位置信息记录中基于期望最大化的名称消重算法

孙晓玲[1] 郑 勉[1] 李伟勤[1] 罗恩韬[2]
(西南石油大学电气信息学院 成都 610500)[1] (中南大学信息科学与工程学院 长沙 410083)[2]

摘 要 在包含位置信息的签到记录中,每条记录仅包含名称和位置(经纬度)两个属性。传统的名称消重算法通过匹配实体的属性值或者计算实体间的名称相似性进行消重,忽略了位置信息的特殊性。为了提高位置信息记录中名称消重的质量,提出了一种基于期望最大化的位置名称消重算法。首先,提出了一种包含核心单词和背景单词的文本名称模型,并给出了计算模型参数值的期望最大化算法。其次,在文本名称模型中引入位置信息,将整个地图划分为若干个网格,分别计算每个网格中核心单词和背景单词的分布情况,并提出了一种考虑位置的文本名称模型。最后,将位置文本名称模型用于位置信息记录中的名称消重,并给出了相应的名称消重算法。实验表明,与传统的名称消重模型相比,提出的位置名称消重模型可以更好地识别出名称中包含的核心词汇,因而在名称消重时具有更好的性能。

关键词 签到,位置,期望最大化,名称消重
中图法分类号 TP319 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.3.043

Expectation Maximization Based Name Deduplicating Algorithm in Spatial Records

SUN Xiao-ling[1] ZHENG Mian[1] LI Wei-qin[1] LUO En-tao[2]
(School of Electrical Engineering and Information, Southwest Petroleum University, Chengdu 610500, China)[1]
(School of Information Science and Engineering, Central South University, Changsha 410083, China)[2]

Abstract In check-in records with corresponding locations, each record only contains the attributes of name and location, i.e., longitude and latitude. Traditional name deduplicating algorithms deduplicate names by matching attributes between two entities or computing similarity between names of the two entities, and thus neglect the particularity of locations. In order to improve the quality of name deduplicating in spatial records, this paper proposed an expectation maximization based name deduplicating algorithm. Firstly, we proposed a text name model containing core and background words, and gave an expectation maximization algorithm for computing parameters of the model. Secondly, we introduced location into the text name model, partitioned the whole world into tiles, computed the distributions of core and background words in each tile, and proposed a text name model including location. Finally, we used the location text name model to deduplicate names in location records, and presented corresponding name deduplicating algorithm. The experiments show that, our proposed algorithm can better recognize core word in a name than related works, and thus performs better while deduplicating name in location records.

Keywords Check-in, Location, Expectation maximization, Name deduplicating

在数据库中,由于缩写、重名、用户的操作失误以及噪音等因素的影响,同一实体可能由多个名称来表示,而这些表示同一实体的多个记录称为重复记录[1]。在数据的清洗与集成[2]中,重复记录的识别与删除是至关重要的一步。对于重复记录的识别,研究人员通常对每一对记录进行评分,计算该记录对相互匹配的概率。由于每条记录可能有多个属性,因此记录对的匹配主要依赖于对应属性值的匹配[3]。

在基于位置的社交网络[4,5]中,每一条记录都对应着一次签到(check-in)行为,该签到行为表明了某个用户在某个时间访问了某个地点。在用户的签到行为中,签到地点的名称由用户手动输入,签到地点的地理位置采用GPS或者网络定位[6]得到相应的经/纬度坐标。由于地点的重名、名称的缩写表示、用户输入错误的数据以及GPS定位自身存在的误差等原因,签到数据中存在着大量的重复记录。图1示出Facebook签到数据[7]中所有包含"Stanford"的位置及其在谷歌地图上的分布情况。

在传统的重复记录识别中,每条记录含有多个属性,因而识别相同的记录很容易;在签到数据中,每条记录仅含有名称和位置两个属性,因而很难识别出重复记录。如果两条记录相隔的距离很远,那么这两条记录几乎不可能指向同一个地

到稿日期:2015-04-07 返修日期:2015-06-11 本文受国家重大专项资助项目(2008zx05026-001-09),国家自然科学基金项目(6140060035),四川省教育厅自然科学一般项目(16ZB0074)资助。

