

大数据下的多源异构知识融合算法研究

张 瑶,李蜀瑜,汤 玥

(陕西师范大学 计算机科学学院,陕西 西安 710119)

摘 要:在大数据环境下,多源异构知识的融合为研究者从众多分散、异构的数据源和知识源中挖掘出隐含的、有价值的和尚未被发现的信息和知识提供了非常有效的手段和方法。针对目前知识融合方法的不足,在对大数据环境下的异构知识融合方法进行深入研究的的基础上,将已有的数据融合算法合理地移植到知识融合中,设计并构造了大数据环境下的多源异构知识融合算法。为进一步提高获取知识的质量,依据知识源粒度的动态选择,提出了一种改进的知识源分解-合并算法,以获得合适粒度大小的知识源集合和尽可能真实可靠的知识。基于 Hadoop 和 MapReduce 框架所构建的实验平台对所提算法进行了实验验证。实验结果表明,所提出的多源异构知识融合算法有效可行,并能够有效显著地提高多源异构知识融合算法的性能。

关键词:大数据;多源异构知识;知识融合;融合算法

中图分类号:TP302

文献标识码:A

文章编号:1673-629X(2017)09-0012-05

doi:10.3969/j.issn.1673-629X.2017.09.003

Research on Heterogeneous Knowledge Fusion Algorithm under Big Data Environment

ZHANG Yao, LI Shu-yu, TANG Yue

(College of Computer Science, Shaanxi Normal University, Xi'an 710119, China)

Abstract: In environment of big data, the integration of multi-source heterogeneous knowledge fusion has provided one of the most effective means and methods for researchers to discover the implicit, valuable and undetected knowledge from a lot of knowledge sources that are dispersed and heterogeneous. Aimed at the shortcomings of the current knowledge fusion methods, based on investigations on them under the big data environment, the existing data fusion methods have been employed, which are transplanted to the knowledge fusion reasonably. A kind of algorithm for multi-source heterogeneous knowledge fusion is proposed. In order to further improve the quality of the acquiring knowledge, an improved algorithm based on the dynamic selection of knowledge source granularity is proposed to obtain the appropriate size of the collection of knowledge sources and the true and reliable knowledge as possible. Its experimental verification is conducted based on the experimental platform constructed by Hadoop and MapReduce framework. Experimental results show that it is effective and feasible and effectively improves the performance of multi-source heterogeneous knowledge fusion algorithms.

Key words: big data; multi-source heterogeneous knowledge; knowledge fusion; fusion algorithm

0 引言

在如今的大数据时代,数据的种类越来越多,数据的规模日益增大。在数据这片汪洋大海中,人们往往不知所措,从多而杂的数据中抽取出有较高利用价值的知识的需求也变得更加迫切。这不仅是企业界也是学术界重点关注的话题^[1]。在大数据环境下,人类对知识服务的探究,已经不仅仅局限于传统的信息和文献服务,而是将研究的眼光更多投放在用户的行为、数

量庞大的碎片化信息、用户之间的关系以及由此而形成的海量的具有实时性的数据、机器数据和非结构化数据等方面^[2]。知识服务的意义和内容,将在大数据的推进下不断发生变化,它将更多地面向知识的不断创新和人类对知识的各方面需求,逐渐转变为知识预测型的服务,将大数据转变为真正的大智慧。

知识融合是基于信息融合发展而成的一个新概念。多源异构知识是由知识自身不断丰富、发展、创

收稿日期:2016-10-17

修回日期:2017-01-20

网络出版时间:2017-07-11

基金项目:国家自然科学基金资助项目(41271387)

作者简介:张 瑶(1992-),女,硕士研究生,研究方向为移动云计算、大数据安全等;李蜀瑜,硕士生导师,副教授,博士,研究方向为移动云计算、大数据安全等。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20170711.1455.060.html>

新、演化而成。多源异构知识融合自身的价值就在于从众多分散、异构的数据源、知识源中挖掘出隐含的、有价值的、尚未被发现的信息和知识(如规则、方法、模型、约束、经验等)。知识融合实现的关键在于融合方法,直接影响融合后知识的内涵、层次以及置信度。

在大数据环境下,由于数据的结构差异大、数据来源广、价值密度较低、更新实时等特点,给知识服务带来了巨大挑战,而多源异构知识的融合为研究者在大 数据环境下进行知识获取、知识组织和利用提供了非常有效的手段和方法。目前的知识融合方法从理论到实践还有很多不足,为此,就大数据环境下的异构知识融合方法展开进一步的深入研究,借鉴数据融合方法,提出了多源异构知识融合算法,并基于知识源的粒度给出了一种改进方法,同时还进行了实验验证。

