International Journal of u- and e- Service, Science and Technology
Vol.8, No. 9 (2015), pp.75-82
http://dx.doi.org/10.14257/ijunesst.2015.8.9.09

Network based Framework for Author Name Disambiguation Applications

Yuechang Liu[1,2] and Yong Tang[1]

[1]*School of Computer Science, South China Normal University*
[2]*School of Computer Science, Jiaying University*
ychangliu@gmail.com, ytang@scnu.edu.cn, chengh3@qq.com

Abstract

With the rapid development of digital libraries, name disambiguation becomes more and more important technique to distinguish authors with same names from physical persons. Many algorithms have been developed to accomplish the task. However, they are usually based on some restricted preconditions and rarely concern how to be incorporated into a practical application.

In this paper, name disambiguation is regarded as the technique of learning module integrated with a knowledge base. A network is defined for the modeling of publication information, which facilitates the representation of differenttypes of relations among the attributes. The knowledge base component serves as the user interface for domain knowledge input. Furthermore, this paper exploits a random walk with restart algorithm and affinity propagation clustering algorithm to finally output name disambiguation results.

Keywords: *Name Disambiguation, Network, Random Walk with Restart, Framework*

1. Introduction

It is common that people's name can usually be written in several forms, *e.g.*, Daniel Smith Weld, D. S. Weld, Dan. Smith Weld and even Dan Weld can actually point to the same people. As well, there are lots of people who have the same names. For example, there are 950 papers with author "Wei Wang" in DBLP which actually corresponds to over 50 people with the name. Author name ambiguity problem widely exists in the digital libraries such as DBLP[1], CiteSeer[2], ACM Digital Library[3], PubMed[4] etc.

Nevertheless, it remains challenging to solve the name ambiguity problem. This problem, referred as name disambiguation in this paper, is usually strongly related to other problems proposed in the literature, *e.g.*, entity resolution, record linkage[1], object distinction[2]. In this paper, we think it as more profound problem. Essentially what we have is only the textual information in the digital library,while concepts like*publication*, *paper*, *author*, *journal*, *conference*, etc are the entities in semantics. How to relate the syntactic words to the semantic entities is never a trivial problem. In this paper, we take the problem as the integration of basic learning technique and domain knowledge utilization. Indeed, if we are only given some textual publication data with no any other knowledge behind the data, we are actually hardly able to distinguish the authors with the same names, not to mention the distinguishing accuracy.

In this paper, a framework for name disambiguation application is proposed which integrates the network based problem representation, Random Walk with Restart (RWR) algorithm and domain knowledge utilization. Specifically, a particular network - bi-relational network structure is defined. The problem and knowledge are the input to the system, which are both represented in XML format. Compared to problem file that provides basic information to construct a bi-relational network, the knowledge file is used

ISSN: 2005-4246 IJUNESST

to further elaborate the network. The elaborated network will then be passed to the closeness calculation and clustering module to get the final result.

The paper is organized as follows. In Section 2, some most related literature are reviewed. For Section3, the preliminary knowledge and basic definition are described. Section 4 is dedicated to the detailed description of our framework. The last section concludes the paper and describes some future directions on this study.

2. Related Work

Among the relevant literature, Jing Xia proposed a network structure also named bi-relational network [3], and a kind of RWR implementation - Iterative Aggregation and Disaggregation (IAD) algorithm. We extend the definition by distinguishing the relationship as similarity and association relation which allows the relations among all different types of nodes. In [2] the authors propose a general object distinction methodology called DISTINCT. DISTINCT assumes a relational structure between neighbor tuples then computes the linkage strengths using two measures - set resemblance and random walks.Fan,*et al.*,proposed another kind of network based name disambiguation method called GHOST [4]. By using only one type of the publication attributes (coautorship) GHOST is claimed to get better performance than DISTINCT. Though GHOST seems quite promising, it is to some extent paradoxical: since the textual equality between author names is suspicious, one cannot presume some names corresponding to same entity and find the entities for other names without explicitly knowledge is given. The inherent drawback hinders GHOST to be a practical framework. Without having such restricted assumption as GHOST, Tang et. al. proposes another probability based framework for name disambiguation based on another network model - Publication Informative Graph (PIG) [5]. Tang's framework based on Random Markov Fields is a unified framework in the sense that it can easily incorporate the content-based similarity and node similarity. Though Tang's model performs quite good in the experiment data set, PIG model is so mathematical and tricky that it is not easy to use for general users.

