基于链接的作者重名处理方法研究与应用*)

吴 斌 徐超群 王文彬 吴 巍

(北京邮电大学智能通信软件与多媒体北京市重点实验室 北京 100876)

摘 要 本文提出了一种适用于中文文献索引数据的实体解析方法。区别于新近的属性+链接结构的聚类方法,本方依据实际问题的特点采用了先属性匹配,然后基于文献合作网络的结构解析的策略。该方法在真实的大数据量文献索引数据上运行获得了良好的效果,并已经运用于数字图书馆的检索系统。

关键词 实体解析,重名分析,信息检索,基于链接的分析

A Link-based Method for Name Disambiguation and its Application

WU Bin XU Chao-Qun WANG Wen-Bin WU Wei

(Beijing Key Laboratory of Intelligent Communications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876)

Abstract The research presents an algorithm that is applicable to name disambiguation of Chinese literature digital library. Differ from the clustering method which considered the node attribute and link structure simultaneity, differ from the state-of-the-art LDA-ER method which employ the LDA model to resolute entities, differ from the DistQC model for resolving name disambiguation, our dedicated algorithm firstly makes an attribute similarity analysis and then detects the reference-entity relationship by considering the links of co-authorship network and the collective relations among the co-authorship network. The algorithm performs better on two datasets from a digital library and has a wide application in digital library.

Keywords Entity resolution, Name disambiguation information retrieval, Link-based analysis

在大规模的文献索引数据库中,我们发现由于记录信息的局限性,无法为每篇文献的作者找到对应真实作者的唯一主键。由于中国作者的姓名只有单一的表述方式(即姓在前,名在后,且没有缩写形式),因此我们可以把针对中文文献索引数据的实体解析问题认为是"重名解析"问题,即多个作者对应同一姓名标识,如"朱东华"可能先后被来自淮南师范学院的"朱东华"或来自北京理工大学经济管理学院的"朱东华"同时使用。

本文提出了一种新的实体解析方法 NDC(Name Disambiguation for Chinese),这种方法专门针对于中文文献索引数据库中的重名问题。NDC 方法的主要贡献在于:

- NDC 是专门针对中文文献索引数据设计的重名分析方法,对中文字符串的匹配处理进行了优化;
- NDC 能够得到理想的 F1 值,在中文的特殊环境中运行所得到的效果优于国外同类方法;
- NDC 具有出色的执行效率,使之能够胜任大数据量的实际应用。

1 相关工作

自从 1969 年 Fellegi and Sunter[1] 提出 Fellegi and Sunter 模型以来,很多学者在统计学、人工智能、数据库等领域开展了实体解析方面的研究。其中比较有影响力的方法有姓名匹配,贝叶斯算法[2],记录链[3,4], 概率模型[12]等等。其中这些方法主要集中在姓名的相似性程度比较。Indrajit 等作者在聚类过程中将节点属性、链接属性以及链接结构等因素作为相似性衡量标准来挖掘图结构中数据实体的方法 GBC-ER[5],采用聚类方法将属性相似性与链接相似性通过加权公式来将两者的效果"并行"加入到聚类过程中来进行实体的解析。Lise Getoor 等人利用 LDA[6] 模型来进行实体解析的方法 LDA-ER[7],Byung 等人利用 DistQC 取得了很好的效果[8]。

2 问题描述

在我们实验的文献索引数据库中,其原始数据信息以论文为单位,记录了论文题名、作者、作者单位、关键词、发表期刊等信息。

下面我们定义几个变量:

标识:从原始数据所抽取出的合作网络的节点,是作者信息的载体,用字母"r"表示;

实体:作者标识所对应的真实的作者,与标识构成一对多的关系,用字母"e"表示;

属性:描述作者的属性,比如地址,关键词等,用字母"a"表示,如 $r_i.a_j$ 表示 r_i 标识的 a_j 属性 ;$r_i \equiv r_j$:表示标识 r_i 与 r_j 对应同一个实体。

我们把问题定义为:

对于给定的作者标识集合 S_r,合作关系集合 $S_c=\{(r_i, r_j)\}$,对于 S_r 中具有某些属性相同、相近或者相似的作者标识 r_i, r_j,求集合 $S_e=\{e_i\}$,其中 S_e 是一个实体集合,其中 e_i 表示与作者标识相对应的实体。

3 重名分析方法

NDC 方法主要由三步组成:第一步为属性匹配,通过对相同标识的作者信息的属性进行匹配,合并匹配成功的标识。第二步为链接分析,通过对作者合作网的结构分析,可以发现具有相同标识的作者与同一个作者实体合作,那么这些具有