1 知识融合相关算法研究

目前关于知识融合还没有一个统一的定义,知识融合的发展是建立在信息融合的基础之上的,在最早的时候,人类关于知识融合的研究大多是将它当作知识工程的一个分支,并且和其他有关的内容结合起来。知识融合的研究内容与信息融合的研究内容有重合部分,所以,在研究知识融合时可以参考信息融合的相关研究结果^[3]。

知识融合算法是知识融合的核心部分。目前,研究人员已经提出了有关知识融合的算法,除知识融合的评价算法外,其他的分别为基于 D-S 理论^[4]、模糊集理论^[5]、主题图^[6]和语义规则^[7]的知识融合算法。

基于 D-S 证据理论的知识融合算法^[8]是由韩立岩提出的,该方法首先进行数学建模,然后实现融合算法,最后对融合结果进行分析预测。但是这种方法会受到单一故障假设的条件限制。姚路等针对这一不足,提出一种将 DSmT 与系统建模相结合的知识融合算法^[9]。周芳等利用模糊集理论解决知识融合问题,基于 Petri 网提出了知识融合的一般模型,并详细介绍了融合模型中的每个步骤,将知识融合算法应用到实际的企业相关问题中^[10]。鲁慧民等在全信息理论^[11]的基础上,通过联合扩展主题图自身的优点,提出了基于扩展主题图相似性算法(ETMSC)^[12]。该算法是针对多源知识融合的,与此同时,提出了层次之间相互对应、阈值选取以及实验确定这三个基本原则。该算法在进行相似性计算时,综合考虑了语用、语义、语法、知识的含义和知识所处的语义环境。

这些知识融合算法面向的应用知识都是有针对性的,其中,基于 D-S 理论的更加侧重关于专家知识的融合,基于语义的则重点是研究非专家的知识融合,基于主题图的主要研究的是专家知识和非专家知识的融

合。目前已有的知识融合算法虽然考虑到知识来源的多样性,但是具体对每个知识的结构分析得不够清楚,而且还有一点不足是没有考虑到源知识本身的可靠性和真实性。从考虑知识的真实概率的角度出发,结合大数据环境,提出了一种多源异构知识融合算法。

2 多源异构知识融合算法

由于知识融合是从不同知识源,如 Freebase、YAGO 等公开的知识库以及互联网网页,抽取知识获得知识三元组,求得知识三元组的真实概率,以做出最佳决策,提供更好的知识服务。而数据融合是解决从不同来源的值,并寻找数据真值的问题。因此,基于知识融合本身的特点,借鉴已有的数据融合算法,将其合理地移植到知识融合中,构造大数据环境下的多源异构知识融合算法。

2.1 多源异构知识融合面临的挑战

数据融合是对从不同来源的数据、信息,加以联合、相关、组织,寻找数据真值。与数据融合相比,对知识融合提出了三大挑战。

(1)数据融合的输入为一个二维数据矩阵,如图 1(a)所示;而知识融合的输入是一个三维矩阵,如图 1(b)所示。新增的一维表示提取器,所以矩阵中的每个单元格表示用相应的抽取器从对应的 Web 源中提取的相应数据项的值。错误在这个过程的每个阶段都有可能发生,不仅来自于 Web 源,也可能来自于提取过程中三元组的识别、实体连接和属性连接。

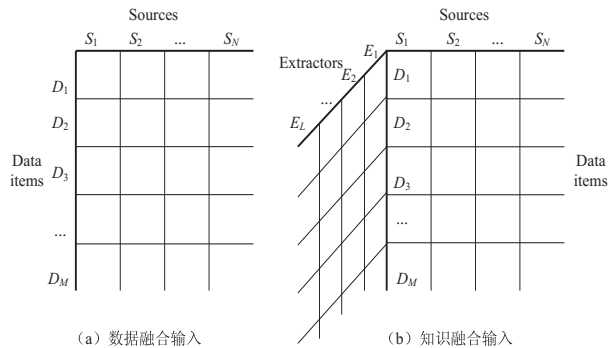


图1 数据融合和知识融合的输入

(2)希望预测概率可以正确地反映三元组真实的可能性。一个基本要求就是单调:具有较高预测概率的三元组应该比一个具有较低预测概率的三元组的真实概率要大些。

(3)知识的规模通常是巨大的。当前在数据融合实验中使用的最大数据集包含 170 K 数据源,400 K 的数据项。知识融合往往需要处理的数据的数量级在各方面都会更大。