3. Preliminary

3.1 Name Disambiguation

In this paper, some name disambiguation problem related conceptsare formally defined as follows.

Let$\mathcal{D}$ be a set of publications $\{p_1, p_2, \ldots, p_n\}$, each of which is represented by a set of attributes$\{attr_1^i, attr_2^i, \ldots, attr_{mi}^i\}$. Intuitively, these attributes can consist in a title, a list of authors, a venue (*e.g.*, a conference or a journal) and so forth. These attributes can be further extended depending on particular case (*e.g.*, publishing time, keywords, venue location, discipline of study, *etc.*,). Moreover, these attributes can also have attributes of their own (for clarification, they can be called as main attributes and sub-attributes respectively). For example, the main attribute"author"can have sub-attributes such as names, affiliations, emails and so on. However, the sub-attributes are only used to characterize main attributes and will not be placed so much importance in this paper. Furthermore, we assume in this paper that a publication has attributes set {title, authors, venue} with their literal meaning. Authors are assumed to have sub-attributes in {email, organization}.

A name disambiguation problem $\mathcal{P}$ is defined to be a function: $\mathcal{P}$: ($\mathcal{D}$, *target_auth*)$\rightarrow$ $2^{\mathcal{D}}$given input of a set of publications$\mathcal{D}$: $\{p_1, p_2, \ldots, p_n\}$ (with the attributes $\{title_i, authors_i, venue_i\}$ for p_i) and a target author *auth* that literally appearsin each of the publications' author list.

Compared to the definition in related literature that map the targetauthor to entity, this paper takes the definition of name disambiguation in "author grouping" style (as proposed in [6]). That is, the task of a name disambiguation problem is to group the publications that is produced by the same author in semantic sense with input author name *auth*.

As an example, we take the one proposed in [4] which is described as follows (the title of each publication is replaced with its venue label like "SIGMOD'12" to save space).

Example 1. Given following set of publications:

{P1: SIGMOD'02, Jiong Yang, Wei Wang, Philip S. Yu, Jiawei Han.,

P2 : SIGMOD'01, Jiawei Han, Jian Pei, Guozhu Dong, Ke Wang,

P3: SIGKDD'04, Jinze Liu, Wei Wang, Jiong Yang,

P4 : VLDB'03, Charu C. Aggarwal, Jiawei Han, Jianyong Wang, Philip S. Y,

P5: ICDM'03, Jinze Liu, Wei Wang,

P6: CIKM'02, Jian Pei, Jiawei Han, Wei Wang,

P7: ICDM'05, Peng Wang, Haixun Wang, Xiaochen Wu, Wei Wang, Baile Shi,

P8: SIGKDD'04, Chen Wang, Wei Wang, Jian Pei, Yongtai Zhu, Baile Shi}.

The name disambiguation task is to group the publications with the same target author named "Wei Wang".

3.2 Bi-Relational Network

Network is a natural and versatile model for many kinds of data. In this paper, we will use the input data for the name disambiguation task to construct a network. After that, Random Walk with Restart (RWR) will be initiated. RWR is an effective and powerful algorithms that runs on networks [7]. Basically RWR is used to compute the similarity (or closeness) between nodes by assigning scores to them. Traditionally RWR runs on general networks that make no difference to all of the nodes or edges. In this paper we also call such general networks *Single-Relational Networks* (SRN). An SRN is formally defined to be N=(V, E, W) where V is the set of nodes, E the edges and W is the function that assigns each edge with a real valued weight. However, a new network model is adopted, which is called *Bi-Relational Network* (BRN).