*)基金项目:国家自然科学基金项目(60402011)和国家科技支撑计划项目(2006BAH03B05)。

相同标识的作者标识对应同一个作者实体。第三步为协同实体分析，在前两部分析处理的基础上，把具有相同合作对的信息加以分析合并，该步与链接分析交叉进行，根据实验统计分析，当链接分析与协同实体分析交互进行4次时，就几乎不再影响F1了。

```
输入：S_r
输出：S_e
定义函数：AttributeMacth(String s1, String s2)表示属性匹配
          LinkAnalysis(S_e, r_i, r_j)表示链接分析
          CollectiveAnalysis(S_e, r_i, r_j)
步骤：for each author in S_r
      { if (isNotExistIn(S_e))
  add(S_e);
  else {findDuplicatedName in S_e
        AttributeMacth(String s1, String s2) } }
  for each entity in S_e
  { LinkAnalysis(S_e, r_i, r_j);
    CollectiveAnalysis(S_e, r_i, r_j); }
```

图1　重名分析方法

下面分别介绍NDC方法中的主要算法：

3.1　属性匹配

针对NAME属性相同(鉴于中文姓名的书写规则单一性，我们认为只有姓名相同的标识才可能对应同一个作者实体，因此本文接下来的篇幅中所有作者标识匹配的前提都是姓名相同，且不再专门指出)的作者标识的其他相关属性进行匹配，对于匹配成功的标识认为是对应同一个作者实体。

对于字符串的匹配算法，采用文[9,10]中提出的字符串匹配算法 *SoftTFIDF*，对于两个作者标识 r_i，r_j，匹配函数 $Match(r_i, r_j, R_s)$ 定义如下：

$Match_1(r_i, r_j, R_s) = \sum SoftTFIDF(r_i.a_i, r_j.a_i) \geqslant R_s$，上式的等号两边均为布尔值，$Match$=true表示匹配成功，反之则失败。右侧是先针对 r_i 和 r_j 求出 *SoftTFIDF* 函数(该函数的值域为[0,1]，值越接近1则表示输入的两个字符串越相似)，并将 *SoftTFIDF* 函数的输出与一个阈值 R_s 进行比较，若大于该阈值则返回true，否则返回false。也就是说，若两个标识的属性通过 *SoftTFIDF* 函数的输出大于等于 R_s，则表示匹配成功。

3.2　基于链接的重名分析

我们对作者标识合作网的链接情况也进行了分析，并将其作为解析作者实体的一个依据。考虑这么一种情况：标识 A 与标识 B 姓名相同，且他们都曾与标识 C 合作过，那么我们就认为标识A与标识B对应了同一个作者实体[11,13,16](当然，这只是一个凭经验的假设，是否能够辅助进行重名分析将在实验中得到验证)，即：

若 $\exists r_x$，使得 $(r_i, r_x) \in S_c$，$(r_j, r_x) \in S_c$

且 $r_i.NAME = r_j.NAME$，则 $r_i \equiv r_j$

我们定义了一个基于链接分析的算法：

```
输入：标识合作网络 S_c, r_i, r_j
输出：r_i≡r_j
定义函数：
    CoAuthor(r_i) 该函数表示所有与标识 r_i 合作过的标识，即在合作网络中与 r_i 相邻的标识集合。
步骤：
    S_1=Φ; S_2=Φ;
    for distance = 1 to 2 {
        if distance mod 2=1 { add CoAuthor(r_i) to S_1; }
        else{ add CoAuthor(r_j) to S_2; }
    if S_1 ∩ S_2≠Φ return true; }
        return false;
```

图2　链接分析算法

3.3　协同实体分析

同时，通过统计发现，在上面提到的两个中文文献数据库中，有95.84%以上文章的作者都是由2个或者2个以上的合作者撰写而成。考虑这么一种情况，标识 A 和标识 B 合作写过一篇或几篇文章，同时与 A 重名的标识 C 和与 B 重名的标识 D 也合作写过一篇或几篇文章。那么，我们将计算 A 与 C 的，B 与 D 的相似性，根据实验发现，如果 A 与 C 以及 B 与 D 的相似性都大于0.5，那么 A 与 C，B 与 D 均对应同一个实体[14,15]。协同实体分析算法定义如下：

```
输入：所有作者合作对集合 blist;
输出：合并了相同作者合作对的合作对集合 alist;
定义函数：collectiveEntityResolution(coAuthorPair, theEqualPair),
该函数用来判断两对具有相同姓名的作者对是否属于同一个作者对。
find(coAuthorPair)，该函数用来判断集合中是否包含 coAuthorPair,
如果包含则返回与 coAuthorPair 内容相同的合作对。
步骤：for each coAuthorPair in the blist
      {if(auAuthorPair not exists in alist)
       add each coAuthorPair in to alist
       else
       theEqualPair = find(CoAuthorPair);
if(! collectiveEntityResolution(coAuthorPair,theEqualPair))
     add coAuthorPair in to the alist; }
```