孙晓玲(1978-),女,硕士,讲师,主要研究方向为数据挖掘、嵌入式技术与信息处理;郑 勉(1980-),女,硕士,讲师,主要研究方向为通信技术与信息处理;李伟勤(1976-),男,博士生,副教授,主要研究方向为无线通信技术;罗恩韬(1978-),男,博士,副教授,主要研究方向为大数据、大数据安全、大数据汇聚计算、多维分析。

点；如果两条记录的地理位置很近，那么需要进一步对记录的名称进行匹配。在个性化字符串匹配中，现有的研究工作主要分为标识符加权模式[8-10]和基于学习的编辑距离[11-14]两类。

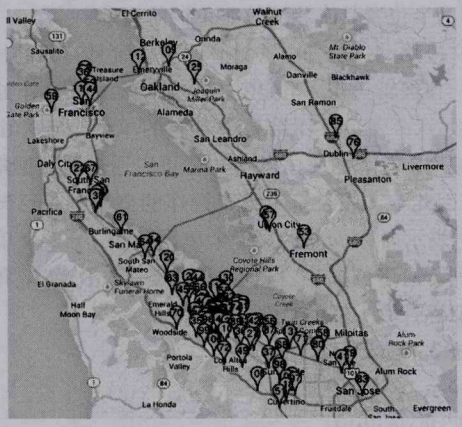

图1 Facebook中包含“Stanford”的签到信息在谷歌地图上的分布

本文研究了签到记录中包含名称和位置(经纬度)两个属性的名称消重方法。对于实体的名称，本文假设名称包含核心单词和背景单词，并且它们服从不同的分布；对于实体的位置，本文将整个地图划分为若干个网格，分析每个网格内核心单词和背景单词的分布情况。本文基于期望最大化方法提出了考虑位置信息的文本名称模型，并应用该模型进行位置记录的名称消重。

1 空间上下文的名称模型

本节首先提出了一种只包含名称的地点匹配模型；然后在该模型中进一步引入位置网格；最后应用提出的模型对签到记录进行匹配，从而实现记录的位置消重。

1.1 文本名称模型

词汇表 W 为所有单词的集合，每个地点的名称包含若干个词汇 $n\subset W$，所有地点的名称集合 $N=\{n|n\subset W\}$。给定地点的名称 n，其核心词汇集合记为 $core(n)$，背景词汇记为 $back(n)$。例如，当 $n=$“Guggenheim Art Museum, Manhattan”时，核心词汇 $core(n)=\{Guggenheim\}$标识了该地点的地理位置，背景词汇 $back(n)=\{$Art, Museum, Manhattan$\}$描述了该地点的位置属性。

对于给定地点的名称，需识别出名称中的核心和背景词汇。为此本文提出如下名称产生模型：令 B 和 C 为词汇表 W 上的两个概率分布，首先依据概率分布 C 从 W 中选取一个单词；其次依据某个固定的分布生成一个整数 $k\geqslant 0$；最后基于概率分布 B 从 W 中选取 k 个单词。核心词汇和背景词汇共同构成了地点的最后名称，背景词汇的个数与数据集中名称的长度分布有关。

给定名称集合 N，通过对 N 的学习达到如下两个目标：a)学习 B 和 C 的概率分布；b)对于每个名称 $n\in N$ 和单词 $w\in n$，学习 w 为 n 的核心词汇的概率。用符号函数 $z(w,n)$ 表示 w 是否为 n 的核心词汇，当 $core(n)=w$ 时 $z(w,n)=1$，否则 $z(w,n)=0$。用 z 表示所有符号函数的集合，那么观测到给定数据 N 的概率为：

$$\Pr(N|B,C,z)=\prod_{n\in N}\prod_{w\in n}C(w)^{z(w,n)}\cdot B(w)^{(1-z(w,n))} \tag{1}$$

通过对式(1)取对数，可以得到如下参数模型：

$$L(B,C,z)=\sum_{n\in N}\sum_{w\in n}z(w,n)\log C(w)+(1-z(w,n))\log B(w) \tag{2}$$

通过对名称集合 N 进行学习，希望得到 $\Pr(N|B,C,z)$ 在取最大值时的参数 B、C 和 z，并令之成为模型的参数。如果 N 中每个名称的核心单词都是已知的(即 z 已知)，那么可以很容易计算出 B 和 C 的估计值。当 z 未知时，可以通过期望最大化(Expectation-Maximization, EM)方法[16]计算上述优化问题。

