Name Disambiguation With Content Mining In Digital Library Using Fuzzy C Means And Linear Discriminant Analysis

[1]K. Kamatchi, [2]N. Suguna

[1] *II ME –CSE, Department of CSE, Akshaya College of Engineering and Technology, Kinathukadavu, Coimbatore- 642109. India*

[2] *Professor & Head, Department of CSE, Akshaya College of Engineering and Technology, Kinathukadavu, Coimbatore- 642109. India*

Abstract

The researchers publish their papers in various publications like IEEE,ACM etc. Their papers can also be retrieved and viewed from internet digital libraries. While searching for any specific paper or domain the user needs to provide some keywords related to his requirements. The list of search results displayed on the screen either by considering the user keywords as a single query or by spliting them into sub queries. The results may not display only the exact domain papers since it splits the user query into keywords and performs search. For example data mining will be considered as data and mining. It displays the results which contains the word data and mining separately as well as search results with data mining as a single word. This increases the precision percentage but decreases the recall percentage. The author names might also be similar in many of the papers and if the user searches with the help of author names then the ambiguity problem remains largely unresolved in years of research. In this paper it incorporates both attributes and relationships between multiple papers and proposes a combination of two algorithms Fuzzy C Means and Linear Discriminant Analysis for text mining with Attribute based and Content Based Approach. The proposed algorithms outperforms than the previous framework by giving better precision and recall percentage of mining from a large database.

Keywords- *Information Search, Digital libraries, Text Mining.*

1. Introduction

The name disambiguation problem can be formalized as partitioning collections of paper details into clusters, with each cluster containing papers authored by the same author, thus disambiguating papers based on their similarities. Name disambiguation is a very critical problem in many knowledge management applications and Digital Libraries like Citeseer and DBLP bibliography and Semantic Web applications like semantic integration and ontology merging. Many knowledge management applications need name disambiguation as the first step. For example, expert finding, people search, expert profiling, and information integration.

Various people may share identical names in the real world. Most of the common male names are used by people repetitively. In many applications such as scientific literature management and information system, the names of authors are used as the identifier to retrieve the information. Ambiguous names will decrease the precision rate of the retrieved information.

The name disambiguation problem in research papers digital library is investigated to retrieve papers with the right author and relevant content. Traditionally, name disambiguation was often undertaken in either a supervised or unsupervised fashion. A general semi-supervised framework to combine the advantages of the supervised and unsupervised methods has been introduced.

2. Related Works

A number of approaches have been introduced to name disambiguation in different applications try to distinguish web pages to different individuals with the same name. Two unsupervised frameworks are presented for solving this problem: one is based on link structure of the Web pages and the other uses Agglomerative/ Conglomerative Double Clustering method. Name disambiguation is conducted on email data, and the authors use a lazy graph walk method based on the links among emails. There are also many works focusing on name disambiguation in publication data. For example, Han et al [7] introduced an unsupervised learning approach using K-way spectral clustering method. The problem has been independently

investigated in different domains and it is also related to web appearance disambiguation, name identification and Object distinction. Despite many approaches introduced, the name ambiguity problem remains largely unresolved.

In general, existing methods for name disambiguation mainly fall into three categories: supervised model, unsupervised model, and constraint model. In the supervised-based approach, a specific classification model is learnt for each author name from the human-labeled training data. Then, the resultant is used to predict the author assignment of each paper. In the unsupervised based approach, clustering algorithms or topic models are employed to find paper. These unsupervised methods use a parameter-fixed distance metric in their clustering algorithm, while parameters of our distance metric can be learned during the disambiguation process. The methods learn a specific model for each author name from the train data and use the model to predict whom a new citation is authored by.

However, the method is dependent on user. It is difficult to train thousands of models for all person names in a large digital library. The other type of related work is semi-supervised clustering introduces a probabilistic model for semi-supervised clustering based on Hidden Markov Random Fields. Their model combines the constraints and distance measures. It defines six kinds of constraints and generates the constraints automatically. The constraint-based approach also utilizes the clustering algorithms. In this user-provided constraints are used to guide the clustering algorithm toward better data partitioning. In addition, many other approaches based on rules mining, citation of papers, author graphs and combinations of the different approaches have been studied.