Definition 1 (Bi-Relational Network, BRN) A **Bi-Relation Network** is a network $N = \{V = V_1 \cup V_2, E = E_1 \cup E_2 \cup E_3, W\}$, where: $V_1 = \{v_1^1, v_2^1, \dots, v_{n1}^1\}$ and $V_2 = \{v_1^2, v_2^2, \dots, v_{n2}^2\}$ are disjoint sets of nodes, and $E_1 = \{< v_i, v_j > | v_i, v_j \in V_1\}$, $E_2 = \{< v_i, v_j > | v_i, v_j \in V_2\}$and $E_3 = \{< v_i, v_j > | v_i \in V_1, v_j \in V_2\}$.E1, E2 and E3 are disjoint sets of edges which respectively represents the edges within V1, within V2 and between V1 and V2. Particularly, edges in E1 and E2 are called ***similarity*** edges and the one s in E3 are named ***association*** edges.

Different from SRN, BRN distinguishes the nodes into two different categories. Then the edges are partitioned into three disjoint sets, respectively represent the relations within the nodes of the same set, and the relations across the node sets (that is, the ***similarity*** set and ***association*** sets respectively).

In [3], the authors also gave a BRN definition, where the relations are defined to be two disjoint sets (illustrated in Figure 2) and similarity relations only reside in one of the node set. In our definition, similarity relations in each node set are allowed. It is obvious that our definition is more general than the one in Figure 2. Indeed, the BRN proposed in [3] can easily be represented by our model, but not vice versa.

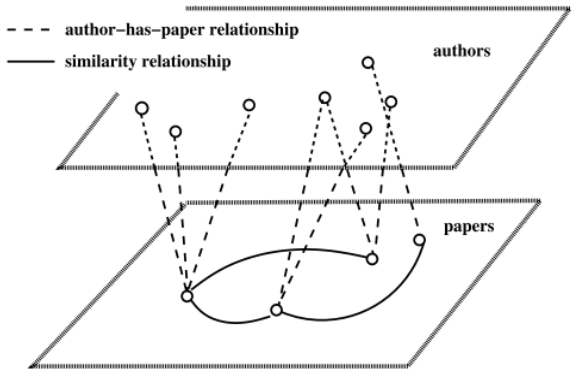

Figure 2. BRN Definition in【JX2009】

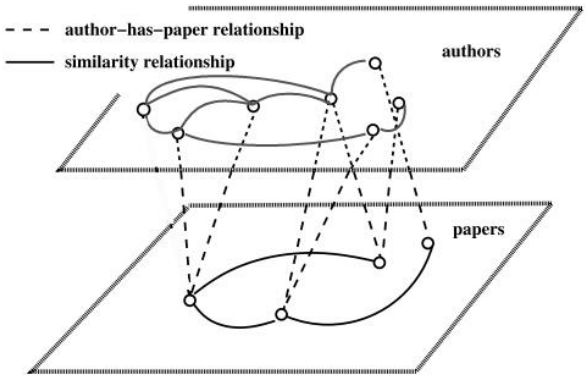

Figure 3. BRN in this Paper

Under BRN definition, the data used in Example 1 can be compiled into a BRN which is pictorially illustrated in Figure 4. In the BRN ofFigure 4, all of the papers are represented as rectangle nodes, and each of the authors corresponds to the ellipse nodes. All of the author nodes are labeled by their abbreviated form, *e.g.*,"Jiong Yang" in P1 is represented as the node labeled "JY1". The weights of the edges are not given explicitly in the example, so they can implicitly taken the value "1". This BRN network can be constructed only by scan the input publication list text. Initially all of the author nodes are viewed as different author entities because we only obtain them by textually scanning. Once we are informed of some knowledge, the network can further refined. As an example, once we are told the fact that "the author "Jiong Yang" appearing in publication P1 is the same author with the same name of P3", we can establish a similarity link between the nodes "JY1" and "JY3" with weights "1" (or, they can be totally merged.) - that means "JY1" and "JY3" are totally trusted to be the same author.