EM方法主要包含E和M两步，并且循环迭代。在E步骤中，第 t 次迭代的参数值 $B^{(t)}$ 和 $C^{(t)}$ 是已知的，根据 $B^{(t)}$ 和 $C^{(t)}$ 计算 z 在所有取值下的期望值 $L(B,C,z)$。由于 $L(B,C,z)$ 是 z 的线性组合，根据期望的线性性可以得到如下条件期望计算公式：

$$\begin{aligned}E(z(w,n)|B^{(t)},C^{(t)})&=\Pr(z(w,n)=1)\\&=\frac{C^{(t)}(w)\prod_{x\in n\setminus w}B^{(t)}(x)}{\sum_{w_2\in n}C^{(t)}(w_2)\prod_{x\in n\setminus w_2}B^{(t)}(x)}=\frac{C^{(t)}(w)/B^{(t)}(w)}{\sum_{w_2\in n}C^{(t)}(w_2)/B^{(t)}(w_2)}\end{aligned} \tag{3}$$

为了对记号进行简化，在下文中用 $z^{(t)}(w,n)$ 来表示 $E(z(w,n)|B^{(t)},C^{(t)})$。在M步骤中，通过最大化式(2)所示的目标函数计算参数值 $B^{(t+1)}$ 和 $C^{(t+1)}$。根据优化函数的求解方法，可以得到如下解：

$$\begin{aligned}B^{(t+1)}(w)&=\frac{\sum_{n|w\in n}(1-z^{(t)}(w,n))}{\sum_{w'\in W}\sum_{n|w'\in n}(1-z^{(t)}(w',n))}\\&=\frac{\sum_{n|w\in n}(1-z^{(t)}(w,n))}{-|N|+\sum_{n\in N}|n|}\end{aligned} \tag{4}$$

$$C^{(t+1)}(w)=\frac{\sum_{n|w\in n}z^{(t)}(w,n)}{\sum_{w'\in W}\sum_{n|w'\in n}z^{(t)}(w',n)}=\frac{\sum_{n|w\in n}z^{(t)}(w,n)}{|N|} \tag{5}$$

通过EM方法求解式(2)中参数 B、C 和 z 的具体过程如算法1所示。

算法1 文本名称模型的EM算法

输入：词汇表W，名称集合N；

初始化：令 $B^{(1)}$ 和 $C^{(1)}$ 为均匀分布；

循环体(循环次数为 $1\leqslant t\leqslant M$)：

E步骤：$z^{(t)}(w,n)\leftarrow\dfrac{C^{(t)}(w)/B^{(t)}(w)}{\sum_{w_2\in n}C^{(t)}(w_2)/B^{(t)}(w_2)}$；

M步骤：

$C^{(t+1)}(w)\leftarrow(\sum_{n|w\in n}z^{(t)}(w,n))/|N|$；

$B^{(t+1)}(w)\leftarrow(\sum_{n|w\in n}(1-z^{(t)}(w,n)))/(-|N|+\sum_{n\in N}|n|)$；

1.2 考虑位置的文本名称模型

在包含位置数据的名称集合中，词汇的分布与地点所坐落的地理空间相关。本文根据地理坐标将世界地图分成网格，并根据名称集合学习每个网格内的词汇分布。

对于名称集合中的每个名称 n，按照其地理坐标，将其与对应的网格 l 相关联，那么每个名称可以表示成 (n,l)。对于每个网格 l，建立相应的背景上下文模型 $B[l]$，用来描述该网格内部背景词汇的概率分布情况。如果名称 n 的背景词汇为 $back(n)$，那么 $B[l]$ 的极大似然估计为 l 中相应单词的相对数量，即

$$B[l]_{ml}(w)=\frac{count(w;l)}{\sum_{w'}count(w';l)} \tag{6}$$

其中，$count(w;l)$ 表示单词 w 在网格 l 中出现的次数。

由于式(6)可能会导致数据的过拟合现象，因此本文采用信息检索中文档模型的平滑技术。通过在 $B[l]$ 中线性插入全局背景模型 B 的极大似然估计，可以得到如下公式：

$$B[l](w)=\lambda B[l]_{ml}(w)+(1-\lambda)B(w) \tag{7}$$

其中全局背景模型 $B(w)$ 表示以 w 作为背景单词的名称个数占所有名称集合的比例。

$$B(w)=\frac{|\{n|w\in back(n),n\in N\}|}{|N|} \tag{8}$$

令 L 表示网格集合，P 表示地点集合，那么每个地点 $p\in P$ 包含一个名称 $p.n$ 和一个网格 $p.l$。通过在文本名称模型中引入位置信息，可以得到算法 2 所示的 EM 算法。