2.2 融合方法选取标准

现有的数据融合方法可以用来解决知识融合的问题。

题。采用了三个标准,从现有的方法中选择合适的数
据融合方法。

(1) 由于知识融合的目标是计算每个三元组的真
实概率,选择的数据融合方法,可以很容易地求出一个
有意义概率。

(2) 由于知识融合的数据规模比传统的数据融合
的数据规模要大三个数量级,选择能按比例放大的基
于 MapReduce^[13] 框架的方法。

(3) 重点放在那些最近研究表明更有效的方法。
例如,文献[14]表明基于贝叶斯方法更优于基于 Web
链路等方法。

2.3 多源异构知识融合方法

按照上述三个标准,选择了三种数据融合方法:
VOTE, ACCU 和 POP ACCU。下面对这三种方法进行
简单的介绍,然后再描述如何使用这三种方法来解决
知识融合问题。

VOTE: 对于每个数据项, VOTE 统计每个值的数
据来源的个数,并且信任来自最多数据源的值。VOTE
作为实验的基准。

ACCU: 采用的是贝叶斯分析方法。算法伪代码如
图 2 所示。对于每一个提供一组值 V_s 的数据源 S , S
的准确度是 V_s 中所有值的平均概率。对于每个数据
项 D 和由 D 提供的值的集合 V_D , 一个值的概率是使
用贝叶斯分析观测其先验概率计算所得。ACCU 假
定: 对于每个数据项 D 只有一个真值; 有 N 个均匀分
布的假值; 数据源之间是相互独立的。

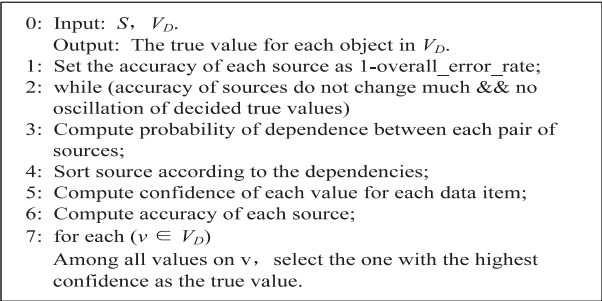


图 2 ACCU 算法

POP ACCU: POP ACCU 通过去除错误的值使均匀
分布的假设扩展了 ACCU; 它从真实数据中计算得出
分布并将其插入到贝叶斯分析中。文献[15]已经证
明 POP ACCU 是单调的, 也就是说在假设数据源和数
据项都是独立的条件下, 增加一个数据源不会降低数
据融合的质量。

2.4 多源异构知识融合体系

采用以上三种数据融合方法解决知识融合问题。
首先, 数据融合方法的输入是二维数据矩阵, 每个
数据源提供相应数据项的值, 而知识融合方法的输入
是三维矩阵, 包含每个数据源通过相应的抽取器抽取

得到的对应数据项的值。为了减小知识融合输入的维
度, 考虑将每对(抽取器, URL)作为数据源。有大量的
数据源表明一个知识三元组不是由 Web 源提供的, 就
是由许多不同的抽取器抽取获得的。

其次, 数据融合方法的输出是由每个提供的值的
二元决策构成的, 而知识融合方法的输出是每个知识
三元组的真实概率。对于 ACCU 和 POP ACCU, 通过
贝叶斯分析计算获得每个知识三元组的真实概率。对
于 VOTE, 采取的计算概率的方法如下: 如果一个数据
项 $D = (s, p)$ 总共有 n 个出处, 一个知识三元组 $T = (s,$
 $p, o)$ 有 m 个出处, 则知识三元组的真实概率为 $p(T) =$
 m/n 。

最后, 使用基于 MapReduce 的框架来扩展上述三
种方法。知识融合的体系结构如图 3 所示。一共有三
个阶段; 每个阶段是一个 MapReduce 的过程, 因此以
并行的方式进行。

第一阶段: Map 步骤是根据相关的数据项将输入
所提取的知识三元组进行划分; Reduce 步骤是运用贝
叶斯分析方法推导并计算出由相同数据项提供的每个
知识三元组的真实概率。

