大数据环境下基于贝叶斯推理的中文地名地址匹配方法

许普乐 王 杨 黄亚坤 黄少芬 赵传信 陈付龙
(安徽师范大学数学计算机科学学院 芜湖 241000)

摘 要 传统的中文地名地址匹配技术难以处理大数据环境下海量、多样和异构的智慧城市地理信息空间中的中文地名地址快速匹配问题。提出了一种Spark计算平台下基于中文地名地址要素的匹配框架及应用智能决策的匹配算法(An Intelligent Decision Matching Algorithm, AIDMA)。首先,从中文地名地址中富含的语义性和中文字符串、数字与字母之间的自然分隔性两个方面进行地址要素解析,构建了融合多距离信息的贝叶斯推理网络,从而提出了基于多准则评判的中文地名地址匹配决策方法。然后,利用芜湖市514967条脱敏后的燃气开户中文地名地址信息库与1770979条网格化社区中的中文地名地址信息库(包含网格化地址的地理空间信息)进行实验与分析。实验结果表明,在处理大规模中文地名地址信息时,相比于传统的中文地名地址匹配方法,该方法能够有效提高单条中文地名地址的匹配效率,同时在匹配度与精确度两个指标上匹配结果更加均衡。

关键词 大数据,Spark,中文地名地址匹配技术,贝叶斯推理
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.09.050

Chinese Place-name Address Matching Method Based on Large Data Analysis and Bayesian Decision

XU Pu-le WANG Yang HUANG Ya-kun HUANG Shao-fen ZHAO Chuan-xin CHEN Fu-long
(School of Mathematics & Computer Science, Anhui Normal University, Wuhu 241000, China)

Abstract Traditional matching technologies of Chinese place-name address is hard to deal with the fast matching problem of Chinese place-name address in matching massive, diverse and heterogeneous geographic information under the big data environment. An intelligent decision matching algorithm(AIDMA) based on computing framework of Spark was proposed. Firstly, geographical elements are analyzed from semantic information and separations of Chinese strings, numbers and letters. Bayesian networks is constructed with three kind of distance combined with multi-criteria decision-making effectively. 514957 desensitized gas account information and 1770979 grid addresses information which includes spatial information of Wuhu City are used to perform the experiments. The conclusions prove that the executed time of each record of AIDMA is reduced to about 2.2s from 1min when compared to traditional algorithms. The matching results are more balanced on matching rate and precise rate. The proposed method possesses the theoretical significance and application value on the road to construct the intelligent countries.

Keywords Big data, Spark, Matching technologies of Chinese place-name address, Bayesian decision

1 引言

近年来,随着智慧城市建设水平的不断提高,高效快捷的中文地名地址匹配技术成为了智慧城市基础数据获取与应用的必然需求[1-2]。传统的中文地名地址匹配技术主要建立在社区网格化[3-4]的基础之上,但是城市计算中时空数据的日益增多,以及多种异构数据源汇聚形成的大数据环境[5-6],对传统中文地址匹配技术在信息的时空精准耦合、智能推理、处理效率以及关联决策等方面提出了新的挑战。如何建立适用于大数据环境的中文地名地址匹配方法,并提供数据分析、定位以及可视化等功能,已成为我国智慧城市发展中的现实需要。

传统的中文地名地址匹配技术[7-8]主要是指将一条文字描述的中文地名地址信息与目标数据库中的地理坐标进行映射的过程。具体而言,即对用户输入的地址信息按照一定的切词、匹配算法,在地理编码数据库中进行查找匹配,根据匹配结果标记相应的空间坐标。由于中文地名地址包含大量语义信息和历史关联信息,因此在处理过程中很难精确、快捷地匹配。从中文地址要素分词解析来看,现有的中文地名地址匹配方法主要分为以下3类:1)基于机械分词的地址匹配算法[9-12]。该类算法主要基于数据库中的字典对中文地址进行分词匹配。在匹配顺序上可分为正向匹配和逆向匹配两种方法。逆向匹配是中文地址处理中较为常用的匹配方法。由于

到稿日期:2017-03-05 返修日期:2017-05-09 本文受国家自然科学基金(61572036),安徽省自然科学基金(1708085MF156),安徽省重大人文社科基金项目(SK2014ZD033)资助。

许普乐(1980-),硕士,副教授,主要研究方向为大数据;**王 杨**(1971-),博士,教授,CCF高级会员,主要研究方向为社会网络、机器学习、大数据,E-mail:wycap@126.com;**黄亚坤**(1992-),硕士,主要研究方向为大数据;**黄少芬**(1993-),女,硕士生,主要研究方向为大数据;**赵传信**(1977-),男,博士,副教授,主要研究方向为物联网;**陈付龙**(1976-),男,博士,教授,主要研究方向为物联网。