Corresponding to the Example 1, the BRN model is depicted as Figure 1.

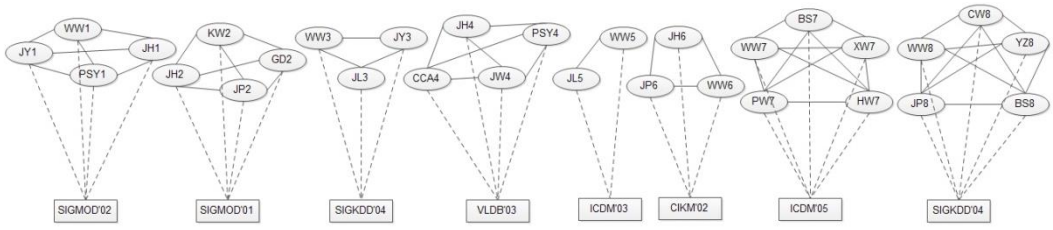

Figure 4. BRN for Example

3.3 Random Walk with Restart (RWR)

Random walk with restart algorithm is defined as Equation 1 as follows [7].

$$r = (1 - c) * W * r + c * e \qquad \text{Eq. 1}$$

RWR can be interpreted as a particle that randomly moves from a location to its neighborhood under a predefined probability 1-c, with the probability c that it may comes back to the starting point itself. With each movement the energy carried by starting node is distributed to its adjacency nodes under the same probability. The object of RWR is to calculate the energy scores (i.e. the r vector) distributed on all of the nodes in the steady state of the particle, with the initial distribution represented by the vector e.

Traditionally RWR is solved by iteration or matrix operation according to the equation transformation as Eq. 2.

$$\begin{aligned} r &= (1 - c) * W * r + c * e \\ &= c * [I - (1 - c) * W]^{-1} * e \end{aligned} \qquad \text{Eq. 2}$$

4. Our Method

4.1 System Architecture

The system architecture is depicted as Figure 5.

As mentioned, the system takes the basic problem description and prior knowledge as input.All of the input will be parsed and compiled into a central BRN structure. Currently the problem and knowledge are represented in XML format. Then, the BRN is established as the core data structure, which is further passed to the next module - closeness calculation to compute the node closeness score with respect to the query node. According to the closeness result, the paper nodes are clustered, which gives the final output of the system.