图3　协同实体分析算法

3.4　标识合并

当我们发现两个或更多标识对应同一实体时，我们可以将其合并，合并后的合作网络将被压缩，剩余的标识将更具代表性(能够一对一地代表实体的标识数量将会增加)，同时，合并过程也将对后续迭代进行的重名分析产生积极的影响。合并标识包括以下几个步骤：选择代表标识；合并标识的论文集合；合并标识的合作标识集合；计算合并后代表标识的合作者数量与论文数量属性。先定义函数如下：

$$Captain(r_i, r_j) = \begin{cases} r_i, r_i.ARTICLE\_NUMBER > r_j.ARTICLE\_NUMBER \\ r_j, r_i.ARTICLE\_NUMBER < r_j.ARTICLE\_NUMBER \\ r_i, (r_i.ARTICLE\_NUMBER = r_j.ARTICLE\_NUMBER) \wedge \\ \quad (r_i.DEPT.LENGTH \geqslant r_j.DEPT.ELNGTH) \\ r_j, (r_i.ARTICLE\_NUMBER = r_j.ARTICLE\_NUMBER) \wedge \\ \quad (r_i.DEPT.LENGTH < r_j.DEPT.ELNGTH) \end{cases}$$

算法描述如下：

```
输入：两个需要合并的标识 r_i, r_j, 合作网络 S_c
输出：合并后的标识 r_c
定义函数：Captain(r_i, r_j)
步骤：r_c = Captain(r_i, r_j);
add r_i.ARTICLE_SET to r_c.ARTICLE_SET;
add r_j.ARTICLE_SET to r_c.ARTICLE_SET;
if r_c=r_i{ ∀(r_j, r_x)∈S_c add (r_i, r_x) to S_c;
            ∀(r_x, r_j)∈S_c add (r_x, r_i) to S_c; }
if r_c=r_j{ ∀(r_i, r_x)∈S_c add (r_j, r_x) to S_c;
            ∀(r_x, r_i)∈S_c add (r_x, r_j) to S_c; }
```

图4　标识合并算法

4　实验描述

本实验采用的两个数据集均来自某大型中文文献索引数据提供。其中一个是包括194,203条论文记录的医药文献索引数据集(以下用dataset1指代该数据集)，另一个是包括358,049条论文记录的冶金文献索引数据集(以下用dataset2指代该数据集)。

4.1　建立测试数据集

为了测试算法的正确率与涵盖率，我们采用人工解析的方法来标注标识与实体的对应关系，并将此对应关系作为测试的依据。我们选择了如下6个作者姓名作为标准测试数据集：

表1 测试数据集

	姓名	标识数目	实体数目	最多和作者数	文章数目
Data set1	叶章群	89	3	164	87
	徐如祥	89	1	156	89
	李莉	438	238	23	3
Data set2	王国栋	73	10	104	57
	刘相华	61	1	93	52
	王芳	99	70	6	5

“叶章群”被入选是因为某个“叶章群”的合作者数量为164，在从dataset1处理得到的S_r中是最大的。而“徐如祥”被入选则是因为某个“徐如祥”的论文数量是89，在从dataset1处理得到的S_r中排名第一，但是因为排第一的“叶章群”已经入选，因此“徐如祥”替补入选。可以认为，“叶章群”和“徐如祥”是dataset1中著名作者的代表。“李莉”的入选是因为NAME为“李莉”的标识数量较大，可以认为是一个被广泛采用姓名的代表。同理，“王国栋”、“刘相华”、“王芳”是从dataset2处理得到的S_r中选出的三个姓名。

4.2 实验方法

NDC方法先考虑节点属性进行匹配(AM, Attribute Match)，并将匹配后认为对应同一实体的标识进行合并；之后在第一次合并的基础上再基于链接结构进行分析(LBA, Link-Based Analysis)，然后对网络中单独存在的各种小团体进行协同实体解析(Collective Entity Resolution)并对找到的对应同一实体的标识再次进行合并。基于链接的分析和协同实体分析反复交叉进行，直到F1不能提高为止。下表为执行效果的F1值对比：

表2 F1值比较

	GBC-ER	LDA-ER	DistQC	NDC
叶章群	0.910	0.933	0.955	0.989
徐如祥	0.932	0.921	0.944	0.981
李莉	0.903	0.924	0.945	0.981
王国栋	0.945	0.959	0.959	0.98
刘相华	0.951	0.967	0.951	0.975
王芳	0.909	0.929	0.939	0.960
平均	0.925	0.939	0.949	0.978

如上表所示，最佳F1值达到了0.978，比GBC-ER方法提高了5.73%，比LDA-ER方法提高了4.15%，比DistQC方法提高了3.06%。NDC方法专门针对中文姓名与单位进行了优化，与国外的同类方法相比，它更适用于“中文”的文献索引数据重名分析。