For example, Wang et al [8] introduced a negative rules-based approach to remove the inconsistencies in the databases and develop two algorithms to identify important properties to create the rules. Davis et al [9] have developed an interactive system which permits a user to locate the occurrences of names in a document. The method used to identify references to a single art object, for example, a particular building in text related to images of that object in a digital collection. McRae-Spencer and Shadbolt [10] presented a graph-based approach to author disambiguation on large-scale citation networks by using author or paper citation, author relationships. This approach achieves a high precision but a relatively low recall. Yu et al [13] have developed supervised approaches to identify the full forms of ambiguous abbreviations within the context they appear. More recently, Chen et al [5] studied how to combine the different disambiguation approaches and introduced an entity resolution framework that combines the results of multiple base-level entity resolution systems into a single solution to improve the accuracy of entity resolution. Wang et al [8] defined an iterative blocking framework where the resolution results of blocks are reflected to subsequently processed blocks. Lee [6] studied the scalability issue of the name disambiguation problem.

Although much progress has been made, due to their limitations existing methods do not achieve satisfactory disambiguation results. Some existing graph clustering methods focus on partitioning the data graph based on the topological structure. some other methods aim to cluster the data graph according to node similarity. A few researchers try to combine the two pieces of information. For example, Zhou et al [5] attempt to combine information based on both vertex attributes (i.e) node similarity and graph topological structure by first constructing an attribute augmented graph through explicit assignments of attribute, value pairs to vertices, and subsequently estimating the pair wise vertices' closeness using a neighborhood random walk model. The pair wise comparisons discard all the topological information. Even though the authors demonstrate that attribute similarity increases the closeness of pair wise vertices in their distance measure, how to balance the contributions of the different information is still problematic. They are only able to conclude that adding attribute similarity information to the clustering objective will not degrade the intra cluster closeness.

Further the experimental data sets contain very few attributes. The first data set only has one attribute and the second data set of DBLP bibliographical data only has two attributes. Much richer node attribute information is required for tackling the name disambiguation problem effectively. The performance of all the aforementioned methods depends on accurately estimating K. Many clustering algorithms such as X-means can automatically find the number K based on some constraint, how such methods can be directly applied to the name disambiguation problem is not known.

In existing methods, the data usually only contain homogeneous nodes and relationships. In our problem definition, multiple different relationships are considered. For example, Main Author, Co Author and Citation between nodes are taken into consideration. The different types of relationships may improve the name disambiguation problem analysis. Modeling the degree of different relationships in a learned method is again a tedious problem.

3. Problem Identification

The problem is illustrated with an example Fig 1 drawn from a real-world system. The researcher profiles are extracted from the web and the publication data are integrated from the online databases such as DBLP,

ACM Digital Library, CiteSeer, etc. In the simplified example Fig 1, each node denotes a paper. Each directed edge denotes a relationship between two papers with a label representing the type of the relationship.

The distance between two nodes denotes the similarity of the two papers in terms of some content-based similarity measurement (e.g., cosine similarity).

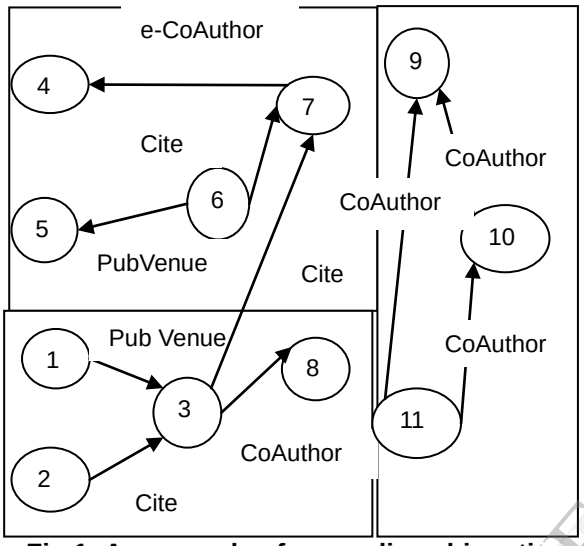

Fig 1- An example of name disambiguation

The rectangle indicates the ideal disambiguation results where 11 papers are assigned to three different authors.

A method based on only content similarity or the distance would be difficult to achieve expected outcome. Hence different types of relationships can help much better with different degrees of relationships. For example, nodes #3 and #8 are related with Co Author. The similarity between these two nodes is not high but the papers can be still assigned to the same author.

Table 1-Attributes of a Paper

Attribute	Description
p_i.title	title of p_i
p_i.pubvenue	published conference/journal of p_i
p_i.year	published year of p_i
p_i.abstract	abstract of p_i
p_i.authors	authors name set of $p_i\{a_i^{(0)},a_i^{(1)}\ldots.a_i^{(u)}\}$
p_i.references	references of p_i

On the contrary, although there is a Citation relationship between nodes #3 and #7, the two papers are assigned to two different authors. An algorithm should be designed for the name disambiguation problem by considering both attribute information of the node as shown in Table 1 and the relationships between nodes. Each paper pi has one or more authors. The author name that has to be disambiguated is described as the principle author and the rest if any exists as secondary authors.