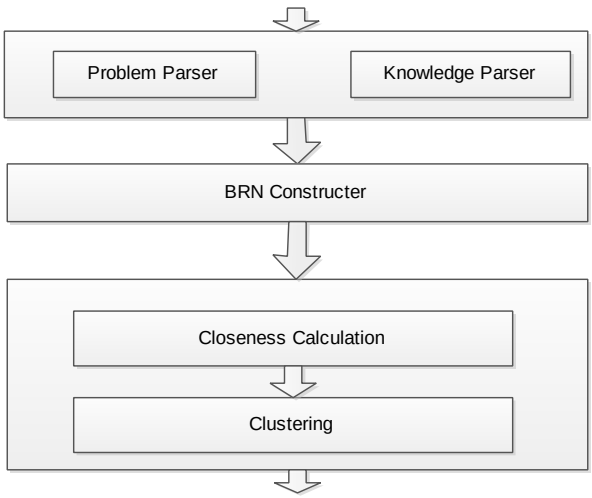

Figure 5. System Architecture

4.2 Problem and Knowledge Representation

As is mentioned above, the problem and domain knowledge is represented in XML format. For the running example, its problem and knowledge is illustrated as Figure 6 (in part) and Figure 7.

```
<?xml version="1.0" encoding="utf-8"?>
<person>
    <personID>5537</personID>
    <FullName>Wei Wang</FullName>
    <FirstName>Wei</FirstName>
    <LastName>Wang</LastName>
    <publication>
        <title>Mining long sequential patterns in a noisy environment</title>
        <year>2002</year>
        <authors>Jiong Yang, Wei Wang, Philip S. Yu, Jiawei Han</authors>
        <jconf>SIGMOD</jconf>
        <id>4031</id>
        <organization>IBM</organization>
    </publication>
    <publication>
        <title>Efficient Computation of Iceberg Cubes with Complex Measures</title>
        <year>2001</year>
        <authors>Jiawei Han, Jian Pei, Guozhu Dong, Ke Wang</authors>
        <jconf>SIGMOD</jconf>
        <id>8598</id>
        <organization>null</organization>
    </publication>
```

Figure 6. XML Representation of Example 1 (in part)

While the problem file is used to construct the initial BRN structure (as Figure 4 for Example 1), the provided knowledge is used for further elaboration of BRN. With the knowledge based described in Figure 7, the resulted BRN is illustrated in Figure 8. According to the provided knowledge, there exist the links between "Wei Wang" in the publication "P1" and "P3", and "Jiong Yang" in "P1" and "P2". In addition, new links between the indirect links (*e.g.*, between "JY1" and "JH2") are also established (via "JH1" node). In implementation layer, these operations on the BRN can be done by just merging the node pair (*i.e.*, "JY1" and "JY2", and "WW1" and "WW3"). The weight on the added edge is computed using multiplication rule of probability.

```
<?xml version="1.0" encoding="utf-8"?>
<KnowledgeBase>
    <mainfile>example1.xml</mainfile>
    <knowledge>
        <type>similarity</title>
        <degree>1.0</degree>
        <precondition>True</precondition>
        <postcondition>EQ(publication(4031).authors("Wei Wang"), publication(18145).authors("Wei Wang"))</postcondition>
    </knowledge>
    <knowledge>
        <type>similarity</title>
        <degree>1.0</degree>
        <precondition>True</precondition>
        <postcondition>EQ(publication(4031).authors("Jiawei Han"), publication(8598).authors("Jiawei Han"))</postcondition>
    </knowledge>
</KnowledgeBase>
```

Figure 7. Knowledge base for Example1

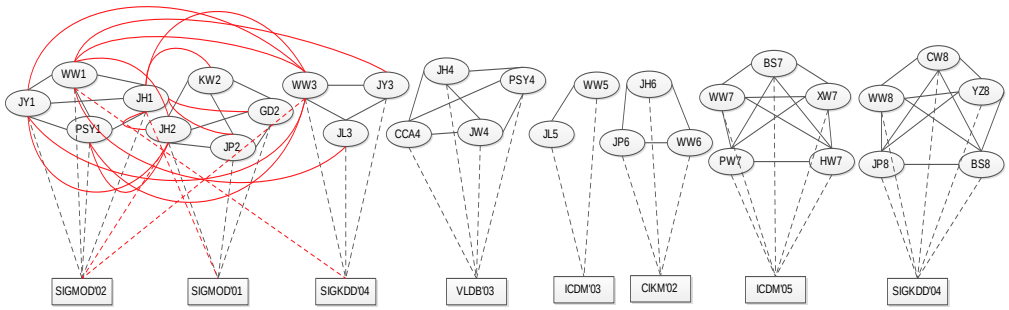

Figure 8. Elaborated BRN for Example 1

4.3 Closeness Calculation andClustering

Under RWR model, two algorithms can be exploited to compute the relative closeness scores between query nodes and other target nodes. In the Example 1 scenario, the query nodes are all of the author nodes named "Wei Wang". To cluster the authors with same name, we need to compute the closeness scores for all of the query authors, i.e. use each of the authors as query node, then invoke the RWR algorithms to compute the closeness scores between it and other authors with same name.Like [8], Figure 9 illustrates the closeness scores between all pairs of the authors with same name corresponding to the Example 1.

