Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 1745-1751
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.5862

AHP based feature ranking model using string similarity for resolving name ambiguity

M. Subathra[a,*], V. Umarani[a]

[a] *Department of Computer Applications, PSG College of Technology, Coimbatore, Tamilnadu, India*

(Communicated by Madjid Eshaghi Gordji)

Abstract

In recent years of Natural Language Processing research, the name ambiguity problem remains unresolved while retrieving the information of author names from bibliographic citations in a digital library system. In this paper, a feature ranking model is investigated that resolve the ambiguity problem with Analytical Hierarchy Process (AHP). The AHP procedure prioritizes and assigns the weights for certain criteria which forms a judgemental matrix called pairwise comparison matrix. The result of the AHP analysis aims to get the preprocessing level using Levenshtein Distance. Finally, the AHP helps to find the co-author criteria as the highest priority than the other criteria taken from the digital library data set.

Keywords: NLP, citations, digital library, levenshtein distance, AHP.

1. Introduction

Author name ambiguity in the case of the bibliographic context of citation is a very hard and treated as unresolved problem in recent years. This problem is mainly faced by the Digital Library System. The ambiguity occurs when there are more distinct authors with the same name or when there are same author under distinct names.

This ambiguity problem decreases the quality and reliability of the information digital library as well as quality of services provided such as information retrieval systems.

To solve this problem and improve reliability in digital library systems it is imperative to disambiguate the citation records. AHP model is used to make multicultural decisions with the help of pairwise comparison matrices to disambiguate the bibliographic citations. The ratio scales are obtained from the priority vectors and Eigenvalues obtained from pairwise comparisons.

[*]Corresponding author

Email addresses: msa.mca@psgtech.ac.in (M. Subathra), vur.mca@psgtech.ac.in (V. Umarani)

Received: Octomr 2021 *Accepted:* November 2021

The rest of this paper is structured as follows. Section 2 describes the existing works that are available in regard to author name disambiguation. Section 3 describes the analytical hierarchy Hierarchy Process Model. Section 4 concludes the paper.

2. Related Works

Anwar et al [1] has applied the weighted graph structure and used Markov clustering to disambiguate the entity names. Grouping web pages related to the same entity using overlapping measures such as web structure, content and local context of the entity names present in different web pages are considered for calculating the term weight. They reported that it requires attention to weakly connected components in the graph and weight determination of high computational complexity.

An et al [2] proposed employing a probabilistic-based Logistic Regression classifier to detect semantic aliases of an entity in the web corpus, using variables like co-occurrence, alias, and social relevancies to calculate the association score between two entities.

Bindu et al [7] have discussed the AHP method with measures of the jury evaluation to evaluate web sites. measures of the jury evaluation. They used a gray hierarchy evaluation model and confirmed the elements of the evaluation matrix.

For author name disambiguation, Cota [3] et al presented a heuristic-based hierarchical clustering methodology with two steps. In the first step, they used references with similar author names and at least one co-author name to construct clusters. The clusters of references are then fused with comparable author names in the second phase, based on the similarity of the citation attributes, such as title, publication, and venue. The information from fused clusters is aggregated (i.e., all words in the titles are grouped together) in each round of fusion, providing more information for the next round. This procedure is repeated until no further fusions are possible, according to the manufacturer. stepwise refining consumed more time for clustering an author name.

Ferreira [5] have investigated about the brief survey of different methods used to solve the author name disambiguation problem in digital libraries and similar systems. This paper proposed a taxonomy of methods which is classified and provides the most representative one. The majority of surveyed methods uses some similarity metrics in citation records, whereas the few methods use supervised and unsupervised learning approaches.

Fernandez et al [4] proposed a novel algorithm incorporated with the extracted features like information about the instance co-occurrence and news trends from news articles for finding disambiguated entities. They used the context of the news item to disambiguate entities, using the information contributed by these news items to rank the set of candidate names for a given named entity. However, there is no certainty that the entity name is present in the semantic context of the news item.

The reference [13] described about the multi-layer clustering system to identify the author names in the digital library system which is ambiguous. Each layer utilizes different methods like fuzzy logic and string similarity metrics and classifies authors and it is appended with more layers of clusters.

Tang et al [9] designed the problems of name disambiguation in a unified framework and propose a generalized probabilistic model. They used a two-step parameter estimation algorithm and estimate the number of author K using a dynamic approach.

Levin et al [6] designed an undirected graph designed with two kinds of vertices and edges to validate whether two references referred to the same author. The vertices represented a reference to an author, occurring in a citation and another represented the citation referred to by itself as well as the first link that represented the reference to the citation and yet another link represented the reference to the same author name. Also, they used social network metrics incorporated with string metrics to find similarity score. However, this type of graph based representation is difficult to model the entire network.

