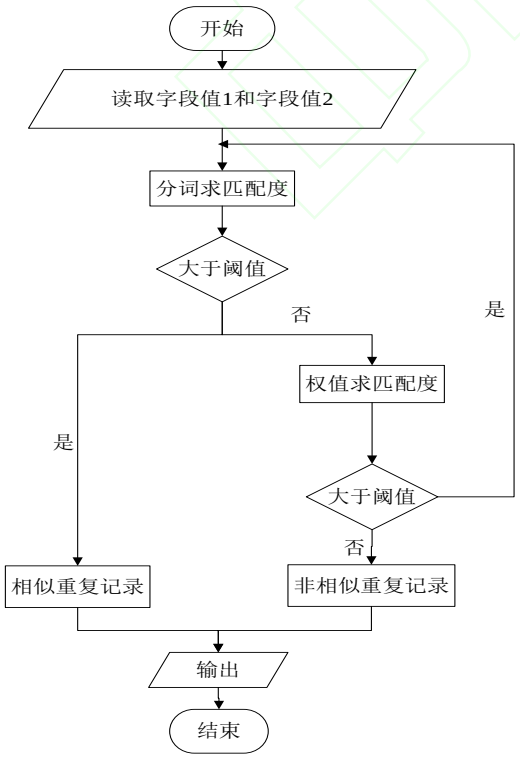

图 1 算法流程图

1）从体检数据库中读取待处理的字段值 1 和字段值 2，使两个字段值不重复，以减少处理的复杂度。

2）利用分词器对体检单位数据进行分词处理，将汉字字符串变为分词串，求取分词匹配度。

本文根据算法的开发语言并结合现有的中文分词器的特点，选用盘古分词器[13]提供的接口完成分词。盘古分词是一个基于.NET 平台的开源中文分词组件，提供 Lucene(.net)和 HubbleDotNet 的接口[14]，采用字典和统计结合的分词算法，提高分词的准确性。由于盘古分词器的字典所包含的数据不全，导致个别分词结果不准确，而体检单位的名称是固定的，因此，可以有针对性的扩展字典中的数据，来进一步提高分词的准确性。

3）利用重心后移规律匹配法求取权值匹配度。

该方法是根据主题重心词通常在词的后半部分的特点，选取自然语言同义词处理技术中重心后移规律匹配法，即根据汉语构词重心后移的特点：字越靠前，作用越小，权值越小[15]。对词语中的字加权处理进行量化，然后计算匹配度，衡量词语的可替换程度[16]。体检数据库中体检单位使用的各种简称也是固定的，绝大多数的同义词含有相同的语素（字），因此，使用此方法可以解决分词过程中出现的问题。

4）结合 2）和 3）中求取分词匹配度和权值匹配度，通过判断是否大于某一阈值来决定是否为相似重复记录。

5）清除相似重复记录，完成体检数据的清洗。

综上所述，改进的字段匹配算法将分词和权值结合，互相取长补短，实现相似重复记录的准确检测，最终完成体检数据的清洗。

4　基于线性函数的数据变换

经过数据规约和数据清洗的数据不能直接进行数据挖掘，必须通过数据变换将体检数据转换为适合数据挖掘的形式。本文以某三甲医院近五年约 80 万人次的体检数据为数据集，提出一种基于线性函数的数据变换方法，将体检数据用唯一标识码结合时间轴的形式表示，实现体检数据的一致性和连续性。

该数据集中，基本体检数据信息中并未录入体检者的身份证号。对于名字相同的体检者，系统无法识别是否是同一个人，因此需要经过一系列的数据清洗过程，找到每条体检数据的关联关系，为每个人增加唯一的查询标识（唯一标识码），最终通过姓名和体检单位两属性确定哪些记录是属于同一个人的。为了让计算机可以识别出基本属性并能直接用来数据挖掘，需要对数据进行变换。以下步骤采用函数来实现体检数据的变换：

步骤 1：生成标识列。令抛物线函数 $Z=AX+BY+C$，其中，设姓名列为 X，体检单位列为 Y，生成标识列为 Z，A，B，C 均为常数；

步骤 2：规范标识列。利用 SQL 语句函数 group by 将标识列 Z 排序分组，并用函数 rank() over 给每个分组添加序号，此时文字标识列转换为数字标识列；

步骤 3：统一标识列。为使标识的每个字段值有相同位数，

新增加一列唯一标识 *W*，由于基本信息近 16 万人次，因此设定函数 *W=Z+1000000*，可将每个体检者的唯一标识码转换为 7 位数的数字。

步骤 4：可视化。利用数据库系统存储过程中的内置函数 patindex()可去掉其前后非数字字符，转换后的数值可以直接进行数据挖掘或者进行数值指标的可视化，体检者可以直观的看出自己历年某一指标的折线趋势图。