A cluster atom is a cluster in which papers are closely connected for example, the similarity > threshold. Papers with similarity less than the threshold will be assigned to disjoint cluster atoms. It is nontrivial to perform these tasks. To formalize the entire disambiguation problem in a framework is not easy. In general Graph models like Markov Random Field are applied to model relational data. In the graph, the papers might be arbitrarily connected by different types of relationships. Performing inference or parameter estimation in such a graph which has arbitrary structure is difficult a little bit. In addition, estimating the number of people K is also a considerable task.

Table 2- Relationships between Papers

R	W	Relation Name	Description
r_1	w_1	CoPub Venue	p_i.pubvenue=p_j.pubvenue
r_2	w_2	CoAuthor	s>0,$a_i^{(r)}=a_j^{(s)}$
r_3	w_3	Citation	p_i refers p_j or p_j refers p_i
r_4	w_4	Constraint	feedback supplied by users
r_5	w_5	e-CoAuthor	extension co-authorship

4. The Proposed Method

Having conducted a thorough investigation, a combination of two algorithms have been proposed to address the above challenges. The disambiguation problem is formalized using Fuzzy C means to cluster the paper details in different categories based on the relationships between papers as shown in Table 2. The proposed approach can achieve better performance in name disambiguation than existing methods because the approach takes advantage of interdependencies between papers and assigned relationships. Name disambiguation problem is formalized in two different ways. e.g., a feature based on the web search engine used with pattern matching instead of keyword search.

The enhancements include:

a. Formulating the name disambiguation problem along with time constraints.

b. An algorithm to solve the keyword search by considering many relationships.

The enhancements denotes how the name disambiguation problem is formalized by using Fuzzy C- Means algorithm which can be used for clustering papers in more than one group based on the relationships between multiple papers. The problem is further enhanced with Linear Discriminant Analysis which aims to add content mining for retrieval of papers.

4.1. Fuzzy C-Means Clustering

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. At first this algorithm was developed by Dunn in 1973 and improved by Bezdek in 1981. It is frequently used in pattern recognition and pattern matching techniques. It uses minimization of the following objective function depicted in equation 1:

$$J_m=\sum_{i-1}^{N}\ \sum_{j-1}^{C} u_{ij}^{m} \| x_i - c_j \|^2 \quad , \quad 1\leq m<\infty \qquad (1)$$

where m is any real number greater than 1. u_{ij} stands for the degree of membership of x_i in the cluster j, x_i stands for the ith of d-dimensional measured data, c_j stands for the d-dimension center of the cluster, and ||*|| stands for the norm to denote the similarity between any measured data and the center. While partitioning, Fuzzy method is employed through an iterative optimization of the objective function. It simultaneously updates the membership u_{ij} and the cluster centers c_j . The degree of membership u_{ij} is calculated by using the equation 2:

$$u_{ij} = \frac{1}{\sum_{k-1}^{c} (\| xi - cj \| / \| xi - cj \|)^{\frac{2}{m-1}}} \qquad (2)$$

The center of the cluster c_j is calculated by using the equation 3:

$$c_j = \frac{\sum_{i-1}^{N} U_{ij}\, m \ . \ xi}{\sum_{i-1}^{N} U_{ij}\, m} \qquad (3)$$

This iteration will stop when $\max_{ij} \{ | U_{ij}(k+1) - U_{ij}(k) | \} < T$, where T is a termination condition between 0 and 1, with k number of iteration steps. This procedure meets a local minimum or a saddle point of J_m. The algorithm is composed of several steps.

The Steps are:
a. Initialize $U= [u_{ij}]$ matrix, $U^{(0)}$
b. At each iteration k-step: Calculate the centers vectors $C^{(k)} = [c_j]$ with $U^{(k)}$ using equation 3.
c. Update $U^{(k)}$, $U^{(k+1)}$ with the help of calculating the value of u_{ij} by using equation 2
d. If $| U^{(k+1)} - U^{(k)} | < T$ then stop; otherwise return to step2.

As already told, data are bound to each cluster by means of a Membership Function, which represents the fuzzy behavior of this algorithm. Appropriate matrix named U whose factors are numbers between 0 and 1 have to be built. It represents the degree of membership between data and centers of clusters. For example, consider a certain data set, suppose to represent it as distributed on an axis.

