I.J. Modern Education and Computer Science, **2016, 3, 28-36**
Published Online March 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2016.03.04

Semi-Supervised Personal Name Disambiguation Technique for the Web

P.Selvaperumal
Manonmaniam Sundaranar University/Department of Computer science and Engineering,
Tirunelveli, India
E-mail: selvaperumal.p@gmail.com

Dr.A.Suruliandi
Manonmaniam Sundaranar University/Department of Computer science and Engineering,
E-mail: suruliandi@yahoo.com

Abstract—Personal name ambiguity in the web arises when more than one person shares the same name. Personal name disambiguation involves disambiguating the name by clustering web page collection such that each cluster represents a person having the ambiguous name. In this paper, a personal name disambiguation technique that makes use of rich set of features like Nouns, Noun phrases, and frequent keywords as features is proposed. The proposed method consists of two phases namely clustering seed pages and then clustering the actual web page collection. In the first phase, seed pages representing different namesakes are clustered and in the second phase, web pages in the collection are clustered with the similar seed page clusters. The usage of seed pages increases the accuracy of clustering process. Since it is difficult to predict the number of clusters need to be formed beforehand, the proposed technique uses Elbow method to calculate the number of clusters. The efficiency of the proposed name disambiguation technique is tested using both synthetic and organic datasets. Experimental result shows the proposed method achieves robust results across different datasets and outperforms many existing methods.

Index Terms—Personal Name disambiguation, Entity name disambiguation, Web page clustering.

I. Introduction

Name queries are quite common in web. Approximately 11-14 % of queries to search engine are personal name queries and one in fourth of the personal name queries involves a celebrity name [1].Name disambiguation is of two types, one involves same persons having multiple names (alias names) [2] and another is different persons sharing the same name. If the query contains just name of the person then obviously it is difficult for a search engine to uniquely identify a person. For a named query "Tom Mitchell" to Google search engine, there are 37 different Tom Mitchell's out of top 100 web page results [3]. Current search engine retrieves all the web pages containing that name, thereby putting the onus of sieving through the web page collection on the user. While a narrow query like "Stephen roberts Professor University of Oxford" gives low recall since not all the web pages of Stephen roberts will have all these words, a query "Stephen roberts" will lead to low precision because of retrieving all the Stephen roberts records. Thus striking a balance between a precision and recall is the ultimate aim for any Information retrieval system.

Named entities in the web are highly ambiguous. A query containing an ambiguous name will return web pages that belongs to all of the persons sharing that name. One possible way to resolve ambiguity is to cluster web pages of different namesakes [4]. There are several people search engines like "people.yahoo.com", "spokeo.com", "peoplesmart.com" with limited capability to locate a person in the web. Web people search is closely related to name disambiguation, as the former is general task of finding a person uniquely from a large number of people sharing the same name in the web and the latter refers to discriminating people using same name. There are many methods proposed for web people search in the web [5] [6]. Name ambiguity is also prevalent in bibliographic databases like DBLP, CiteSeer, PubMed and Medline. For example, author Christopher M. Bishop may appear in multiple publications with different abbreviation forms like C.Bishop, Christopher Bishop, Christopher Michael Bishop or even in a misspelled form like Christophor bishop. For example, the author "David S. Johnson" may appear in multiple publications under different name abbreviations such as "David Johnson", "D. Johnson", or "D. S. Johnson", or a misspelled name such as "Davad Johnson". A search of author name Wang in the DBLP bibliographic database throws a match of 333 authors having it as a part of their name. This shows the prevalence of author name ambiguity in bibliographic database. Name disambiguation problem is also familiar in biomedical texts where genes and proteins often share the same name [7].

If people sharing the same name have different affiliations, then disambiguation process is slightly easier. There may be persons sharing both name and affiliations. In such cases, the disambiguation process is complex

since a lot of words do co-occur between them. Information retrieval is challenging for the persons whose relative presence in the web is lesser compared to other persons sharing the same name. The current search engine will retrieve top ranked web pages which in majority of the cases are influential persons in the web. Thus existence of an influential person overwhelms results of other. (Compare "Julia Roberts" actress and "Julia Roberts" Professor Western Kentucky University)[8]. This means that if the interested person presence in the web is low(his web page rank is low and other persons sharing his name have relatively high presence) it may be possible that this person's web pages may not figure in to the top results and makes it difficult to make it in the search results. One possible solution for this problem is clustering the web pages before ranking process. Since each cluster represents an individual sharing the name, during ranking process, giving due weightage to each cluster will boost other people's web pages which are normally ranked low. Thus results returned after the clustering process will compose of a harmonic mixture of web pages of different persons irrespective of their page rank. User can then select interested person from it, thus reduces the burden of users.

