Exploring The Use Of Hybrid Similarity Measure For Author Name Disambiguation

Tasleem Arif

Abstract: Name disambiguation has become one of the hard to crack problem in a virtual setup. With each passing day more and more entities with identical features are emerging online making it quite difficult to distinguish them. Digital libraries face similar problems in differentiating publications of similar looking authors. This leads to incorrect attribution of publications, thus making the entire effort of indexing publications of individual authors ineffective. This paper proposes a two stage hybrid similarity computation mechanism that combines the best of both the worlds. The proposed method use a token-based similarity score in this first stage of comparison and based on the results of the first stage it uses a character-based similarity score in the second stage. Experimental results obtained on standard datasets indicate that the proposed technique shows a lot of improvements over the existing methods.

Index Terms: Name disambiguation, token-based, string-based, hybrid similarity, digital libraries, publications, metadata.

1 Introduction

Digital libraries and other online literature management services index publications of researchers with each group or page referring to publications of different authors. This task may seem to be a straightforward one but that is not the case [1]. Because of an inherent problem in name identification, publications one author is either split into multiple groups or publications of more than one author are grouped together. The former is commonly called as split-citation problem and latter is called mix-citation problem. Collectively these two give birth to what is called as author name ambiguity [2]. To this end solutions need to be explored to resolve this ambiguity and the solution is called as author name disambiguation. A typical author name disambiguation mechanism performs two fundamental tasks, similarity matching and record grouping. Majority of the solutions proposed so far for resolving name ambiguity use some sort similarity computation between different attributes of the candidate publications. In majority of the cases digital libraries index publications with metadata like author(s), publication title, venue (journal/conference), etc. With some additional efforts attributes like affiliation(s) of author(s), their e-mail ID(s), etc. can also be obtained. A brief description and usefulness of each of the commonly used publications attributes for author name disambiguation is provided in [3]. Each of these attributes has its own role in differentiating authors but few of them like their names, affiliations and their publication destinations are comparatively more discrete in identifying ambiguous authors. Thus the focus of similarity computation in majority of the studies proposed so far is a combination of any of these attributes. String similarity measures fall broadly under two categories, token-based or character-based [4]. Jaccard Similarity, Cosine Similarity, etc. are token-based similarity measures, whereas, Levenstein Distance, Jaro-Winkler Distance, etc. are character-based similarity measures. Variations of these similarity measures can also be found. A detailed discussion on these string similarity measures can be found in [5]. It has been observed that none of the string similarity measure is capable of comparing all types of publication attributes at its own. In this context for use in author name disambiguation, a combination of character-based and token-based similarity metrics can be explored. In this work we explore the use of mixture of character-based and token-based similarity measures in two stages for comparing various publication features.

2 Background

Author name disambiguation is a use case of name matching task. To perform name disambiguation in digital citations, various publication attributes in any two candidate citation-records have to be compared and distances/similarities have to be calculated. String comparison techniques play an important role in name disambiguation as these techniques are used to compare the distances/similarities between author names, affiliations, e-mail IDs, publication titles, etc. of any two citation-records [5]. Majority of the techniques for name disambiguation [6] use any one of the string similarity measures.

2.1 Problems with Existing Similarity Measures

Errors may creep in while the input is provided through any of the input devices. These errors are called as typographical errors or typos whereas, in some cases the errors of data entry are deliberate, such as, not concerned about using exactly the same spelling for a named entity. It may be possible for a human eye to detect typos or observe deliberate variations of same string but it is difficult for a machine to recognize these errors at its own. String comparison functions are designed to perform comparisons on input string and identify their similarity or difference. For errors of data entry like typographical errors or abbreviations, whether inadvertent or deliberate (due to differences in the conventions being followed), character-based similarity metrics work efficiently. But their efficiency decreases for larger strings [4]. Token-based similarity measures, however, work efficiently for larger strings treating them as bags of words. In token-based similarity measures the order of the words in the string does not matter [4].