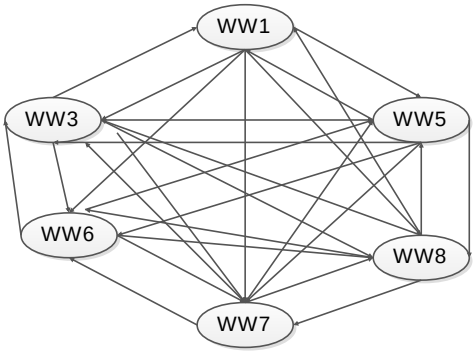

Figure 9. Closeness Network for Example 1

Based on the resulted closeness matrix, clustering phase will give the final result for the disambiguation. Considering that we actually don't know how many clusters there are in the authors with the same name, it is not suitable to use those clustering algorithms (*e.g.*, K-means, K-medoids, *etc*.,) that need a pre-specified cluster number K. like [3], *Affinity Propagation* (AP) clustering algorithm [9] is adopted in this paper.

5. Conclusions and Future Work

In this paper, a framework based on network and knowledge utilization for author name disambiguation is proposed. Different from other literature, this paper considers name disambiguation application as the integration of a network module and a knowledge base, where a network module is the basic model and reasoning mechanism for the problem representation, the knowledge base is used the user interface of domain knowledge input. The framework proposed is more practical and more suitable for real-world application.

For the next step, more study will be dedicated to the more sophisticated and effective random walk with restart variant algorithms. Moreover, how to exploit more domain knowledge in the system is another problem that worth studying.

Acknowledgments

The research is supported by following projects: the National Science Foundation (No. 60970044 and No. 61272067) in part and Jiaying University Science Project "Research and Application of Collaboration Mechanism in Social Networks".

References

[1] B. Tansel, D. G. Brizan and A. U. Tansel, "A Survey of Entity Resolution and Record Linkage Methodologies," Communications of the IIMA, pp. 41–50, **(2006)**.
[2] X. Yin, J. Han, P. S. Yu and I. T. J. Watson, "Object distinction: Distinguishing objects with identical names by link analysis," in In ICDE'07, **(2007)**.

[3] J. Xia, D. Caragea and W. Hsu, "Bi-relational Network Analysis Using a Fast Random Walk with Restart," in Ninth IEEE International Conference on Data Mining, ICDM '09,**(2009)**, pp. 1052–1057.
[4] X. Fan, J. Wang, X. Pu, L. Zhou and B. Lv, "On Graph-Based Name Disambiguation," J. Data and Information Quality, vol. 2, no. 2,**(2011)**, pp. 10:1–10:23.
[5] J. Tang, A. C. M. Fong, B. Wang and J. Zhang, "A Unified Probabilistic Framework for Name Disambiguation in Digital Library," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 6,**(2012)**June, pp. 975–987.
[6] A. A. Ferreira, M. A. Gonçalves and A. H. F. Laender, "A Brief Survey of Automatic Methods for Author Name Disambiguation", SIGMOD Rec., vol. 41, no. 2,**(2012)**, pp. 15–26.
[7] H. Tong, C. Faloutsos and J.-Y. Pan, "Fast Random Walk with Restart and Its Applications," in Proceedings of the Sixth International Conference on Data Mining, Washington, DC, USA,**(2006)**, pp. 613–622.
[8] B. Malin, "Unsupervised name disambiguation via social network similarity," in In Proceedings of the SIAM Workshop on Link Analysis, Counterterrorism, and Security,**(2005)**, pp. 93–102.
[9] B. J. Frey and D. Dueck, "Clustering by Passing Messages Between Data Points," Science, vol. 315, no. 5814,**(2007)**February, pp. 972–976.

Authors

Yuechang Liu, Ph. D in computer software and theory, postdoctoral researcher in School of Computer Science, South China Normal University. His research interest include knowledge engineering, social network analysis .

Yong Tang, Professor and Ph.D supervisor in computer science in School of Computer Science, South China Normal University. His research interests include temporal database, social computing and computer supported collaborative work.

`