算法 2 位置-文本名称模型 EM 算法

输入：词汇表 W，地点集合 P 和网格集合 L；

初始化：令 $B^{(1)}$，$C^{(1)}$ 和 $B[l]^{(1)}$（$\forall l\in L$）为均匀分布；

循环体（循环次数为 $1\leqslant t\leqslant M$）：

E 步骤：$z^{(t)}(w,p)\leftarrow\frac{C^{(t)}(w)/B[p.l]^{(t)}(w)}{\sum_{w_2\in p.n}C^{(t)}(w_2)/B[p.l]^{(t)}(w_2)}$；

M 步骤：

$C^{(t+1)}(w)\leftarrow(\sum_{n|w\in p.n}z^{(t)}(w,p))/|P|$；

$B^{(t+1)}(w)\leftarrow(\sum_{p|w\in p.n}(1-z^{(t)}(w,p)))/(-|P|+\sum_{p\in P}|p.n|)$；

$B[l]_{ml}^{(t+1)}(w)\leftarrow\frac{\sum_{p|w\in p.n,p.l=l}(1-z^{(t)}(w,p))}{\sum_{w'\in W}\sum_{p|w'\in p.n,p.l=l}(1-z^{(t)}(w',p))}$；

$B[l]^{(t+1)}(w)=\lambda B[l]_{ml}^{(t+1)}(w)+(1-\lambda)B^{(t+1)}(w)$；

1.3 位置消重

本小节将（位置）文本名称模型与字符串的编辑距离相结合，从而进行位置记录的消重。给定同属于一个网格内的两个地点，本文的目标是判断这两个地点的名称是否指向相同的商铺或者兴趣点。

在 1.1 节，假设每个名称中仅含有一个核心单词，然而实际的地点名称中可能含有多个核心单词。此处取消这一约束条件，使每个名称中可能含有多个核心单词。给定位置 p，假设 $p.n$ 中的每个单词都是以概率 α 来源于核心词汇分布 C，以概率 $1-\alpha$ 来源于背景单词分布 $B[p.l]$（$p.l$ 表示位置 p 所处的网格），那么可得

$$\Pr(core(w,p))=\frac{\alpha C(w)}{\alpha C(w)+(1-\alpha)B[p.l](w)} \tag{9}$$

其中，$core(w,p)$ 表示单词 w 为位置 p 的核心单词。为了简化符号的表示，在下文中用 $c(w,p)$ 来表示 $\Pr(core(w,p))$。

对于位置 p，用 $core(p)$ 表示 p 的核心单词随机变量，那么两个位置 p_1 和 p_2 具有相同核心单词的充要条件为：

a) p_1 和 p_2 的名称对称差中的单词全部来源于背景；

b) p_1 和 p_2 的名称中包含的共有单词在两个名称中都为核心单词，或者都为背景单词。

将上述两个条件进行形式化描述，可以得到

$$\begin{aligned}\Pr(core(p_1)=core(p_2))&=(\prod_{w\in p_1\setminus p_2}1-c(w,p_1))\times(\prod_{w\in p_2\setminus p_1}1-c(w,p_2))\times\\&(\prod_{w\in p_1\cap p_2}c(w,p_1)c(w,p_2)+(1-c(w,p_1))(1-c(w,p_2)))\end{aligned} \tag{10}$$