3 Proposed Hybrid Similarity Measure

To overcome such problems we evaluated the use of a mix of similarity measures and different thresholds for different attributes. After testing a number of combinations and thresholds, we found that a combination of token-based similarity measure and character-based similarity measure is more effective than using any one of these approaches in isolation. In our hybrid approach, we used Cosine-Similarity in the first stage and if the value for it was above a defined threshold, we used Jaro-Winkler Similarity in the second stage. For two strings ‘s’ and ‘t’ Cosine Similarity and Jaro-Winkler Similarities are defined using Equation (1) and (2) respectively. Equation (3) is used to define Equation (2).

$$CosineSimilarity(s,t) = \cos(\theta) = \frac{s \cdot t}{\|s\| \, \|t\|} = \frac{\sum_{i=1}^{n} s_i \times t_i}{\sqrt{\sum_{i=1}^{n} (s_i)^2} \times \sqrt{\sum_{i=1}^{n} (t_i)^2}} \quad (1)$$

$$Jaro - Winkler(s,t) = Jaro(s,t) + \left(lp\left(1 - Jaro(s,t)\right)\right) \quad (2)$$

Where *l* specifies the length of the longest common prefix of *s* and *t*, and *p* is a scaling factor (constant). In Winkler's implementation *l*=4 and *p*=0.1 and

$$Jaro(s,t) = \frac{1}{3} \cdot \left(\frac{|s'|}{s} + \frac{|t'|}{t} + \frac{|s'| - T_{s',t'}}{2|s'|}\right) \quad (3)$$

Where *s'* is the number of characters in *s* that are common with characters in *t* in the same order as they appear in *s*, *t'* is the number of characters in *t* that are common with characters in *s* in the same order as they appear in *t*, and *Ts'*,*t'* is half the number of transpositions for *s'* and *t'* [4].

3.1 Affiliation Comparison

Affiliation refers to the working place of an author. Comparing affiliation strings for similarity is a difficult task because various factors like name variations, abbreviations, typing mistakes, identical indications of different institutions, etc. can cause the same affiliation to appear different in different strings or cause different affiliations to look similar [7]. Previous studies have tried to solve these problems by creating authority files [8, 9] by converting affiliations into canonical form, by determining frequency of shared affiliation tokens [10], by normalizing affiliation strings [7], etc. Canonical forms of affiliations appear to be a good proposition for finding a match between affiliation strings effectively. However, the method proposed by French and company is semi-automatic which means that creating and automating authority files for a real life database is not an easy task [10]. The similarity between affiliation strings on the basis of common words as in [10] can be obtained by using Jaccard-Similarity. We have observed that this measure fails in a number of cases, e.g. for the following two name variations of one affiliation string, this method returns very low value of Jaccard-Similarity (0.375000) after removing stop-words as done in [10]:

- "*Department of Computer Engineering, A.M.U. Aligarh.*"
- "*Computer Engineering, Aligarh Muslim University, Aligarh.*"

Before comparing the affiliation strings, we removed stop-words from them using the approach followed by [10] for such strings. We also tested our method with affiliation strings used in previous studies. Using our proposed affiliation similarity measure, we identified the two affiliation strings "*Duke University Medical Center*" and "*Duke University Medical Center and Duke Clinical Research Institute*" correctly as the variation of same string, as has also been done earlier by [7]. We use the proposed methodology to calculate the similarity in three publication features viz. author names, affiliations and publication venue titles. The task of finding similarity between affiliations and publication venue titles is different from finding similarity between author names because these two features have to be normalized before they could be subjected to comparison. The model proposed in [7] is country centric as the emphasis of the model is to deal and disambiguate institutions primarily located in the United States. Another issue with this method is that the expansion of acronyms is not well defined.