中文的词义中心通常相对靠后，因此此方法的分词精度与匹配度相对较高[12]。2）基于统计分词的地址匹配方法[13-14]。统计分词主要研究上下文信息，若两个中文字符同时出现的频率较高，则可以推断其构成一个词的概率也较高。该方法仅根据中文地址的语料频度进行分词，无需字典库支持，因此进行大规模地址匹配时，其能有效降低算法的复杂性。3）基于自然语言分词的地址匹配算法[15-16]。由于中文地址包含了丰富的语义信息，同一地址可能存在多种表述形式，因此结合语义信息进行地址要素解析能够有效提高匹配的精确。

上述匹配方法主要关注匹配的精确度，忽略了在大数据环境下当前地址数据本身富含的语义多样性、场景多元化以及采集源异构等特征。因此在面对大规模中文地址匹配场景时，传统方法在处理效率、匹配准确度等方面难以满足实际应用的需要。本文基于 Spark 大数据处理框架，提出了一种基于贝叶斯推理的中文地名地址匹配方法。首先，结合语义要素和中文字符与数字、字母的自然分离性进行地址解析预处理，从而降低了推理匹配计算的规模；然后，根据解析后的中文地名地址信息给出中文串和数字的匹配距离与覆盖距离的定义，将匹配问题转化为多准则决策问题。在多种距离信息的基础上，进一步建立贝叶斯网络，结合多准则决策模型进行中文地名地址的智能匹配。

2　问题描述及地址匹配框架

为了对城市非结构化信息的时空进行耦合，目前主要通过社区地址网格化的手段使相关中文地址与坐标匹配得更加准确、高效。网格化地址匹配与传统的中文地名地址匹配的定义存在一定差异，需要针对具体的数据逻辑抽象模型进行预处理、计算及匹配检验。下文将针对智慧城市社区网格化中的中文地名地址匹配问题给出相关定义说明。

2.1　问题分析与相关说明

一般而言，地址匹配[7]可以被定义为将两个不同数据库中的街道地址进行关联的过程。待匹配数据库中存储了地址信息和坐标信息，目标数据库中包含了待匹配数据库中一条相同标记地址特征信息或强语义关联性的地址信息。地址匹配系统试图将上述不同数据库的地址信息进行匹配处理，再进行相关信息的记录，并共享匹配记录的其他信息，从而提高相关业务的质量与效率。图 1 给出了一种典型的地址匹配输出结构的示意图。

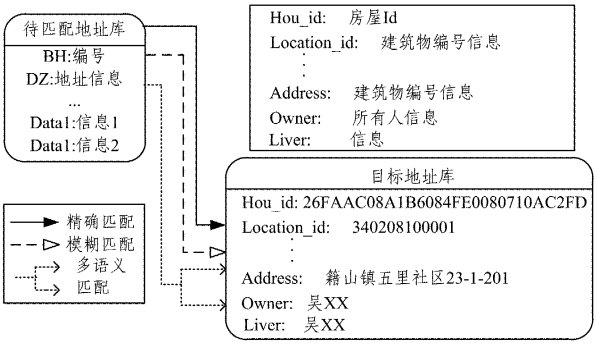

图 1　一般中文地名地址匹配输出结构示意图

从图 1 的地址匹配数据结构集中可以看出，匹配结果主要包含精准匹配、模糊匹配以及多语义匹配 3 类输出结果。由于中文地名地址要素复杂多样，多语义匹配特征在处理中文地址时表现更优。

2.1.1　中文地名地址要素解析

中文地名地址通常由以下地址要素组成：行政要素、基本约束信息和位置信息[17-18]，具体表述如下：

〈标准地址〉::=〈行政要素〉〈基本约束信息〉〈位置信息〉

〈行政要素〉::=〈国家〉〈城市〉〈行政区县〉

〈基本约束信息〉::=〈街道〉|〈集镇〉|〈工业区〉|〈自然村〉

〈位置信息〉::=〈建筑物编号〉〈门牌号〉|〈标志物〉

图 2 给出了一个示例地址“安徽省芜湖市弋江区柏庄春暖花开小区 88 幢 11 单元 703 室”的结构要素。

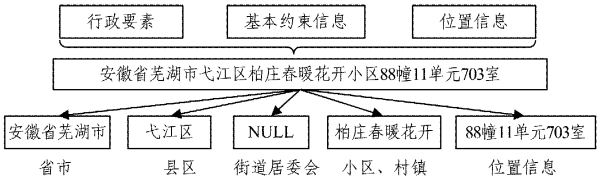

图 2　中文地址要素结构示例

通常的地址要素解析方法难以同时满足语义解析、解析效率、解析准确性等要求。中文地名地址的复杂语义特性及录入方式的多样化使得同一地址的表现形式复杂多样。本文主要结合中文地址的语义特性进行解析，并提出了中文字符、数字及字母分离的解析方法，从而降低了智能推理匹配的计算规模。在算法设计及实现方面，基于已有的成熟的语义分词框架进行地址要素解析，并将解析后的结果存入字典库。事实上，由于数字信息在中文地名地址中所占的信息量比重较大，因此通过初步模糊匹配得出预处理结果集的方式能够减小低数据处理的规模。