NDC与GBC-ER[5]，LDA-ER[8]，DistQC[7]的执行效率进行了对比，结果见下表(单位：秒)：

表3 执行效率的比较

	GBC-ER	LDA-ER	DistQC	NDC
叶章群	221.452	124.677	78.547	3.458
徐如祥	205.674	129.575	81.632	2.467
李莉	231.257	145.692	91.786	3.968
王国栋	195.640	123.253	77.649	1.031
刘相华	133.210	83.922	52.871	0.594
王芳	121.330	76.438	48.156	3.171
平均	148.525	113.926	71.774	1.771

从上表中我们可以看到，NDC方法对于中文文献索引数据的重名分析具有良好的执行效率。在1.66 GHz Dell E1505 Intel Core Duo T2300机器上，我们将NDC方法应用于包含931172个作者标识的dataset1，以AM+LBA2+CER方式运行只需37分43.59秒，平均每秒处理约411个标识。由上表可以得出，NDC方法比GBC-ER，LDA-ER，DistQC处理数据的效率更高。因此NDC方法其他方法更适用于大数据量的实际应用。

结论 在本文中，我们提出了一种针对中文文献索引数据的重名分析方法NDC，该方法分为三步：考虑标识的属性匹配(AM)、基于链接的分析(LBA)以及协同实体分析(Collective Entity Resolution)。经过实验对比我们发现NDC方法能够获得较为理想的F1值，同时执行效率也能满足实际应用的要求。该方法已经用于数字图书馆的新型检索系统的数据预处理。

但我们的方法仍然无法寻找出所有的标识与实体对应关系，如在属性匹配中未被找到的，同时又未曾用相同的标识信息与其他标识合作两次或两次以上的，在基于链接分析中无法被找到，在协同实体分析中也没有办法识别的。还有只独立撰写了一篇论文的标识等等。所以在下一步工作中，我们将尝试加入更多的因素比如关键词、期刊信息等辅助信息来提高属性匹配的正确率，优化链接分析算法和协同实体分析算法使之能够有效地涵盖所有标识来提高F1值。

参考文献

1 Fellegi I P, Sunter A B. A Theory for Record Linkage. Journal of the American Statistical Association, 1969, 64: 1183
2 Han Hui, Xu Wei, Zha Hongyuan, Lee Giles C. A Hierarchical Naive Bayes Mixture Model for Name Disambiguation in Author Citations. In: Proceedings of the 2005 ACM symposium on Applied computing, 2005
3 Han H, Giles C L, Zha H, Li C, Tsioutsiouliklis K. Two supervised learning approaches for name disambiguation in author citations. In: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital libraries, 2004
4 Winkler W. Matching and Record Linkage. In: Brenda G. Cox, ed, Business survey methods. Wiley, 1995
5 Bhattacharya I, Getoor L. Entity Resolution in Graph Data. University of Maryland Technical Report CS-TR-4758, October 2005
6 Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, Jan. 2003, 3: 951～991
7 Won On B, Elmacioglu E, Lee D, Kang J, Pei J. Improving Grouped-Entity Resolution using Quasi-Cliques. In: Proceedings of the 2006 IEEE International Conference on Data Mining (ICDM'06). HongKong, 2006
8 Bhattacharya I, Getoor L. A Latent Dirichlet Model for Unsupervised Entity Resolution. SIAM Data Mining Conference. MaryLand, 2006
9 Cohen W W, Ravikumar P, Fienberg S E. A comparison of string distance metrics for name-matching tasks. In: Proceedings of the IJCAI, 2003
10 Warner J W, Brown E W. Automated name authority control. In: Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital libraries (JCDL01), 2001
11 Bhattacharya I, Getoor L. Deduplication and Group Detection Using Links. KDD Workshop on Link Analysis and Group Detection. Seattle, WA, August 2004
12 Torvik V I, Weeber M, Swanson D R, Smalheiser N R. A probabilistic similarity metric for Medline records: A model for author name disambiguation. Journal of the American Society for Information Science and Technology, 2005, 56(2): 140～158
13 Holzer R, Malin B, Sweeney L. Email Alias Detection Using Social Network Analysis. In LinkKDD 2005. Chicago, August 2005
14 Kalashnikov D V, Mehrotra S, Chen Z. Exploiting relationships for domain-independent data cleaning. In SIAM International Conference on Data Mining (SIAM SDM). Newport Beach, CA, USA, April 2005
15 Wang Houfeng, Mei Zheng. Chinese multi-document personal name disambiguation. HIGH TECHNOLOGY LETTERS, 2005, 11(3): 280～283
16 Getoor L, Diehl C P. Link Mining: A Survey, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2005, 7(2)