3.2 Venue Comparison

Finding the similarity or difference between publication venue titles is compounded by the existence of different variations of a single publication venue title. In addition to the variations there are always some venue titles which have a semantic relationship between them e.g. APWeb/WAIM and APWeb, ICDM and ICDM Workshops, etc. Previous studies have exploited publication venue titles for author name disambiguation in different ways: using TF-IDF model [11, 12], using same published venue referred to as co-venue [13]. However these methods consider only same venues but discard different but related venues. These latent relationships between publication venues help a lot in achieving high disambiguation performance. In order to accommodate these concerns we used a very simple and efficient mechanism of finding similar publication venues or publication venues having a latent relationship between them, and, at the same time differentiating apparently similar but different publication venues. To achieve better results we first removed common terms and stopwords from publication venue titles, where

stopwords = {one letter words like J } ∪ {small words like the, of, on, and, int, inf, intl} ∪ {international, journal, conference, society, ieee, acm, transactions, system, workshop, information} ∪{number like publication or conference year, journal volume, issue or journal number}

Experimental Results:

In this section we show the results of experiments conducted for comparison of authors and affiliations. Table-1 presents a comparison of similarity values between two name strings obtained by using commonly used string similarity measures. We compute the similarity values between two name strings using Jaccard, Cosine, and JaroWinkler string similarity measures. Analysis of the values obtained clearly reveals that no single similarity/distance function can decide whether two name strings are similar or different. It is evident from the values listed in this table that Jaccard Similarity measure fails in a number of cases where exactly similar values are obtained for two exactly similar strings and two different strings. If we take into account Cosine and JaroWinkler similarity values, there are some cases where the name refers to the same person but it has low JaroWinkler similarity and high cosine similarity whereas in some others, we have low Cosine similarity and high JaroWinkler similarity. Thus, it can be concluded that neither Cosine similarity nor JaroWinkler similarity is capable of finding a matching or different author name individually. In order to overcome these problems we use hybrid similarity measure to compare author names and affiliations where we use cosine similarity in the first stage and JaroWinkler in the second stage only when cosine similarity is above a threshold.

TABLE 1

VALUES OF DIFFERENT STRING SIMILARITY MEASURES

String1	String2	Similarity Values		
		Jaccard	Cosine	Jaro Winkler
M. Asger	Asger M.	1.000	1.000	0.810
Mohammed Asger	M. Asger	0.333	0.500	0.548
M. M. Sufyan Beg	M. Asger	0.250	0.577	0.536
Bing Liu	B. Liu	0.333	0.500	0.828
Bing Liu	Lin Liu	0.333	0.500	0.870
Rakesh K. Kumar	Rakesh Kumar	0.667	0.817	0.980
Rakesh K. Kumar	R. K. Kumar	0.333	0.708	0.472
Rakesh K. Kumar	A. Kumar	0.000	0.000	0.557
Rakesh S. Kumar	Rakesh Kumar Singh	0.500	0.667	0.877
Yenji Tang	Jie Tang	0.333	0.500	0.808
J A Walsh	Ajay Gupta	0.000	0.000	0.532
A G Sharpe	Ajay Gupta	0.000	0.000	0.600
Rashid Ali	Rashid Al-Ali	0.333	0.500	0.976
Rashid Ali	Rashid J. Al-Ali	0.250	0.409	0.930
Wajid Ali Khan	Rashid Ali	0.250	0.409	0.733
Wahid Ali Khan	Ali Shaikhali	0.250	0.409	0.795
Jin Zhang	Qian Zhang	0.333	0.500	0.819
Jin Zhang	Qing-Yu Zhang	0.250	0.409	0.710

TABLE 2

PREDICTION RESULTS OF THE PROPOSED APPROACH FOR AUTHOR NAME SIMILARITY COMPUTATION

String1	String2	Prediction Results	
		Predicted	Actual
M. Asger	Asger M.	Same	Same
Mohammed Asger	M. Asger	Same	Same
M. M. Sufyan Beg	M. Asger	Different	Different
Bing Liu	B. Liu	Same	Same
Bing Liu	Lin Liu	Same	Different
Rakesh K. Kumar	Rakesh Kumar	Same	Same
Rakesh K. Kumar	R. K. Kumar	Same	Same
Rakesh K. Kumar	A. Kumar	Different	Different
Rakesh K. Kumar	Rakesh Kumar Singh	Same	Same
Yenji Tang	Jie Tang	Similar	Different
J A Walsh	Ajay Gupta	Different	Different
A G Sharpe	Ajay Gupta	Different	Different
Rashid Ali	Rashid Al-Ali	Same	Different
Rashid Ali	Rashid J. Al-Ali	Different	Different
Wajid Ali Khan	Rashid Ali	Different	Different
Wahid Ali Khan	Ali Shaikh Ali	Different	Different
Jin Zhang	Qian Zhang	Same	Different
Jin Zhang	Qing-Yu Zhang	Different	Different