2.2　中文地名地址匹配框架

大数据驱动下的地址匹配算法在数据规模、运算能力和数据多样性上与传统地址匹配算法存在较大差异。首先，在数据规模上，面向大数据环境的地址匹配需要在短时间内高效地处理海量数据，其中目标地址库的数据规模一般较为庞大；其次，地址信息的复杂性和多样性使得运算模式以及计算框架需要具有智能化的特点。综合考虑，采用分布式的计算框架能够有效提高海量地址匹配计算的效率，图 3 给出了基于分布式计算的中文地址匹配框架图。

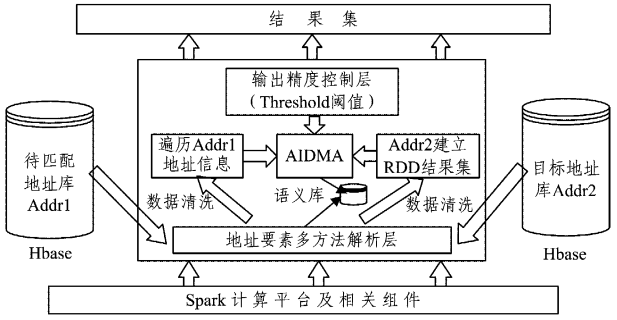

图 3　中文地名地址匹配框架

首先，针对匹配地址库与目标地址库进行数据预处理清洗与地址要素解析工作，并将语义解析结果导入语义库；然后基于目标地址信息建立弹性分布式数据集（Resilient distri-

bute Datasets,RDD),并根据英文和数字信息对 RDD 过滤,将预处理后的数据作为 3.3 节贝叶斯推理决策匹配算法 AIDMA 算法的输入;最后在采用精度控制的情况下输出决策结果。

3 智能匹配模型及相关算法

3.1 理论基础

3.1.1 贝叶斯网络

定义 1(贝叶斯网络,Bayesian Networks,BN) B 表示一个有向无环图(Directed Acyclic Graph,DAG)(V,E)。其中,节点集 V 代表随机变量,有向边集 E 代表节点间的依赖关系。通常每个节点只有一个概率分布,若用 $P(X)$ 表示根节点的概率分布,则非根节点的条件概率分布为 $P(X|\pi(X))$[19]。

若进行定量分析,BN 则是联合概率分布的分解表示。设网络中的变量描述为 $X_1,X_2,\cdots,X_n$,则每个变量的概率分布相乘可得到联合分布,如式(1)所示:

$$P(X_1,X_2,\cdots,X_n)=\prod_{i=1}^{n}P(X_i|\pi(X_i)) \tag{1}$$

其中,当 $\pi(X_i)=\varnothing$ 时,$P(X_i|\pi(X_i))$为边缘分布。

如图 4 所示,$\pi(B)=\pi(E)\neq\varnothing$,$\pi(A)=\pi(B,E)$,$\pi(M)=\pi(A)$,$\pi(J)=\pi(A)$。条件概率分布 $P(A|BE)$描述 A 如何依赖 B 和 E,$P(M|A)$和 $P(J|A)$分别定量刻画 M 与 J 如何依赖于 A,变量 B,E 不相互依赖。

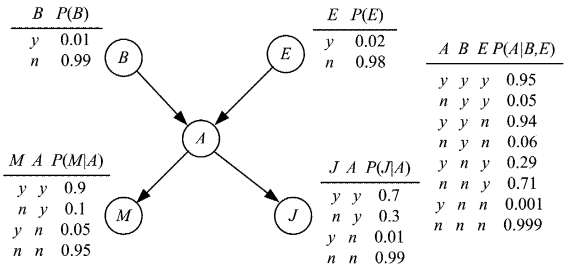

图 4 贝叶斯网络示例

可以将中文地名地址匹配视作在多个待选集合中决策出最佳匹配项。决策目标通常需要综合考虑多个事件之间的相互关系,BN 难以直接解决中文地名地址匹配决策问题[20],因此建立如式(2)所示的多准则评价决策模型:

$$\max_{x\in X} f(x) \tag{2}$$

式(2)表明,决策的本质是找出 $f(x)$的最优解。设 X 是决策方案集,记 $X=(C_1,C_2,\cdots,C_n)^{\mathrm{T}}$,则 $x=(c_1,c_2,\cdots,c_n)^{\mathrm{T}}$ 为某个决策方案描述。