Shen et al [8] proposed a fuzzy set based Absolute Order-of-Magnitude (AOM) model incorporated with link based properties such as cardinality and uniqueness has been used for alias detection. The properties are constructed with qualitative descriptions of label set which are semantically defined by the collections of fuzzy sets. To generalize the performance of the proposed model, the fuzzy sets are expressed as an aggregation of weights from link based properties and triangular membership function is used for the qualitative description of the decision boundary of AOM.

Tang et al [9] proposed formalized the unified probabilistic model based on Markov random fields for disambiguation in author publication data set. The publication details with relations could be designed as an undirected network. Each paper consisted of venue, co-authors, references, abstract, and year of publication as the feature set and the formalized relationship between the papers. Also, all the features are integrated and the similarity weights are found between the papers. The Bayesian information criterion measures are used to estimate the number of groups of authors. The same author details are used for bipartite graph based social network in person name disambiguation. They have an assumption that different namesakes have the different social group and they used social network information as a predominant classification feature to identify the different namesakes. The relational data of social network is used to identify a specific person. Initially, it represented each document by the social network snippet of a specific namesake and then, bipartite graph based similarity measure is used to merge these snippets. As a result, different namesakes are identified and their social networks are generated. However, the structure-based approach is not suitable for disambiguating small groups of author publications.

Vechtomova et al [10] proposed the candidate names for a given entity as a query for the named entity extraction from the web search. Similarity measures such as tf-idf, PMI, and Pearson's coefficient have been used to rank the candidates name for refining the web query result.

Veloso et al [11] proposed the use of a supervised rule-based associative classifier to infer the authors and their references in a digital library dataset. This classifier incorporates author names, work title, publication, venue, and title as features for inference rule to infer the exact author of a reference, generated with the help of a strong association between the bibliographic features. The weighted score function is found using these rules for author name disambiguation. Using active learning, a new author data is inserted into the training data during the disambiguation process to detecting ambiguous authors not found in the training phase.

Wu et al [12] have introduced a Dempster Shafer Theory based hierarchical agglomerative clustering algorithm used for author name disambiguation with the affiliation feature for finding the pairwise similarities and generating candidate pair of clusters with the use of Jaccard and Levenshtein distance.

3. Analytical Hierarchy Process Model

It is a method to derive ratio scales from paired comparisons. It is a decision-rule model which makes judgements based on a pairwise comparison matrix. AHP model uses the principal Eigenvectors and the consistency index is derived from the principal Eigenvalue in order to derive ratio scales to make decisions. AHP not only helps us to arrive a good decision, but also makes a clear choice of decision.

The Six step procedure of an AHP model is as follows,

Defining the Decision problem The first step is to define the problem for which decision to be made

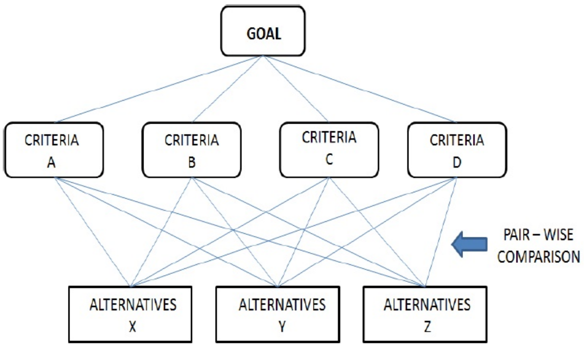
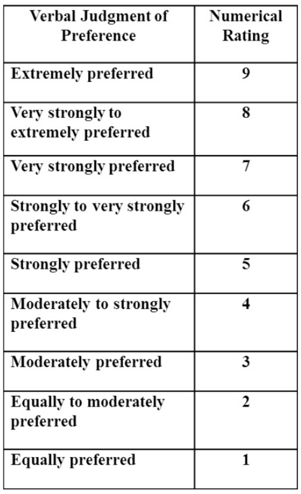


Figure 1: AHP Decision Hierarchy Model

Table 1: AHP 1-9 ratio



Verbal Judgment of Preference	Numerical Rating
Extremely preferred	9
Very strongly to extremely preferred	8
Very strongly preferred	7
Strongly to very strongly preferred	6
Strongly preferred	5
Moderately to strongly preferred	4
Moderately preferred	3
Equally to moderately preferred	2
Equally preferred	1

Setting up a decision hierarchy The next step in the AHP is to develop a graphical representation of the problem in terms of the overall goal, the criteria, and the decision alternatives. The Decision Hierarchy model of AHP is shown in Fig.1.