接下来介绍如何在式(10)中引入字符串编辑操作。假设字符串编辑操作集合 E 包含缩写操作（如 North West → NW）、连接操作（North West → NorthWest）和字母编辑（gooogle → google）。给定字符串编辑操作集合 E 以及概率函数 $\pi:E\rightarrow[0,1]$，位置名称的产生模型如下：给定位置 p，p 中的每个单词相互独立地以概率 α 来源于核心单词模型 C，以概率 $1-\alpha$ 来源于背景单词模型 $B[p.l]$；然后以概率 $\pi(e)$ 应用编辑操作 $e\in E$ 对生成的名称进行字符串编辑操作。最后，计算两个位置具有相同核心单词的概率。

本文应用动态规划方法计算两个位置具有相同核心单词的概率。在编辑操作 e 下，用 $l(e)$ 表示该编辑操作处理过的单词个数，$r(e)$ 表示操作结果中含有的单词个数。由于每个字符串编辑操作 e 的使用概率 $\pi(e)$ 可以通过对数据集进行分析得出，因此本文假设每个编辑操作的概率都是已知的。用 $p.n[i,j]$ 表示名称 $p.n$ 的第 i 个单词到第 $j-1$ 个单词构成的子序列。对于位置 p_1 和 p_2（单词个数分别为 l_1 和 l_2），它们的相似性计算方法如算法 3 所示。

算法 3 Similarity(p_1, p_2)

输入：p_1，p_2；

Return DP(p_1, p_2, 0, 0)；

函数 DP(p_1, p_2, i, j)：

1. 如果 $i=l_1$ 并且 $j=l_2$，那么 Return 1；
2. 初始化 ret←0；
3. 遍历 E 中的每个操作 e：
4. 令 $W_1\leftarrow p_1.n[i,i+l(e)]$，$W_2\leftarrow p_2.n[j,j+r(e)]$；
5. 如果 $e(W_1)=W_2$，那么令
6. ret←ret+$\pi(e)\cdot$DP(p_1, p_2, $i+l(e)$, $j+r(e)$)；
7. Return ret；

在算法 3 中，如果字符串操作集合 E 中仅包含插入、删除和复制 3 种操作，那么算法 3 等价于式(10)。为此，可以通过增加字符串操作符对式(10)进行扩展。当某个单词出现在第一个名称中而没有出现在第二个名称中时，删除操作符以概率 $1-c(w,p_1)$ 将其从第一个名称中删除；当某个单词没有出现在第一个名称中反而出现在第二个名称中时，插入操作符以概率 $1-c(w,p_2)$ 将其插入到第一个名称中；当某个单词同时出现在两个名称中时，可以以概率 $c(w,p_1)c(w,p_2)$ 进行复制操作，也可以以概率 $(1-c(w,p_1))\times(1-c(w,p_2))$ 将其删除后再插入进来。

2 实验结果与分析

2.1 位置数据库与对比算法

为了对算法的性能进行评估，实验选用了 WikiMapia[15] 和 Facebook Plcaes[7] 两个公开的位置数据库。WikiMapia 是一个开放的内容协同地图项目，该项目的目的是对世界上所有的地理对象进行标记和描述。WikiMapia 数据库包含两千万个对象，每个对象包含名称和地理位置（经/纬度）。实验选取了 WikiMapia 数据库中美国地区的数据，大约为一百万个对象。Facebook Plcaes 是 Facebook 社交网络提供的位置签到服务。对于 Facebook 位置数据库，实验同样选取了美国地区的数据。

实验中，将 1.1 节提出的文本名称模型记为名称模型，将 1.2 节提出的考虑位置的文本名称模型记为位置名称模型，将这两个模型与传统的基于编辑距离的相似性模型[13] 和 TF-IDF 模型[8] 进行对比。

2.2 参数评估

在该组实验中，根据未标识的名称数据对名称中的单词进行分类，从而识别出这些名称中的核心单词。在结果的验

证中，将分类结果与已经标识的标签进行对比。本文提出的位置名称模型主要包含两个参数 λ（见式(7)）和 α（见式(9)）。在评估参数 λ 时，令 $\alpha=0.5$，通过调整 λ 的取值观察算法的准确率。位置名称模型在 WikiMapia 和 Facebook 两个数据库中的准确率随着参数 λ 的变化分别如图 2 和图 3 所示。在这两幅图中，算法的准确率在 $\lambda=0.9$ 处达到了峰值，分别为 0.91 和 0.87。