3.1.2 相关距离的定义

由于相似度或距离能够用于分析两段文字或个体间的差异大小,因此据此可评判出两个实体是否相同或同属一类。在处理中文地名地址要素匹配时,需要采用适合于 BN 推理的距离计算方式。

定义 2(覆盖距离,Coverage Distance,CD) CD 定义为在由有限元素组成的有序集合中,相同元素与长度较短集合的长度比值。例如 A,B 为两个有限元素组成的有序集合,$|A|$,$|B|$分别表示 A 和 B 的集合长度,$A\cap B$ 表示进行集合交集操作,则覆盖距离可用式(3)表示:

$$cd=\frac{|A\cap B|}{\min(|A|,|B|)},A\neq\varnothing,B\neq\varnothing \tag{3}$$

定义 3(匹配距离,Matching Distance,MD) MD 是指两个分别由有限元素组成的有序集合中,从第一个元素逐一进行匹配,直至元素不相等时的距离长度与较短集合的长度比值。设 A,B 为两个中文地名地址集合,$A\wedge B$ 表示从第一个元素起进行连续匹配操作直至元素不相等时产生的集合,则该距离可表示为:

$$md=\frac{|A\wedge B|}{\min(|A|,|B|)},A\neq\varnothing,B\neq\varnothing \tag{4}$$

3.2 智能推理匹配模型

实际匹配过程中,通过对目标数据库的地址进行模糊匹配来获得预选集合,此时,原匹配问题可以转化为由多种距离评判的多准则评判问题[20],即待匹配地址为决策人从预选集合中构成的不同决策方案中选取对自己利益最大化的地址方案。

对于预处理后的地址匹配推理决策问题,决策事件 ζ 表示针对待匹配地址在预选集合中决策出最符合用户需要的地址方案,即决策收益最大化。假定 ζ 的状态空间为 $S=\{pd_1,pd_2,\cdots,pd_n\}$,$pd_i$ 表示预选地址集。定义决策目标集合为 δ,则 $\delta=$(匹配优势,匹配损失)$^{\mathrm{T}}$ 构成了决策准则向量。基于上述定义,构建了图 5 所示的基于 BN 的智能匹配模型。

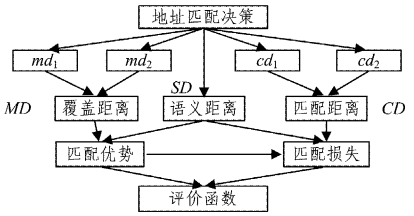

图 5 基于 BN 的智能匹配模型

该问题可表述为求下式的最优解:

$$\begin{aligned}&\max_{\delta=(cd_1,cd_2,md_1,md_2)^{\mathrm{T}}} f((cd_1,cd_2,md_1,md_2)^{\mathrm{T}})\\&=\max_{\delta=(cd_1,cd_2,md_1,md_2)^{\mathrm{T}}} f((p(cd_1|\zeta)p(\zeta),p(cd_2|\zeta)p(\zeta),\cdots,\\&\quad p(cd_n|\zeta)p(\zeta))^{\mathrm{T}})\end{aligned} \tag{5}$$

对于式(5),文献[20]给出了该 BN 推理决策模型的求解方法。所求决策向量的概率状态为 δ 的条件概率,如式(6)所示:

$$p(cd_i)=p(cd_i|\zeta)p(\zeta) \tag{6}$$

则式(5)可替换为:

$$\begin{aligned}&\max_{\delta=(cd_1,cd_2,md_1,md_2)^{\mathrm{T}}} f((cd_1,cd_2,md_1,md_2)^{\mathrm{T}})\\&=\max_{\delta=(cd_1,cd_2,md_1,md_2)^{\mathrm{T}}} f((p(cd_1|\zeta)p(\zeta),p(cd_2|\zeta)p(\zeta),\cdots,\\&\quad p(cd_n|\zeta)p(\zeta))^{\mathrm{T}})\end{aligned} \tag{7}$$

在求解上式最优解的过程中,定义 P 为节点的后验概率分布,δ 和 ζ 之间的概率关联关系通过基于距离来计算结果以及 BN 模型来计算。记 π 和 λ 为从父节点和子节点获得的相关信息量。在推理过程中主要以单节点为中心,根据 π 和 λ 计算 P,之后根据式(8)对相邻的节点概率状态进行更新[21]。

$$P(X)=\alpha\lambda(X)\pi(X) \tag{8}$$

其中，$\lambda(X)=\prod\lambda(X)$，$\pi(X)=\pi(Pa(X))p(X|Pa(X))$。$\alpha$ 是归一化参数，$Pa(X)$表示 X 的父节点信息。根据上述知识，给出各节点之间的概率关系。