通过以上 4 个步骤，生成了体检者的唯一标识码，实现了体检数据用唯一标识码结合时间轴的形式表示。

5　实验评估

本文所涉及的预处理方法，采用 *C#*语言和 SQL Server 2008 数据库进行了实现。所用数据集来源于某三甲医院最近五年约 80 万人次的体检数据库，该数据库主要包括体检者基本信息表和体检数据表。

5.1　数据规约方法的实验评估

5.1.1 实验目的及方法

目的：检验体检数据经过数据规约能否减少不相关及冗余信息。

方法：以某三甲医院近五年约 80 万人次的体检数据为例，其中，四万左右人次的体检者仅有基本信息而缺乏体检数据，因此在不影响数据挖掘结果的前提下需要将这四万左右无效数据剔除。由于体检者的基本信息和体检数据是靠流水号关联的，可利用相差法剔除有基本信息而没有体检数据的记录。原始体检数据库中含有 11 个表，经过筛查选择 4 个基本表（基本信息表和体检数据表）。

5.1.2 实验结果及分析

原始体检数据大小：5391M ，经数据规约方法处理后的数据大小：2585M。

通过处理前后对比，数据量减少了 52%，验证了基于压缩方法的数据规约可大大减少体检数据的不相关及冗余信息。

5.2　数据清洗方法的实验评估

5.2.1 实验目的及方法

目的：验证本文提出的基于分词和权值的字段匹配算法的准确性是否高于传统的基于字符的字段匹配算法。

方法：首先，以基本信息表的“体检类型”为关键字进行非重复排序；然后，分别采用改进算法和传统算法对检索结果的相邻记录进行匹配度计算，并根据阈值选取算法选出的 *K* 值来判断，匹配度大于 *K* 的即为相似重复记录；最后，消除相似重复记录。

本文采用召回率、正确率和 *F*-测度值三个指标对基于分词和权值的字段匹配算法的优劣进行评判。召回率 *r* 反映了被正确判定为正例数占总正例数的比率；正确率 *p* 反映了被正确判定为正例数占被判定为正例数的比率；*F*-测度值 *f* 综合了召回率和正确率的结果，用于综合反映整体指标。具体如公式（7）所示：

$$\begin{aligned} r &= a/(a+c)\times 100\% \\ p &= a/(a+b)\times 100\% \\ f &= 2p\times r/(p+r)\times 100\% \end{aligned} \qquad (7)$$

其中，*a* 被正确判定为正例的个数；*b* 表示被错误判定为正例的个数；*c* 表示将正例排除在外的个数。

5.2.2 阈值选取

本文所提算法的准确程度取决于阈值 *K*。为了得到优化的 *K* 值，用分词匹配度算法求相邻记录的匹配度，根据匹配度值的特征，本文从 0%到 100%以 10%间隔递增设定 *K* 值，以检测相似重复记录的正确程度来计算不同 K 值对应的召回率、正确率和 *F*-测度值，变化趋势如图 2 所示：

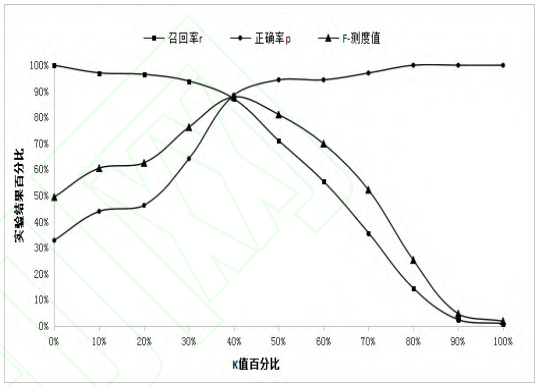

图 2 不同 K 值下的召回率、正确率和 *F*-测度值

根据实验结果可知：随着 *K* 值增大，召回率呈下降趋势，正确率呈上升趋势，*F*-测度值则呈先上升后下降的趋势。由于 *F*-测度值可很好的衡量召回率和正确率，因此，*F*-测度值最大时，可得到优化的 *K* 值，此时 *K=40%*。

5.2.3 实验结果及分析

用基于分词和权值的字段匹配算法和传统算法求相邻记录的匹配度，分别计算阈值 *K* 为 40%时对应的召回率、正确率和 *F*-测度值。两种算法的对比结果如表 1 所示。

表 1 算法的对比实验结果

字段匹配算法	召回率	正确率	F-测度值
传统字段匹配算法	87.18%	88.48%	87.82%
改进后字段匹配算法	93.41%	93.92%	93.66%

从实验结果指标可知，基于分词和权值的字段匹配算法在召回率、正确率和 *F*-测度值上分别高于传统算法 6.23%、5.44%和 5.84%。基于分词和权值的字段匹配算法在计算相邻记录匹配度上准确性更高，可以更好的完成体检数据的清洗工作——检测、消除相似重复记录。