Employing the pairwise comparison Pairwise comparisons are fundamental building blocks of the AHP.

The AHP employs an underlying scale with values from 1 to 9 to rate the relative preferences for two items.

Estimating relative weights of elements The weighted matrix must then be determined. Now that we have a comparison matrix, we can compute the priority vector, which is the matrix's normalised Eigenvector.

Check the consistency The AHP provides a method for measuring the degree of consistency among the pairwise judgments provided by the decision maker in order to achieve quality in the decision made.

Come to a final destination based on the results Based on the priorities obtained from the resulting decision is made to achieve the required goal.

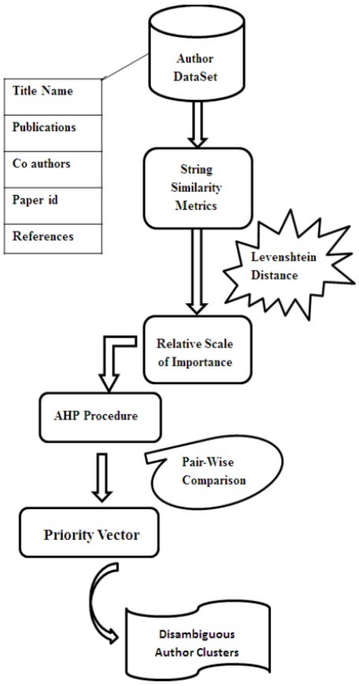


Figure 2: AHP Model

3.1. Levenshtein Distance (LD)

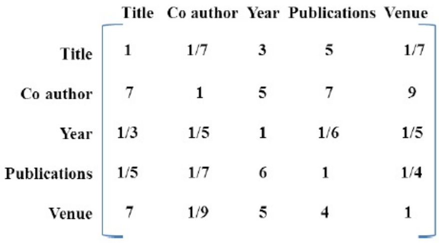
Levenshtein Distance (LD) is a string metric for measuring the differences between two sequences, which it refers to as the source string (s) and the target string (t). The greater the Levenshtein distance we obtain, the more different the strings are. The Levenshtein distance algorithm has been used in many areas like Spell checking, Speech recognition, DNA analysis.

3.2. Algorithm

Input: Two Strings
Output: Distance Between two input Strings
Step 1: Input two String s (source string), t (target string). p, q be the length of s and t
If $p = 0$, return q and exit.
If $q = 0$, return p and exit.
Construct a matrix $q \times p$.
Step 2: Initialize the row values as 0 to p and column values as 0 to q.
Step 3: Evaluate each character of s (source) (I from 1 to p) and each character of t (j from 1 to q).
Step 4: If $s[I]$ equal $t[j]$, the cost is 0.
If s $[I]$ not equal $t[j]$, the cost is 1.
Step 5: Make the value of mat $[I, j]$ equal to the minimum of: mat $[I-1, j]+1$; mat $[I, j-1]+1$; mat $[I-1, j-1]$+ cost.
Step 6: After all the iteration steps (3, 4, 5, 6) is completed, the distance between s and t is found in cell mat $[P, q]$.

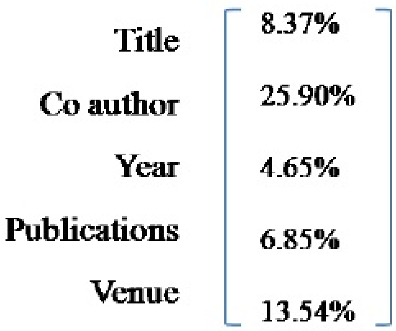
The model for the author disambiguation problem is shown in Fig 2.

Step 1: The framework inputs the author dataset from digital library system, which has a set of attributes like Title, name, paper id, references, co-author, publications and venue.



$$\begin{array}{r|ccccc|} & \text{Title} & \text{Co author} & \text{Year} & \text{Publications} & \text{Venue} \\ \text{Title} & 1 & 1/7 & 3 & 5 & 1/7 \\ \text{Co author} & 7 & 1 & 5 & 7 & 9 \\ \text{Year} & 1/3 & 1/5 & 1 & 1/6 & 1/5 \\ \text{Publications} & 1/5 & 1/7 & 6 & 1 & 1/4 \\ \text{Venue} & 7 & 1/9 & 5 & 4 & 1 \end{array}$$

Figure 3: Pairwise Matrix



$$\begin{array}{r|c|} \text{Title} & 8.37\% \\ \text{Co author} & 25.90\% \\ \text{Year} & 4.65\% \\ \text{Publications} & 6.85\% \\ \text{Venue} & 13.54\% \end{array}$$

Figure 4: Priority Vector

Step 2: The Relative scale of importance is being calculated using Levenshtein Distance.