对于基于上述距离构建的 BN 网络，在不同地址构成的决策方案中，匹配优势和匹配损失的概率高低有一定的关联性差异。当匹配率较低时，匹配损失较大；当匹配损失较低时，匹配率较高。相关概率关系如表 1 所列。

表 1　概率影响关系

地址方案	匹配优势		匹配损失	
	高	低	高	低
1	p_1	$1-p_1$	q_1	$1-q_1$
2	p_2	$1-p_2$	q_2	$1-q_2$
…	…	…	…	…
n	p_n	$1-p_n$	q_n	$1-q_n$

其中，$p_i=(MD_i+SD)/2$，$q_i=(CD_i+SD)/2$。根据上述 BN 模型，构建针对匹配优势和匹配损失的评价函数，如式(9)所示：

$$f(\delta)=\mu*(\text{匹配优势})+(1-\mu)*(\text{匹配损失}) \tag{9}$$

$\delta=(\mu,1-\mu)$，根据给定的 BN 模型结构和概率参数计算。对于一个待匹配决策地址，匹配损失的概率为：

$$P(\text{匹配损失})=p(\text{匹配损失}|\text{匹配优势})pr(\text{匹配优势})+p(\text{匹配损失}|\text{匹配劣势})pr(\text{匹配劣势}) \tag{10}$$

进一步比较不同决策方案的评判函数值 $f_i(\delta)$，即可获得最优方案。

3.3　算法描述与分析

针对上述计算模型，给出面向大数据处理的智能中文地名地址匹配算法描述，并对该算法进行相关分析。

算法 1　贝叶斯推理决策匹配算法 AIDMA

```
1. Input:addr1,DataSet(addr2),μ(阈值因子)
2. Output:ResultSet(top-kmatching)
3. Begin
4.   ad1←PreSove(addr1);
5.   rdd←Read_from_hbase(Sets(addr2));
6.   dev1←Dev(sets(ad1));
7.   filter_rdd←rdd.filter(num&cha);
8.   For each elements in rdd
9.        dev2←Dev(per filter_rdd);
10.       Calculate(Sem_dis);
11.       If max(f(md1,md2,cd1,cd2))>μ
12.            ResultSets←Result.add(current_filter_rdd);
13.       End if
14.  End For
15. End
```

该算法将不同来源的地址信息进行数据清洗和地址要素解析后的信息作为 AIMDA 的输入。其中，将构建的 RDD 数据集作为目标地址库，并将其缓存至内存中进行过滤，最后得到 AIMDI 的决策集合。算法通过一次遍历即可完成 top-k 地址的匹配。AIMDI 算法中单条地址匹配复杂度为 O(n)。

4　实验及结果分析

为验证本文方法的可行性，下面针对上述大数据环境下的贝叶斯推理决策匹配算法(AIDMA)进行相关实验验证。实验环境采用基于 Spark 的计算平台，数据存储形式为 HDFS 文件形式，CDH 部署的集群环境主要为：配置内存均为 32GB 的 8 台计算节点，存储磁盘大小为 50GB，1 台 CDH 管理机的节点配置同上；此外，配置 1 台物理内存大小为 16GB 并安装 yarn 和 Zookeeper 的服务器进行资源管理调度。具体的集群组件配置图如图 6 所示。

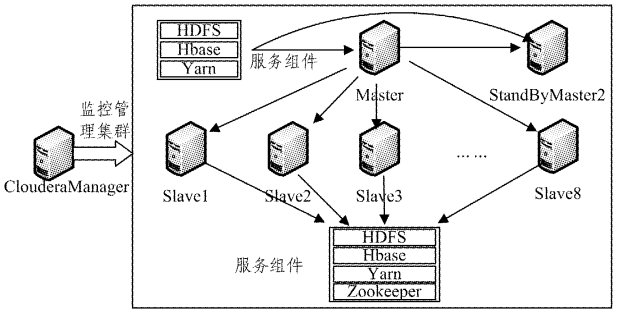

图 6　基于 CDH 部署的 Spark 集群组件配置图

实验数据集由两部分组成，一部分是对个人信息脱敏后的芜湖市燃气开户信息，共计 514967 条地址记录；另一部分是社区网格化后的中文地名地址信息，进行个人信息脱敏后共计 1770979 条，该数据主要存储在 Hbase 数据库中，同时此数据库具有较高的读写性能。采集的原始网格地址信息在数据质量方面相对较差，需要对这两种地址库进行匹配关联。所提算法主要与文献[12]中的基于地址库支持方法、文献[14]中的统计分析方法以及常规模糊匹配与精准串匹配方法进行性能对比。在验证不同方法的有效性方面，主要采用的评价指标如下。