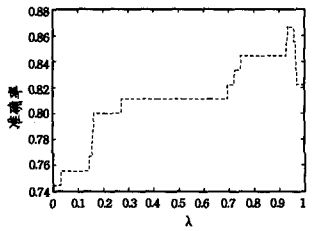

图 2　位置名称模型的参数 λ 评估(WikiMapia)

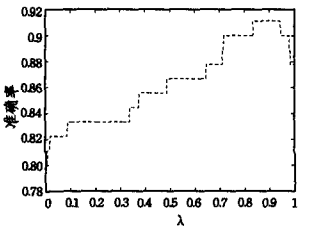

图 3　位置名称模型的参数 λ 评估(Facebook)

在评估参数 α 时，令 $\lambda=0.9$，通过调整 α 的取值观察算法在识别核心单词时的 F 值（准确率和召回率的调和平均值）。从图 4 可以看出，当 $\alpha\approx0.5$ 时，位置名称模型在 WikiMapia 和 Facebook 两个数据库下的 F 值都是最大的。

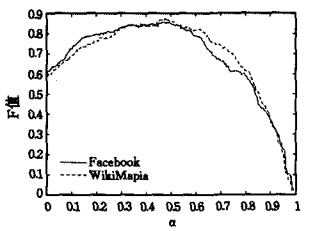

图 4　位置名称模型的参数 α 评估

2.3　实验结果

在算法的性能对比中，位置名称模型中的参数分别取值为 $\lambda=0.9$，$\alpha=0.5$。首先，通过实验对比了位置名称模型与 TF-IDF 模型在核心单词识别时的性能。在结果的验证中，将分类结果与已经标识的标签进行对比，以衡量算法的准确率和召回率。图 5 和图 6 分别为两种算法在 WikiMapia 和 Facebook 两个数据库上的测试结果。从这两幅图中可以看出，TF-IDF 模型在 WikiMapia 数据库上的性能要好于 Facebook，其原因为 Facebook 数据库中包含的位置信息量大，并且位置信息的粒度多种多样。此外，位置名称模型在两个数据中的性能都好于 TF-IDF 模型。

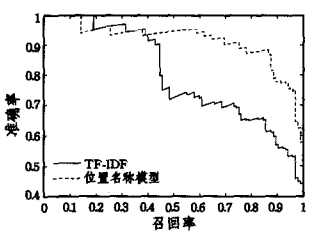

图 5　算法关于核心单词的识别对比(WikiMapia)

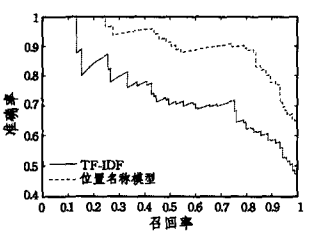

图 6　算法关于核心单词的识别对比(Facebook)

最后，分别应用名称模型、位置名称模型、基于编辑距离的相似性模型和 TF-IDF 模型 4 种方法对 Facebook 数据库中的重复记录进行消重。图 7 为采用上述 4 种方法进行位置记录消重实验得到的准确率和召回率分布图。在 4 种算法的对比中，基于编辑距离的相似性消重方法的性能是最差的，TF-IDF 模型稍好，本文提出的名称模型和位置名称模型的性能明显高于其它两种方法。此外，位置名称模型由于在名称模型的基础上进一步考虑了位置信息在位置数据库中的重要性，因此消重结果要好于名称模型。

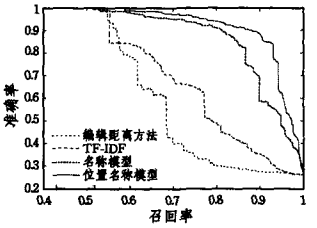

图 7　算法在 Facebook 中对位置名称记录的消重结果对比

结束语　在数据的清洗与集成中，重复记录的识别与删除是至关重要的一步。由于签到记录中的每个实体仅包含名称和位置两个属性，这给名称消重带来了巨大的挑战。本文提出了一种基于期望最大化的位置名称消重算法。首先，提出了一种包含核心单词和背景单词的文本名称模型，并给出了模型参数值计算的期望最大化算法。接下来，在文本名称模型中引入位置信息，将整个地图划分为若干个网格，分别计算每个网格中核心单词和背景单词的分布情况，并提出了一种考虑位置的文本名称模型。最后，将位置文本名称模型用于位置信息记录中的名称消重，并给出了相应的名称消重算法。实验表明，与传统的名称消重模型相比，本文提出的位置名称消重模型可以更好地识别出名称中包含的核心词汇，因而在名称消重时具有更好的性能。