5.3　数据变换方法的实验评估

5.3.1 实验目的

目的：检验体检者经过数据变换能否成功添加唯一标识码。

5.3.2 实验方法及结果分析

方法：用基于线性和内置函数的数据变换将体检数据的形式化表示转换为数字化及图形化，使体检者历年体检数据以唯一标识码结合时间轴的形式表示。具体结果如图 3 所示：

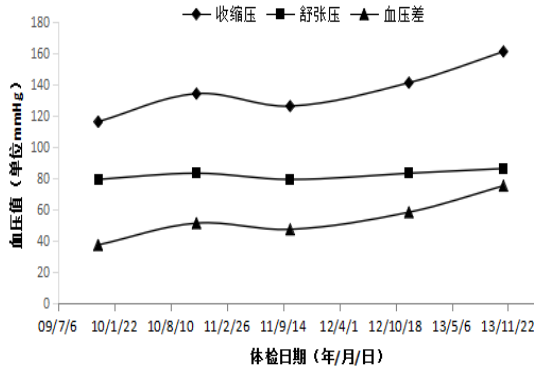

图 3 血压指标折线图

从处理后的结果可直观看出：可查看体检者历年体检数据指标的变化趋势，验证了经过基于线性函数的数据变换成功的给体检者添加唯一标识码。

综上所述，数据规约是数据清洗的基础，而数据变换是数据清洗结果的呈现。三者互补，共同完成体检数据预处理工作。

6 结束语

体检数据的预处理可以改善数据质量，提高数据挖掘的效率，帮助体检者查看历年体检信息及对应的体检项目指标，了解各指标的变化趋势。本文从数据规约、数据清洗和数据变换三个方面来实现体检数据的预处理：根据基于压缩方法的数据规约降低了算法的时间和空间复杂度；提出了基于分词和权值的字段匹配算法，提高了数据清洗的准确率；根据基于线性函数的数据变换将体检者的历年体检数据用唯一标识码结合时间轴的形式表示。

为了更好的完成体检数据的预处理，可以进一步开展两方面的工作：一方面加强体检机构的数据标准化，提高体检数据的准确性和规范性；另一方面，进一步优化本文提出的基于分词和权值的字段匹配算法，目前的算法计算权值时采用重心后移法，下一步的工作需综合考虑部分重心词前移的情况，进一步提高算法的准确性。

参考文献

[1] 徐宾, 时利群. 健康体检对早期预防和发现疾病的重要性分析[J]. 现代预防医学, 2012, **39**(19): 5033-5034.

[2] 朱玲, 赵艳燕, 马兰军, 等. 健康体检数据整理分析思路及实践体会[J]. 中华健康管理学杂志, 2010, **4**(5): 257-259.

[3] 张茜, 朱振昕, 孟文佳, 等. 大型纵向监测健康体检数据的统计分析策略[C]//中国卫生统计学年会会议论文集. 2011.

[4] 张茜. 大型纵向监测健康管理队列设计及其统计分析策略研究[D]. 济南: 山东大学, 2013. 05.

[5] 李晓菲. 数据预处理算法的研究与应用[D]. 成都: 西南交通大学, 2006.

[6] 骆芝英, 陈晓洁, 俞春晓, 等. 体检机构检后信息档案存储和运行存在的问题和对策[J]. 中华健康管理学杂志, 2014, **8**(5): 359-360.

[7] Nouri V, Akbarzadeh T, Rowhanimanesh M R. A hybrid type-2 fuzzy clustering technique for input data preprocessing of classification algorithms[C]//Proc of IEEE International Conference on Fuzzy Systems. 2014: 1131-1138.

[8] Sreenivas P Srikrishna C V. An analytical approach for data preprocessing[C]//Proc of International Conference on Emerging Trends in Communication, Control, Signal Processing and Computing Applications. 2013: 1-12.

[9] 王红玉. 国内外健康体检的发展现状及分析[J]. 中国疗养医学, 2014(3): 214-215.

[10] 赵楠, 梁英, 谭志军, 等. 中外健康体检项目比较分析[J]. 现代生物医学进展, 2013, **13**(11): 2135-2141.

[11] 王曰芬, 章成志, 张蓓蓓, 等. 数据清洗研究综述[J]. 现代图书情报技术, 2007(12): 50-56.

[12] 陈挺, 郭颖, 刘云超. 中文字段匹配算法[J]. 计算机工程, 2003, **29**(13): 119.

[13] Eaglet. 盘古分词-开源中文分词组件[EB/OL]. (2010-08). http://pangusegment. codeplex. com/.

[14] 马军, 杨维明, 周民. 采用 Lucene. Net 与盘古分词器的网上书城站内搜索方法[J]. 电脑知识与技术, 2015, **11**(20): 184-187.

[15] 陆勇, 章成志, 侯汉清. 基于百科资源的多策略中文同义词自动抽取研究[J]. 中国图书馆学报. 2010, **36**(1): 56-62.

[16] 崔韬世, 麦范金. 词语相似度计算方法分析[J]. 网络安全技术及应用, 2012(5): 55-56.