(1)匹配率[22]

$$mr=\frac{N}{M}*100\% \tag{11}$$

其中，M 为需要匹配的地址条数，N 为能够准确匹配获得的地址条数。

(2)准确率[22]

$$pr=\frac{K}{M}*100\% \tag{12}$$

其中，M 为理想匹配的地址条数，K 为正确匹配的地址条数。

4.1　实验及分析

图 7、图 8 给出了从匹配率和匹配结果集的准确率方面对算法进行有效性分析的结果。图 9 给出了匹配过程中阈值 threshold 对匹配率及准确率的影响。

图 7 主要给出了 AIDMA 与其他 4 种算法在匹配率上的结果比较。对比结果表明，模糊匹配算法与精确匹配算法分别处于两侧的位置。精确匹配算法主要是对相关分词后的地址要素进行传统串匹配操作，其中富含大量语义信息的地址数据被过滤掉。当数据规模较小时，匹配率相对较高，当数据规模为 20%时，匹配率为 0.42。模糊匹配算法则忽略了中文地址的语义信息，采用传统的模糊串匹配操作，其总体匹配率很高，达到了 0.95。随着数据规模的增长，地址的数量增大，两种算法的匹配率均有一定幅度的下降，主要原因是数据规模增大的同时，“脏数据”地址信息(如空地址、错序地址信息数)也在增多。

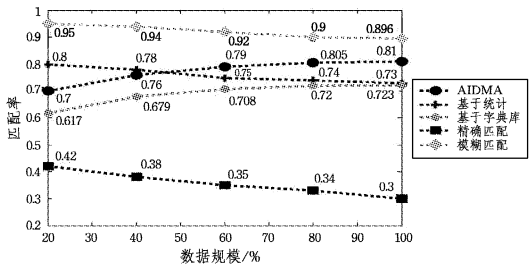

图 7 匹配率比较示意图

AIDMA 与基于统计的匹配算法在匹配率上都处于 0.8 左右。随着数据规模的增大，AIDMA 的匹配率波动小，尤其在数据规模达到 60%以后，其匹配率处于稳定的状态，即达到0.8。此外，观察统计算法的曲线信息可以看出，该算法的匹配率随着数据规模的增加呈现出与精确匹配和模糊匹配算法相同的下降趋势，主要原因也是因为脏数据的逐渐增大。而基于字典库支持的匹配算法的曲线变化趋势与 AIDMA 相似：随着数据规模的增大，呈现出了一定的上升趋势。当数据规模达到 60%以后，字典库的信息相对比较稳定，因此在匹配率上也较稳定。从表面上来看，这两种算法的匹配率并没有随着脏数据量的增加而降低，但随着数据规模的不断增长，AIDMA 的匹配率会逐渐收敛于匹配上限值。

图 8 反映了不同算法在匹配准确率指标上的变化情况。从图 8 可以看出，精确匹配虽然匹配率低，但其准确率最高，这主要是由算法的特性决定的。模糊匹配在匹配率上占优，但在准确率方面处于劣势。这两种匹配算法在单一指标上表现出较大优势，但从实际应用角度来看，难以在多指标上进行均衡匹配。统计理解与字典库匹配算法的精确度在 0.5 左右。从曲线趋势的角度看，统计理解的算法呈现逐渐下降的趋势，因为该算法在每次匹配时仅仅依赖于当前的中文地址串信息。而对于字典库支持算法，在数据规模增大时，字典信息逐渐完善，其匹配率会随之增大；当数据规模达到 60%时，准确率与匹配率的趋势相似，字典库相对稳定时，准确率也趋于稳定，达到 0.584。

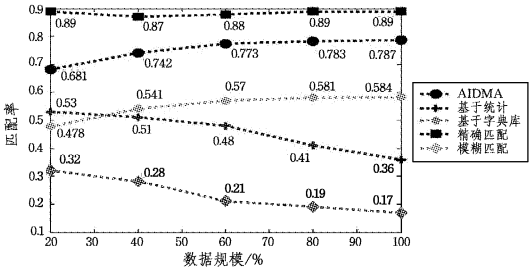

图 8 准确率比较示意图

AIDMA 算法基于解析的多种距离信息构建贝叶斯网络，结合了多准则决策推理进行匹配计算；在地址解析与匹配时考虑了中文地址富含的语义信息。整体来看，AIDMA 在匹配率上虽然不及精确匹配算法，但随着数据规模的增长，其准确率呈上升趋势，且趋于稳定，同时在匹配率上也比较稳定。综合图 7 与图 8 来看，AIDMA 算法在匹配率与准确率上都表现出较好的均衡性。随着数据规模的增大，AIDMA 算法在两项指标上趋于稳定。