Fig 2- Data distribution on axis

Looking at the Fig 2, we may identify two clusters in proximity of the two data concentrations. We will refer to them using 'A' and 'B'. In the first approach shown in this in the k-means algorithm - we associated each datum to a specific centroid; therefore, this membership function appears like a form as shown in Fig 3:

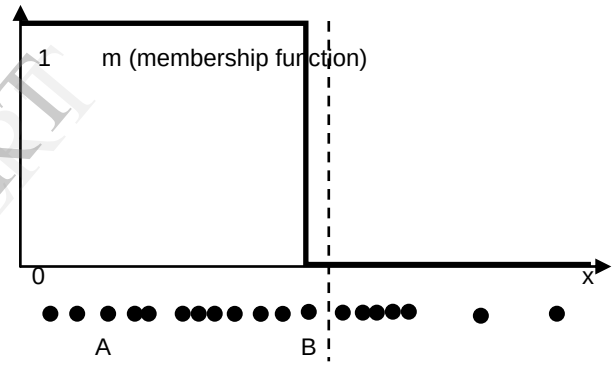

Fig 3- Membership Function

In the FCM approach, the same datum does not belong exclusively to a cluster alone and it can be placed as a middle one. The membership function which follows a smoother line indicates that every datum may belong to several clusters with different values of the membership coefficient.

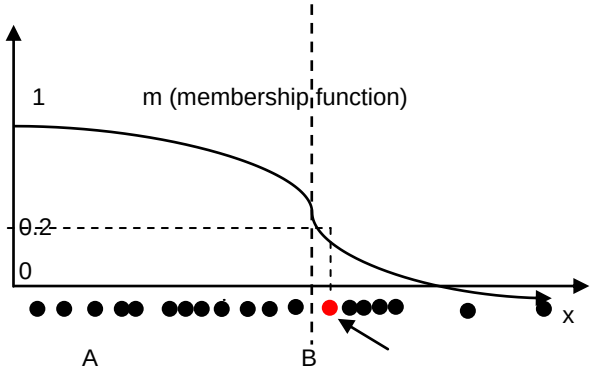

Fig 4- Datum marked on Red

In the Fig 4, the datum shown as a red marked spot

belongs more to the B cluster rather than the A cluster. The value 0.2 of 'm' indicates the degree of membership for the datum A. Matrix U is introduced to represent this and its factors are the ones taken from the membership functions:

$$U_{N\times C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ \cdots & \cdots \\ 0 & 1 \end{bmatrix} \quad (4)$$

$$U_{N\times C} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ \cdots & \cdots \\ 0.9 & 0.1 \end{bmatrix} \quad (5)$$

The number of rows and columns depends on how many data and clusters we are considering. More exactly we have C = 2 columns (C = 2 clusters) and N number of rows. C denotes the total number of clusters and N denotes the total number of data. In the examples above we have considered the k-means matrix equation (4) and FCM matrix equation (5) cases. In the first matrix specified in (4) the coefficients are unitary in order to indicate the fact that each datum can belong only to one cluster. The other properties are to be satisfied as specified in equations (6), (7) and (8):

$$u_{ij} \in [0,1] \quad \forall\ i,j \quad (6)$$

$$\sum_{j-1}^{C} \cup ik = 1 \quad \forall\ i \quad (7)$$

$$0 < \sum_{i-1}^{N} \cup ij < N \quad \forall\ N \quad (8)$$

4.2. Linear Discriminant Analysis

Linear discriminant analysis (LDA) and the related Fisher's linear discriminant are methods used various fields such as statistics, machine learning and pattern matching. to find a linear combination of features which characterizes or separates two or more classes of objects. The results are combined and used as a linear classifier used for dimensionality reduction before later classification.

LDA is closely related to ANOVA (ANalysis Of VAriance) and regression analysis to express one dependent variable as a linear combination of other features or measurements. The dependent variable is a categorical variable (i.e.) the class label in LDA. Logistic regression and probit regression are also similar to LDA and uses categorical variable. ANOVA has the categorical independent variables and a continuous dependent variable where as the discriminant analysis has continuous independent variables and a categorical dependent variable.

LDA is also closely related to Principal Component Analysis (PCA) and factor analysis in that they both look for linear combinations of variables to denote the data. The difference between the classes of data is modeled by LDA. Any difference in class is not taken by PCA. Factor analysis builds the feature combinations based on differences rather than similarities. Linear Discriminant analysis is different from factor analysis in that it is not an interdependence technique: a distinction between independent variables and dependent variables also called criterion variables must be made.LDA works when the measurements made on independent variables for each observation that are continuous quantities.