参 考 文 献

[1] Ye Huan-zhuo, Wu Di. A Survey of Approximately Duplicated Data Cleaning Method [J]. New Technology of Library and Information Service, 2010(9): 56-66(in Chinese)
叶焕倬，吴迪. 相似重复记录清理方法研究综述[J]. 现代图书情报技术，2010(9)：56-66

[2] Guo Zhi-mao, Zhou Ao-ying. Research on Data Quality and Data Cleaning: a Survey[J]. Journal of Software, 2002, 13(11): 2076-2082(in Chinese)
郭志懋，周傲英. 数据质量和数据清洗研究综述[J]. 软件学报，2002，13(11)：2076-2082

[3] Pang Xiong-wen, Yao Zhan-lin, Li Yong-jun. Efficient duplicate records detection method for massive data[J]. Journal of Huazhong University of Science and Technology(Natural Science Edition), 2010, 38(2): 8-11(in Chinese)
庞雄文，姚占林，李拥军. 大数据量的高效重复记录检测方法[J]. 华中科技大学学报(自然科学版)，2010，38(2)：8-11

[4] Zheng Y. Location-based social networks: Users[M]// Computing with Spatial Trajectories. Springer New York, 2011: 243-276

[5] Ye M, Yin P, Lee W C. Location recommendation for location-based social networks[C]// Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2010: 458-461

[6] Sakib M N, Bin Halim J, Huang C T. Determining Location and Movement Pattern Using Anonymized WiFi Access Point BSSID [C]// 2014 7th International Conference on Security Technology (SecTech). IEEE, 2014: 11-14

(下转第 251 页)

Mechatronics 21,2011:509-522
[4] Riener R, Lunenburger L, Jezernik S, et al. Patient-cooperative strategies for robot-aided treadmill training: first experimental results[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering,2005,13(3):380-394
[5] Yang Da-peng, Jiang Li, Zhao Jing-dong, et al. High intelligent prosthetic hand control based on EEG signal[J]. Journal of Jilin University(Engineering and Technology Edition),2008,38(5):1225-1230(in Chinese)
杨大鹏,姜力,赵京东,等. 基于脑电信号的高智能假手控制[J]. 吉林大学学报(工学版),2008,38(5):1225-1230
[6] Okamura J, Tanaka H, Sankai Y. EMG-based Prototype Powered Assistive system for Walking Aid [C]// Proceedings of Asian Symposium on Industrial Automation and Robotics (ASIAR99). Bangkok, Thailand,1999:229-234
[7] Racine, Louis Charles J. Control of a Lower Extremity Exoskeleton for Human Performance Amplification [D]. University of California, Berkeley,2003
[8] Yano H, Kaneko S, Nakazawa K, et al. A New Concept of Dynamic Orthosis for Paraplegia: the Weight Bearing Control (WBC) Orthosis [J]. Prosthetics and Orthotics International, 1997,21(3):222-228
[9] Neuhaus P, Kazerooni H. Design and control of Human Assisted Walking Robot [C]// Proceeding of 2000 IEEE International Conference on Robotics & Automation. San Francisco, 2000: 563-569
[10] Kazerooni H, Steger R. The Berkeley Lower Extremity Exoskeleton [J]. Journal of Dynamic Systems, Measurement, and Control,2005,128(3):14-25
[11] http://www.hocoma.com/products/lokomat/
[12] Homoca. LokomatPro Brochure [OL]. http://www.hocoma.com/fileadmin/user/Dokumente/Lokomat/bro_L6_120416_en_A4.pdf
[13] Esquenazi A, Talaty M, Packel A, et al. The ReWalk powered exoskeleton to restore ambulatory function to individuals with thoracic-level motor-complete spinal cord injury[J]. American Journal of Physical Medicine & Rehabilitation, 2012, 91(11): 911-921
[14] http://www.medgadget.com/2008/03/rewalk_exoskeleton.html
[15] Bogue R. Exoskeletons and robotic prosthetics: a review of recent developments[J]. Industrial Robot: An International Journal,2009,36(5):421-427
[16] Kong K, Jeon D. Fuzzy control of a new tendon-driven exoskeletal power assistive device[C]// Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM 2005). 2005:146-151
[17] Agrawal S K, Banala S K, Fattah A. A Gravity Balancing Passive Exoskeleton for the Human Leg[OL]. http://www.roboticsproceedings.org/rss02/p24.pdf
[18] Bin Niu. Study on the design and Control of a Wearable Exoskeleton Leg for Huaman's Walking Power Augmentation[D]. Hangzhou: Zhejiang University,2006
[19] Fei Ye-yun. Design and Study of Lower Artificial Limb Exoskeleton Based on the Control of EMG for Rehabilitation Physic [D]. Hangzhou: Zhejiang University,2006
[20] Zhang Jie. Study on the exoskeleton Leg for training paraplegic patients[D]. Hangzhou: Zhejiang University,2007
[21] Zhang Yu. Development and Research of Ankle-foot Rehabilitation Exoskeleton Orthosis[D]. Hangzhou: Zhejiang University, 2010
[22] Pan Hui-ju. A new method to study the relationship between mechanical behavior of human motion device system and sports injury[J]. Journal of Zhejiang Normal University(Natural Science Edition),1995,18(4):4-7(in Chinese)
潘慧炬. 人体运动器系力学行为与运动损伤关系研究的新方法[J]. 浙江师范大学学报(自然科学版),1995,18(4):4-7