AIDMA 中的阈值 μ 在实际应用过程中主要用于均衡匹配率和准确率。图 9 给出了在不同数据规模下 μ 的平均值与匹配率和准确率的曲线变化图。实验结果表明，随着阈值 μ 的增大，准确率增加，匹配率降低。对不同数据规模的 μ 进行实验，结果显示两条曲线在 μ 为 0.6 左右时相交，可以将该点定义为匹配率与准确率的均衡点。

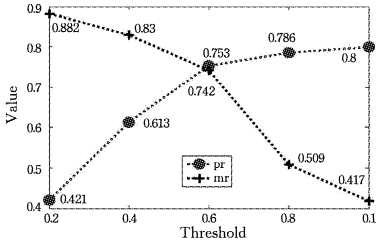

图 9 Threshold 示意图

4.2 算法效率分析

针对上文提到的多种中文地址匹配算法，为了分析 AIDMA 算法效率，下文主要从 AIDMA 算法处理不同规模的数据对执行效率进行分析，并给出了不同算法在较小规模（以20%为例）时与本文算法的执行效率的比较结果。

图 10 给出了不同数据规模下 AIDMA 算法对平均每条地址的匹配时间代价。图中曲线说明随着数据规模的增长，基于 Spark 计算的地址匹配的时间代价逐渐减小，从而体现出了其在处理大规模数据时的优势。基于 Spark 计算框架进行分布式计算编程时，首条待匹配地址对目标地址库的地址数据建立 RDD（弹性分布式数据集），并将其缓存至内存中，在下一条地址进行匹配时直接从内存中读取，有效减少了查找时间。此外，每次计算集群的启动时间较长，当数据规模较小时，集群的启动时间占据了较大比例；当数据规模大时，单条地址匹配计算的时间趋于稳定。

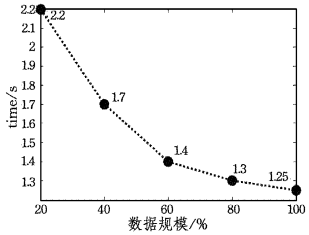

图 10 不同数据规模下AIDMA 算法的执行时间分析

表 2 不同算法对平均每条地址的执行时间的对比

Algorithms	Time
AIDMA	2.2s
基于统计	1min
基于字典库	1.5min
精确匹配	1.2min
模糊匹配	1.3min

表 2 列出了对比算法在小规模（20%）的数据集下与本文算法对单条地址的处理时间的对比。除 AIDMA 外，其他 4 种算法的实验环境均为集群中的单台服务器配置。从处理时间的单位来看，AIDMA 为秒，而其他算法为分钟，因此 AIDMA 更适用于大数据环境下的中文地址匹配。从数值上分析，其他 4 种算法的执行时间都在 2min 内，相差较小；从执行效率上来看，其他 4 种算法仅适用于处理小规模的中文地址匹配。

结束语 大数据环境下的中文地址匹配技术对推动智慧社区、智慧城市发展提供了有力的技术支持。本文基于成熟

的 Spark 大数据处理框架，根据中文地址的语义信息进行地址要素解析，并提出了中文字符串、数字和字母分离的解析预处理方法，有效降低了匹配过程中的数据规模；其次，根据定义的多种距离信息构建了贝叶斯网络，结合多准则决策模型，给出了在预选集合中的智能匹配算法。最后，基于芜湖市 514967 条个人信息脱敏后的燃气地址与社区网格化地址库中的地址（共计 1770979 条）进行匹配验证，主要从匹配率、准确率以及算法效率进行分析。实验结果显示，相比于传统匹配算法，本文提出的算法在匹配率和精确率上具有更好的均衡性。特别是在处理大规模数据时，相比于传统地址匹配算法，该算法在一定程度提升了性能、效率，对于实际的中文地名地址信息处理具有一定的应用价值。

参 考 文 献

[1] REMERO，BARRIGA，MOLANO. Big Data Meaning in the Architecture of IoT for Smart Cities [C]// International Conference on Data Mining and Big Data. Springer International Publishing，2016：457-465.

[2] DELMASTRO F，ARNABOLDI V，CONTI M. People-centric computing and communications in smart cities [J]. IEEE Communications Magazine，2016，54(7)：122-128.

[3] LIU D，PEI Y，LI C. Research on Establishment of Grid-based Intelligent Community Synergistic Service Platform [J]. Bulletin of Surveying and Mapping，2015，3(12)：98-100.

[4] PU Z，XU L. Research to the Community Resources Integration Under Grid City Management [J]. Asian Social Science，2009，4(7)：64-68.