The results of our hybrid name comparison similarity methodology are shown in Table-2. From the analysis of the prediction results it can be observed that our proposed hybrid methodology is capable of achieving high degree of prediction accuracy. It has been observed that authors express their affiliations in different ways, sometimes they use full version of the affiliation and in some cases only the abbreviation. There are number of examples that can be used to test the efficiency of a proposed string comparison technique. In order to show the efficiency of our affiliation similarity computation methodology we took affiliation variations of a single actual affiliation from Figure-1 of [8]. These affiliation variations are reproduced in Table-3.

TABLE 3

SIMILARITY MEASURES RAW AFFILIATION STRINGS FOR 'UNIVERSITY OF VIRGINIA' IN ASTROPHYSICS DATASET (ADS) [14].

Variation Number	Affiliation String
1	Univ. of Virgina, Charlottesville, VA, US
2	Univ. of Virginia, Charlottesvill, VA, US
3	Univ. of Virginia, Charlottesville, VA, US
4	Univ. of Virgina, Charlottsville, VA, US
5	Univ. of Virgina, VA, US
6	University of VA., Charlottesville
7	University of Virginia, Charlottesville, VA, US
8	University of Virginia, Charlottesville, Virginia, US
9	University of Virginia, Virginia, US
10	Virginia Univ.,Charlottesville,VA,US
11	Virginia, University, Charlottesville,VA
12	Virginia Univ.
13	Virginia Univ.,Charlottesville
14	Virginia Univ.,Charlottesville, VA
15	Virginia Univ.,Charlottesville, VA US
16	Virginia University, Charlottesville
17	Virginia University, Charlottesville, VA
18	Virginia, University
19	Virginia, University, Charlottesville
20	Virginia, University, Charlottesville, VA
21	Virginia, University, Charlottesville, Va.

In this case, the proposed two-stage affiliation matching technique was able to treat 85.71 percent variations i.e. variation number 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 17, 19, 20 and 21 as being a representative of a single affiliation string. However, our method could not correlate remaining 14.29 percent variations i.e. variation number 9, 12 and 18 with other variations of the given affiliation string. Table-4 lists the comparison results. There seems to be significant improvement over the results of the other methods mentioned above. In our affiliation matching approach there is no need to create an authority file or a dictionary for normalization of affiliation strings or for expansion of acronyms.

TABLE 4
AFFILIATION SIMILARITY RESULTS

String1: 'University of Virginia' String2	Prediction Results	
	Predicted	Actual
Univ. of Virgina, Charlottesville, VA, US	Same	Same
Univ. of Virginia, Charlottesvill, VA, US	Same	Same
Univ. of Virginia, Charlottesville, VA, US	Same	Same
Univ. of Virgina, Charlottsville, VA, US	Same	Same
Univ. of Virgina, VA, US	Same	Same
University of VA., Charlottesville	Same	Same
University of Virginia, Charlottesville, VA, US	Same	Same
University of Virginia, Charlottesville, Virginia, US	Same	Same
University of Virginia, Virginia, US	Different	Same
Virginia Univ.,Charlottesville,VA,US	Same	Same
Virginia, University, Charlottesville,VA	Same	Same
Virginia Univ.	Different	Same
Virginia Univ.,Charlottesville	Same	Same
Virginia Univ.,Charlottesville, VA	Same	Same
Virginia Univ.,Charlottesville, VA US	Same	Same
Virginia University, Charlottesville	Same	Same
Virginia University, Charlottesville, VA	Same	Same
Virginia, University	Different	Same
Virginia, University, Charlottesville	Same	Same
Virginia, University, Charlottesville, VA	Same	Same
Virginia, University, Charlottesville, Va.	Same	Same