(上接第 241 页)

[7] Chang J, Sun E. Location 3: How users share and respond to location-based data on social networking sites[C]// Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. 2011:74-80
[8] Tata S, Patel J M. Estimating the selectivity of tf-idf based cosine similarity predicates[J]. ACM SIGMOD Record, 2007, 36 (2):7-12
[9] Xu Yi-zhen, Wang Yong-cheng. A Fast Algorithm for Matching Multiple Patterns[J]. Journal of Shanghai Jiaotong University, 2002,36(4):516-520(in Chinese)
许一震,王永成. 一种快速的多模式字符串匹配算法[J]. 上海交通大学学报,2002,36(4):516-520
[10] Sun De-cai, Sun Xing-ming, Zhang Wei, et al. A Fitter Algorithm for Approximate String Matching Based on Match-Region Features[J]. Journal of Computer Research and Development, 2010,47(4):663-670(in Chinese)
孙德才,孙星明,张伟,等. 基于匹配区域特征的相似字符串匹配过滤算法[J]. 计算机研究与发展,2010,47(4):663-670
[11] Bilenko M, Mooney R J. Adaptive duplicate detection using learnable string similarity measures [C]// Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2003:39-48
[12] Oncina J, Sebban M. Learning stochastic edit distance: Application in handwritten character recognition [J]. Pattern recognition,2006,39(9):1575-1587
[13] McCallum A, Bellare K, Pereira F. A conditional random field for discriminatively-trained finite-state string edit distance[C]// Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05). Arliagton, Virginia: AVAI Press,2005
[14] Huang Lin-sheng, Deng Zhi-hong, Tang Shi-wei, et al. A Chinese organization's full name and matching abbreviation algorithm Based on edit-distance[J]. Journal of Shandong University (Natural Science),2012,47(5):43-48(in Chinese)
黄林晟,邓志鸿,唐世渭,等. 基于编辑距离的中文组织机构名简称-全称匹配算法[J]. 山东大学学报(理学版),2012,47(5):43-48
[15] Fritz S, McCallum I, Schill C, et al. Geo-Wiki: An online platform for improving global land cover[J]. Environmental Modelling & Software,2012,31:110-123
[16] Moon T K. The expectation-maximization algorithm[J]. Signal Processing Magazine, IEEE,1996,13(6):47-60
[17] Chen Qing-zhi, Chen Guo-long, Guo Wen-zhong, et al. A Hybrid Clustering Algorithm for Information Security Evaluation Log Data[J]. Journal of Chongqing Institute of Technology(Natural Science),2009,23(10):77-82,118(in Chinese)
陈庆枝,陈国龙,郭文忠,等. 信息安全评估日志数据的一种混合聚类算法[J]. 重庆工学院学报(自然科学),2009,23(10):77-82,118