[5] LI D R，CAO J J，YAO Y. Big data in smart cities [J]. Science China Information Sciences，2015，58(10)：1-12.

[6] HASHEM I A T，CHANG V，ANUAR N B，et al. The role of big data in smart city [J]. International Journal of Information Management，2016，36(5)：748-758.

[7] GOLDBERG D W，WISON J P，KNOBLOCK C A. From text to geographic coordinates：the current state of geocoding [J]. Urisa Journal，2007，19(1)：33-46.

[8] DRUMMOND W J. Address Matching：GIS Technology for Mapping Human Activity Patterns [J]. Journal of the American Planning Association，1995，61(61)：240-251.

[9] SUN Y，CHEN W. Address Matching Technology Based on Word Segmentation [C]// China Geographic Information System Association Annual Meeting. 2007：1-12.

[10] MA Z，LI Z，SUN W，et al. An Automatic Geocoding Algorithm Based on Address Segmentation [J]. Bulletin of Surveying and Mapping，2011，4(2)：59-62.

[11] TIAN Q，REN F，HU T，et al. Using an Optimized Chinese Address Matching Method to Develop a Geocoding Service：A Case Study of Shenzhen，China [J]. ISPRS International Journal of Geo-Information，2016，5(65)：1-17.

[12] WEI J，ZHONG Z. An Approach to Address Matching Based on Confidence [J]. Science of Surveying and Mapping，2015，40(1)：122-125.

[13] HUANG K，MA S. Chinese Web Page Classification Based on Statistical Word Segmentation [J]. Journal of Chinese Information Processing，2002，16(6)：25-31.

[14] XIAO J. Method of Recognition and Match of Place Name Based on Statistic [J]. Journal of Geomatics Science and Technology，2014，31(4)：408-412.

[15] SONG Z. Address matching algorithm based on chinese natural language understanding [J]. Journal of Remote Sensing，2013，17(4)：788-801.

[16] MA L，GONG J. Application of Spatial Information Natural Language Query Interface [J]. Geomatics and Information Science of Wuhan University，2003，28(3)：301-305.

[17] ZHANG X. A knowledge-based agent prototype for Chinese address geocoding [C]// Geoinformatics 2008 and Joint Conference on GIS and Built environment：Advanced Spatial Data Models and Analyses. 2008：1-10.

[18] JING Z，QI L. Research on the application of geocoding [J]. Geography and Geo-Information Science，2003，3(19)：22-25.

[19] QIN B，WANG Q Y，LI C. Effective Strategy for Sensitive Analysis of Bayesian Networks[J]. Journal of Chinese Systems，2016，37(4)：732-737.

[20] GE S，XIA X. An Intelligence Decision Model Based on Probabilistic Influence Analysis [J]. Computer Engineering，2016，42(6)：213-217.

[21] PEARL J. Fusion，propagation，and structuring in belief networks [J]. Artificial Intelligence，1986，29(3)：241-288.

[22] YAO X，LI X，PENG L. A Novel Fuzzy Chinese Address Matching Engine Based on Full-text Search Technology [C]// Proceedings of Science. 2015：1-9.

（上接第 255 页）

[12] SUN G M，WANG S. Compute adaptive fast recommendation algorithm statisfied user interest drift[J]. Application Research of Computers，2015，32(9)：2669-2673. (in Chinese)
孙光明，王硕. 满足用户兴趣漂移的计算自适应快速推荐算法[J]. 计算机应用研究，2015，32(9)：2669-2673.

[13] LI C，LIANG C Y，MA L. A Collaborative filtering recommendation algorithm based on Domain nearest neighbor[J]. Journal of Computer Research and Development，2008，45(9)：1532-1538. (in Chinese)
李聪，梁昌勇，马丽. 基于领域最近邻的协同过滤推荐算法[J]. 计算机研究与发展，2008，45(9)：1532-1538.

[14] WANG X M，ZHANG X M. Collaborative recommendation algorithm based on contribution factor[J]. Application Research of Computers，2015，32(12)：3551-3554. (in Chinese)
王兴茂，张兴明. 基于贡献因子的协同过滤推荐算法[J]. 计算机应用研究，2015，32(12)：3551-3554.

[15] LEE H C，LEE S J，CHUNG Y J. A Study on the Improved Collaborative Filtering Algorithm for Recommender System[C]// Acis International Conference on Software Engineering Research，Management & Applications. IEEE Computer Society，2007：297-304.

[16] DEAN J，GHEMAWAT S. MapReduce：simplified data processing on large clusters [J]. Communications of the ACM，2008，51(1)：107-113.

[17] ZHAO Z D，SHANG M S. User-Based Collaborative-Filtering Recommendation Algorithms on Hadoop [C]// International Conference on Knowledge Discovery and Data Mining. IEEE，2010：478-481.