5 CONCLUSIONS

Author name ambiguity is an important problem and considerable number of efforts has been made to resolve it. Majority of the methods proposed so far are 'unsupervised' using some sort of comparison functions to find the similarity between candidate publications which requires comparing them on the basis of their constituent attributes. It was observed that majority of unsupervised name disambiguation techniques use either token-based or character-based string similarity measures. This limits the performance of the name disambiguation. In this paper we proposed a two stage hybrid string similarity computation mechanism that exploits the advantage of both exact and fuzzy string matching. Experimental results obtained on author-name and author-affiliation similarity computation are very encouraging. This proves the efficiency of the proposed approach for name matching tasks in general and author name disambiguation in particular.

REFERENCES

[1] Tang, L. and Walsh, J. P. (2010) "Bibliometric fingerprints: Name disambiguation based on approximate structure equivalence of cognitive maps." *Scientometrics*, 84(3), pp. 763-784.

[2] Lee, D., On, B.-W., Kang, J. and Park, S. (2005) "Effective and scalable solutions for mixed and split citation problems in digital libraries." In *Proceedings of the 2nd International Workshop on Information Quality in Information Systems*, Baltimore, MD, USA, ACM Press, pp. 69-76.

[3] Arif, T., Ali, R. and Asger, M. (2015) "A multistage hierarchical method for author name disambiguation." *International Journal of Information Processing*, 9(3), pp. 92-105.

[4] Bilenko, M. and Mooney, R. J. (2003) "Adaptive duplicate detection using learnable string similarity measures." In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, USA, pp. 39-48.

[5] Arif, T. (2015) "Social network extraction using web mining techniques." *Ph.D. Thesis, Department of Computer Sciences, BGSB University Rajouri*. Available online at: http://shodhganga.inflibnet.ac.in:8080/jspui/bitstream/10603/56053/4/chapter-3.pdf

[6] Ferreira , A. A., Gonçalves M. A., and Laender, A.H.F. (2012) "A brief survey of automatic methods for author name disambiguation." *ACM SIGMOD Record*, 41(2), pp. 15-26.

[7] Jonnalagadda S. and Topham, P. (2010). "NEMO: Extraction and normalization of organization names from PubMed affiliation strings." *Journal of Biomedical Discovery and Collaboration*, 5, pp. 50-75.

[8] French, J., Powell, A., Schulman, E. and Pfaltz, J. (1997) "Automating the construction of authority files in digital libraries: a case study." *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, 1324, pp. 55–71.

[9] French, J. C., Powell, A., & Schulman, E. (2000) "Using clustering strategies for creating authority files." *Journal of the American Society for Information Science*, 51, pp. 774–786.

[10] Torvik, V.I., Weeber, M., Swanson, D.R., & Smalheiser, N.R. (2005). "A probabilistic similarity metric for Medline records: A model for author name disambiguation." *Journal of the American Society for Information Science and Technology*, 56(2), pp. 140–158.

[11] Han, H., Zha, H. and Giles, C. L. (2005) "Name disambiguation in author citations using a k-way spectral clustering method."In *Proceedings of Joint Conference on Digital Libraries*, Denver, Colorado, USA, pp. 334–343.

[12] Cota, R.G., Ferreira, A.A., Nascimento, C., Gonçalves, M.A., and Laender, A.H.F. (2010) "An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations." *Journal of the American Society for Information Science and Technology*, 61(9), pp. 1853–1870.

[13] Tang, J., Fong, A.C.M., Wang, B., and Zhang, J. (2012) “A unified probabilistic framework for name disambiguation in digital library.” *IEEE Transactions on Knowledge and Data Engineering*, 24(6), pp. 975-987.

[14] Accomazzi, A., Eichhorn, G., Kurtz, M.J., Grant, C.S. and Murray, S.S. (1997) “The ADS article service data holdings and access methods.” In G. Hunt and H. Payne, editors, *Astronomical Data Analysis Software and Systems VI*, 125 of A.S.P. Conference Series, pp. 357-360.