5. Experimental Results

We evaluated the proposed method by combining the FCM along with LDA algorithms. We created a data set, which includes 25 real author names and 300 papers. In these names, some names are only associated with a few persons.

Initially the paper details are added as nodes and relationships between papers are depicted through edges. The clustering of papers with similar details is done with the help of FCM and the retrieval is achieved by using relationships and similarity. After clustering, Linear Discriminant Analysis is used to improve the recall percentage by retrieving papers based on content instead of keyword search alone.

6. Conclusion

Till now the name disambiguation problem has yet unresolved. To improve its performance instead of searching papers based on their attributes, content mining has been introduced with the help of Linear Discriminant Analysis. It retrieves papers based on pattern matching techniques. The system is implemented with Fuzzy C Means algorithm for clustering papers with their relationships of attributes and it uses keyword search. In Future the system can be enhanced with Title search instead of keyword search by splitting the user query into keywords. In addition to that time information will also be included when the same author submits more than one paper.

7. References

[1] Jie Tang, A.C.M. Fong, Bo Wang, and Jing Zhang, **"**A Unified Probabilistic Framework for Name Disambiguation in Digital Library" - IEEE Transactions On Knowledge And Data Engineering, VOL. 24, NO. 6, JUNE 2012 pp.975-987

[2] Sugato Basu, Mikhail Bilenko, Raymond J. Mooney, **"**A Probabilistic Framework for SemiSupervised Clustering" - Proceedings Of The Tenth ACM SIGKDD International Conference On Knowledge Discovery And Data Mining (KDD-2004), Seattle, WA, AUGUST 2004, pp. 59-68.

[3] Ron Bekkerman, Andrew McCallum, "Disambiguating Web Appearances of People in a Social Network" - ACM, Chiba, Japan, MAY 10-14,2005, pp. 463-470.

[4] Chris Buckley, Ellen M. Voorhees, "Retrieval Evaluation with Incomplete Information" - ACM, Sheffield, South Yorkshire, UK, JULY 25–29, 2004, pp. 583-560.

[5] Zhaoqi Chen, Dmitri V. Kalashnikov, Sharad Mehrotra, "Adaptive Graphical Approach to Entity Resolution" - ACM, Vancouver, British Columbia, Canada, JUNE 17–22, 2007, pp. 978-987.

[6] Hui Han, Lee Giles, Hongyuan Zha,Cheng Li, Kostas Tsioutsiouliklis, "Two Supervised Learning Approaches for Name Disambiguation in Author Citations" - ACM, Tucson, Arizona, USA, JUNE 7–11, 2004, pp. 296-305.

[7] Hui Han, C Lee Giles, Hongyuan Zha, "Name Disambiguation in Author Citations using a Kway Spectral Clustering Method " - ACM, Denver, Colorado, USA, JUNE 7–11, 2005, pp. 334-343.

[8] Lili Jiang, Jianyong Wang, Ning An , Shengyuan Wang, Jian Zhan , Lian L, "GRAPE: A Graph-Based Framework for Disambiguating People Appearances in Web Search " - Ninth IEEE International Conference on Data Mining, OCT 2009, pp. 199-208.

[9] Xin Li Paul Morie Dan Roth, "Identification and Tracing of Ambiguous Names: Discriminative and Generative Approaches " - University of Illinois, Urbana 2009, pp.156-165.

[10] Duncan M. McRae-Spencer, Nigel R. Shadbolt, "Also By The Same Author: AKTiveAuthor, a Citation Graph Approach to Name Disambiguation " - ACM, Chapel Hill, North Carolina, USA. JUNE 11-15 2006, pp. 354-355.

[11] Yang Song1, Jian Huang2, Isaac G. Councill2, Jia Li3,1, C. Lee Giles2,1 "Efficient Topic-based Unsupervised Name Disambiguation " – ACM,Vancouver, British Columbia, Canada. JUNE 2007, pp. 342-351.

[12] Yuanyuan Tian, Richard A. Hankins, Jignesh M. Patel, "Efficient Aggregation for Graph Summarization " – ACM,Vancouver, British Columbia, Canada. JUNE 2008, pp. 102-111.

[13] Xiaoxin Yin, Jiawei Han, Philip S. Yu, "Object Distinction: Distinguishing Objects with Identical Names " –IEEE 2007, pp. 1242-1246.