第二阶段: Map 步骤将已经由出处获得的概率的
知识三元组进行划分; Reduce 步骤是依据出处所包含
的知识三元组来计算它的准确度。重复前两个阶段直
至收敛。

第三阶段: Map 步骤是划分所提取的知识三元组;
Reduce 步骤是将由不同出处得到的相同的知识三元
组进行去重, 第三阶段输出最终结果。

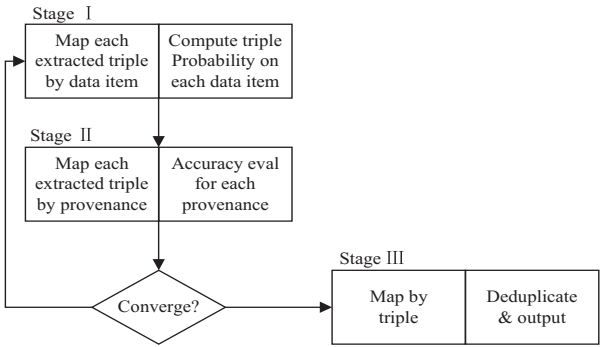


图 3 MapReduce 实现 ACCU 和 POP ACCU

3 多源异构知识融合方法的改进

针对 ACCU 和 POP ACCU 的融合方法, 从知识源
的质量角度出发, 提出一种改进算法。该算法可以动
态选择知识源的粒度大小, 得到合适粒度大小的知识
源集合, 作为以上融合算法的输入参数, 以提高知识三
元组真实概率的准确度和有效性。

理想情况下, 希望用最好的粒度大小知识源。例
如, 由于一个网页可能与其他网页有不同的精确度,

所以很自然地将每个网页看作是一个独立的源。甚至可以定义一个源作为在特定网页上的特定谓词,这样可以估算一个关于特定种类的谓词的网页可信度。然而,当定义来源过于精准的话,可能有太少可靠的数据来估算它们的准确度;相反,可能存在一些数据源,它们有太多的数据都在最后的粒度上,这样可能会导致计算瓶颈。

为了解决这个问题,需要动态选择知识源的粒度。对于粒度过小的知识源,可以在层次结构上回退到比较粗糙的级别,使得可以借用相关页面之间的统计强度。对于粒度过大的知识源,可以选择将其拆分成多个知识源,然后独立地评估它们的准确度。当做归并时,目标是在不降低效率的条件下提高评估的统计质量。当做分解时,目标是在没有显著改变评估结果的前提下有效提高数据偏斜。

为了使效果更精准,把知识源定义为一个特征向量:<网站,谓词,网页>,并按照最一般到最特殊的情况进行排序。然后在一个层次结构上安排这些知识源。例如,<wiki.com>是<wiki.com,date of birth>的父亲,而<wiki.com,date of birth>是<wiki.com,date of birth,wiki.com/page1.html>的父亲。定义以下两种操作:

分解:当分解一个较大的知识源时,希望可以将其随机分解为大小相似的子知识源。具体就是,令一个大小固定的知识源 M ,是期望的最大尺寸,将三元组均匀分布到大小小于最大尺寸的桶,每个桶代表的是一个子知识源。将 M 设置为一个比较大的值,这样那些不需要分解的知识源就不会分解,同时不会导致计算瓶颈。

合并:当合并小的知识源时,希望只合并那些有共同特征的知识源,例如这些知识源共有相同的谓词,或者来自于相同的网站。因此,在结构层次上只合并那些有相同的根源或属于同一个分支的子源,将其设置为一个很小的值,这样能降低合并的范围,不需要合并的知识源就不会合并,同时还保持足够的统计强度。

例如,考虑以下三个知识源:<website1.com,date_of_birth>,<website1.com,place_of_birth>,<website1.com,gender>,每个只含有两个特征,不足以用来进行质量评估。可以通过移除第二特征把它们合并到它们的父亲源,然后得到一个大小一定知识源<website1.com>,该知识源可以为质量评估提供很多的知识。

有两种情况需要考虑:一是当合并了小的知识源,但是得到的父亲源可能并没有期望的大小,它可能还是太小,这时,需要反复迭代合并父亲源,以达到期望的大小;二是当合并的结果过于庞大,大大超出了期望的大小,这时就要将这些合并的源再做分解。用来动

态选择知识源的粒度大小的知识源分解-合并算法(SplitAndMerge)的伪代码如下:

输入: S 为具有最好粒度的知识源; m/M 为期望的最小/最大知识源的大小。

输出: S' 为一个具有期望大小的知识源的集合。

Begin

```
1:  $S' \leftarrow \emptyset$  ;//将最终知识源初始化为空集
2: For  $S \in S$  do;//遍历知识源集合中的每个知识源
3:  $S \leftarrow S \setminus \{S\}$  ;
4: If  $|S| > M$  then //知识源大小大于期望的最大值
5:  $S' \leftarrow S' \cup \text{SPLIT}(S)$  ;//进行知识源分解操作
6: else if  $|S| < m$  then //知识源大小小于期望的最小值
7:  $S_{\text{par}} \leftarrow \text{GETPARENT}(S)$  ;//进行知识源合并操作
8: if  $S_{\text{par}} = \perp$  then;//已经到达了层次结构的顶部
9:  $S' \leftarrow S' \cup \{S\}$  ;
10: else
11:  $S \leftarrow S \cup \{S_{\text{par}}\}$  ;//继续迭代合并
12: else
13:  $S' \leftarrow S' \cup \{S\}$  ;
14: Return  $S'$  ;//输出最后得到新的知识源集
```

4 实验结果与分析

利用文献[16]中的知识抽取方法获取实验数据,抽取结果如表1所示。

表1 知识抽取结果及抽取质量

	#Triples	#Webpages
TXT1	274 M	202 M
TXT2	31 M	46 M
TXT3	8.8 M	16 M
TXT4	2.9 M	1.2 M
DOM1	804 M	344 M
DOM2	431 M	925 M
DOM3	45 M	N/A
DOM4	52 M	7.8 M
DOM5	0.7 M	0.5 M
TBL1	3.1 M	0.4 M
TBL2	7.4 M	0.1 M
ANO	145 M	53 M

所涉及的知识一部分是来源于已有的一些高品质的知识库,如 Freebase、YAGO 等,另一部分是来自于互联网上的最新知识。使用 Hadoop 构建知识融合的实验平台。另外,为了更好地评估所提出的多源异构知识融合方法中不同算法的性能,利用大型 Matlab 对几组数据进行处理,比较模块化度并绘制相应的结果。

首先,给出评价实验结果的一个指标:校准曲线。校准曲线绘制的是预测概率与真实概率之间的变化。为了计算真实概率,把知识三元组分成 $l+1$ 桶:第 i ($0 \leq i \leq l-1$) 桶包含预测概率在 $[i/l, (i+1)/l)$ 知识三元组,第 $l+1$ 桶包含概率为 1 的知识三元组。实验中设 $l=20$,然后计算每个桶的真实概率。理想的

情况是预测概率应该与真实概率相同,这样的理想曲线是由(0,0)到(1,1)。

利用多源异构知识融合方法得到知识三元组真实概率的结果和预测结果,绘制了校准曲线,如图 4 所示。结果显示应用 POP ACCU 的多源异构知识融合算法的结果最贴近理想曲线,效果最好。