Step 3: The AHP procedure is being used to make multi-criteria decision. Here any one criteria are taken and the AHP procedure is applied to it.The relative scale of importance is applied between two items and then the same process is carried out to other pairs to compute pair-wise comparison matrix is shown in Figure 3.

Step 4: After preparing pairwise comparison matrix, perform the calculations according to the pairwise algorithm to compute the priority vector (Eigen Value).

As per the results obtained from the priority vector for the given criteria, the Fig 5 symbolizes that the co-author reaches the highest priority value.

Step 5: Finding out the priority vector, perform consistency checks in order to measure the pairwise judgement is consistent. If inconsistent, again revise the previous matrix and follow the AHP procedure as in previous steps.

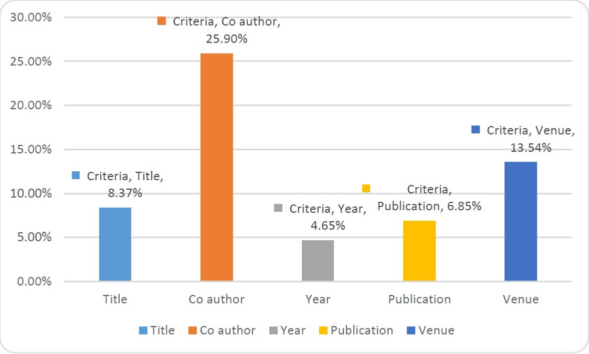


Figure 5: Sample Entity with priority values

4. Conclusion

This paper discussed about how the decision of profitable feature selection is made when there is multiple criteria using AHP with Levenshtein distance. Also, a pairwise comparison matrix of the AHP to identify the author names in digital library is investigated. It can be concluded that these techniques efficiently cluster author citations using unsupervised learning approaches to receive more accurate results.

References

[1] T. Anwar and M. Abulaishy, *Namesake alias mining on the web and its role towards suspect tracking*, Info. Sci. 276 (2014) 123–145.

[2] N. An, L. Jiang, J. Wang, P. Luo, M. Wang and B. N. Li, *Towards detection of aliases without string similarity*, Info. Sci. 261 (2014) 89–100.

[3] R.G. Cota, A.A. Ferreira, C. Nascimento, M.A. Gonçalves and A.H. Laender, *An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations*, JASIST 61(9) (2010) 1853–1870.

[4] N. Fernandez, J.A. Fisteus, L. Sanchez and G. Lopez, *Identity rank: named entity disambiguation in the news domain*, Expert Syst. Appl. 39(10) (2012) 9207–9221.

[5] A. Ferreira, M. Gonçalves and A. Laender, *A brief survey of automatic methods for author name disambiguation*, ACM SIGMOD Record 41 (2012) 15–26.

[6] F.H. Levin and C.A. Heuser, *Evaluating the use of social networks in author name disambiguation in digital libraries*, J. Inf. Data Manag.1(2) (2010) 183–198.

[7] B. Madhuri, S.T. Rao, M. Padmaja and A.J. Chandulal, *Evaluating website based on the Grey clustering theory combined with AHP*, Int. J. Eng. Tech. 2(2) (2010) 71-76.

[8] Q. Shen and T. Boongoen, *Fuzzy orders-of-magnitude-based link analysis for qualitative alias detection*, IEEE Trans. Knowledge Data Engin. 24(4) (2012) 649–663.

[9] J. Tang, A.C.M. Fong, B. Wang and J. Zhang, *A unified probabilistic framework for name disambiguation in digital library*, IEEE Trans. Knowledge Data Engin. 24(6) (2012) 975–987.

[10] O. Vechtomova and S.E. Robertson, *A domain-independent approach to finding related entities*, Inf. Proces. Manag. 48(4) (2012) 654–670.

[11] A. Veloso, A.A. Ferreira, M.A. Gonçalves, A.H. Laender and W.Jr. Meira, *Cost-effective on-demand associative author name disambiguation*, Inf. Proces. Manag. 48(4) (2012) 680–697.

[12] H. Wu, B. Li, Y. Pei and J. He, *Unsupervised author disambiguation using Dempster–Shafer theory*, Scientomet. 101 (2014) 1955–1972.

[13] J. Zhu, *A Multiple-Layer Clustering Approach to the Name Ambiguity Problem*, School of ITEE, University of Queensland, Australia, Fang Yang, School of Business, Renmin University of China, China.