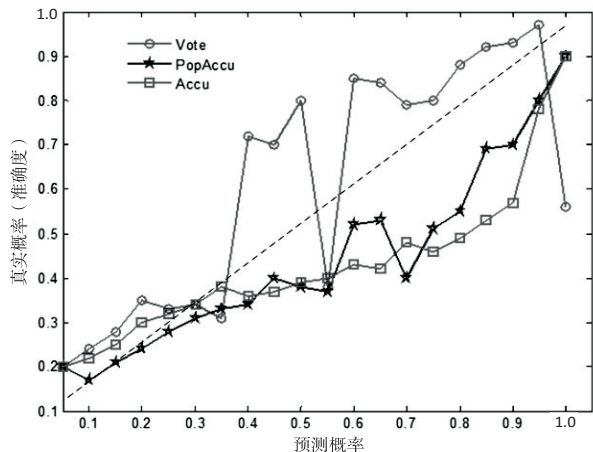


图 4 不同融合方法的校准曲线

再将抽取获得的结果按照改进方法求得知识三元组的真实概率,并绘制校准曲线,如图 5 所示。结果显示,改进算法确实可以在一定程度上提高多源异构知识融合算法的性能。

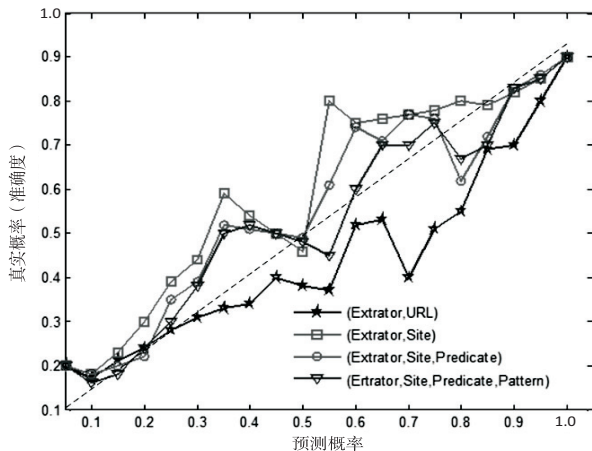


图 5 改进后的校准曲线

5 结束语

多源知识融合是对知识进行融合、处理,进而提高知识的内涵、品质、置信度。针对目前知识融合方法的不足,结合大数据背景,借鉴数据融合算法,提出了一种多源异构知识融合算法,以求出知识三元组的真实概率,并依据知识源的粒度提出了相应的改进算法。应用 Hadoop 构建实验平台,并基于 MapReduce 框架,实现了多源异构知识融合算法,并对改进方法进行了实验验证。实验结果表明,改进算法可以有效提高多源异构知识融合算法的性能。

参考文献:

- [1] 苏苏宁. 面向知识服务的知识组织理论与方法[M]. 北京: 科学出版社, 2014.
- [2] 唐晓波, 魏巍. 知识融合: 大数据时代知识服务的增长点[J]. 图书馆学研究, 2015(5): 9-14.
- [3] 缙锦. 知识融合中若干关键技术研究[D]. 杭州: 浙江大学, 2005.
- [4] Valin P, Djiknavorian P, Bosse E. A pragmatic approach for the use of Dempster-Shafer theory in fusing realistic sensor data[J]. Journal of Advances in Information Fusion, 2010, 5(1): 32-40.
- [5] Werro N. Fuzzy set theory[M]//Fuzzy classification of online customers. [s. l.]: Springer International Publishing, 2015: 7-26.
- [6] Lu J, Ma J, Zhang G, et al. Theme-based comprehensive evaluation in new product development using fuzzy hierarchical criteria group decision-making method[J]. IEEE Transactions on Industrial Electronics, 2011, 58(6): 2236-2246.
- [7] Okoye K, Tawil A R H, Naeem U, et al. A semantic rule-based approach towards process mining for personalised adaptive learning[C]//High performance computing & communications, IEEE international symposium on cyberspace safety & security, IEEE international conference on embedded software & systems. [s. l.]: IEEE, 2014: 929-936.
- [8] 韩立岩, 周芳. 基于 D-S 证据理论的知识融合及其应用[J]. 北京航空航天大学学报, 2006, 32(1): 65-68.
- [9] 姚路, 康剑山, 曾斌. 结合 DSMT 理论和系统建模的知识融合算法[J]. 火力与指挥控制, 2014, 39(12): 88-91.
- [10] 周芳, 刘玉战, 韩立岩. 基于模糊集理论的知识融合方法研究[J]. 北京理工大学学报: 社会科学版, 2013, 15(3): 67-73.
- [11] 何华灿. 人工智能基础理论研究的重大进展-评钟义信的专著《高等人工智能原理》[J]. 智能系统学报, 2015(1): 163-166.
- [12] 鲁慧民, 冯博琴, 李旭. 面向多源知识融合的扩展主题图相似性算法[J]. 西安交通大学学报, 2010, 44(2): 20-24.
- [13] Odia T, Misra S, Adewumi A. Evaluation of Hadoop/MapReduce framework migration tools[C]//Asia-Pacific world congress on computer science and engineering. [s. l.]: IEEE, 2015: 1-8.
- [14] Li X, Dong X L, Lyons K, et al. Truth finding on the deep web: is the problem solved? [J]. Proceedings of the VLDB Endowment, 2012, 6(2): 97-108.
- [15] Dong X L, Berti-Equille L, Srivastava D. Truth discovery and copying detection in a dynamic world[J]. Proceedings of the VLDB Endowment, 2009, 2(1): 562-573.
- [16] Reuss P, Althoff K D, Henkel W, et al. Semi-automatic knowledge extraction from semi-structured and unstructured data within the OMAHA project[C]//International conference on case-based reasoning. [s. l.]: Springer International Publishing, 2015: 336-350.