Closely related fields of personal name disambiguation includes Multi document personal name resolution [9], cross document co-reference resolution, Word sense disambiguation, word sense discrimination entity resolution and Word sense disambiguation etc. All these problems are clearly distinct from name disambiguation problem even though they share some similarity. In co-reference resolution, the task is to extract co-reference chain for a personal name, which is quite different from personal name disambiguation task where the task is to cluster documents of a same person together. Entity resolution refers to resolving the references to object of real-world entities. Word sense disambiguation refers to selecting the appropriate sense of the word in a given context. For example the word bank refers to river bank as well as financial bank if it is used as noun and if it is used as verb is also means depending on. Word sense disambiguation can rely on dictionaries containing different senses a word whereas Name disambiguation is difficult to solve by creating or maintaining such a dictionary. To resolve ambiguous words, approaches like using dictionaries, ontologies containing list of words and their senses are used. Name disambiguation is distinct from these problems as it is not possible to find list of persons sharing the same name beforehand. Alias name extraction is another closely related problem to name disambiguation, where the task is to extract all the surnames that refers to a name has to be extracted [10].

The Web People Search (WePS) Evaluation is a evaluation campaign focused on name disambiguation problem particularly in the web [11]. WePS workshop involved primarily two kinds of tasks, clustering and attribute extraction. Clustering involves grouping web pages of each individual who are sharing the same name and attribute extraction involves extracting attributes pertaining to each person like date of birth, location, education. Weps-3 which held on 2010 has given 300 personal names along with first 200 top ranked web pages of each person. Like previous editions it also hand web page clustering, attribute extraction along with a third task of online reputation management where the task was to discriminate between the ambiguous company names. WePS workshop conducted in 2010 involved extracting 16 kinds of "attribute values" of target individuals including Birthplace, occupation, affiliation, award, phone degree etc. Web People search (WePS) provides annotated dataset to evaluate different methods of name disambiguation for training and testing. A total of 300 person names were used in Weps-3, compared to 30 names used in WePS-2 obtained names randomly from the US Census, Wikipedia and computer science conference program committees. In WePS-3, 300 person names are provided with the top 200 documents retrieved from the search engine for each person name. The task is to cluster the documents such that each cluster should represent an ambiguous personal name.

Solving the name ambiguity has a wide spectrum of applications ranging from information retrieval to question and answering. Name disambiguation helps in removing ambiguity of names thereby increasing accuracy of a number of systems. In question and answering system, [12] disambiguating a named entity contained in the query will increase accuracy of answers. Similarity in many social network extraction system [13], name disambiguation is the first step. During ontology integration, a personal name may appear in one or more ontologies. Disambiguating personal names before ontology integration is helpful in constructing an informative ontology.

Name disambiguation can be solved either by using local methods or by using global methods [4]. The local method includes using word level features like words [14], bi-grams, sentence level features [15] , extracting biographic information[16], named entities [17] etc. Global methods includes using external corpora like using Wikipedia for name disambiguation [19] [19].Disambiguation can also be done by two ways. One is disambiguation routine in the server side and another is disambiguation at the middleware. Name disambiguation in the web involves clustering web pages such that each cluster belongs to one of the group of person sharing that ambiguous name [15] [20]. It is a hard clustering problem, meaning that each document that needed to be clustered should strictly be in a maximum of one cluster. It should also be noted that, there may be some web pages that belongs to more than one clusters and thus soft clustering is needed. Over the years, name disambiguation is done using keywords, phrases, sentence level features, extracting biographic information etc. Most of the current system uses either biographic information extracted from web pages [16] or use of external knowledge base like Wikipedia[21][18], web directories[22].

1.1 Related Work

Most of the method for personal name disambiguation includes extracting vector space model based features

from web pages (like key phrases, words, bi-grams etc) and performing Hierarchical agglomerative clustering to form 'n' number of clusters where 'n' refers to number of persons sharing that ambiguous name. In personal name disambiguation process, the real challenge is finding the value of 'n'.

Zhao Lu et al, [23] proposed an ontology based name disambiguation process. They first constructed personal ontology for every ambiguous name, temporary instances were created from the features extracted from the web. Comparing similarity between these instances with instances from the constructed ontology solves the ambiguity of the personal name. Masaki Ikeda et al, [24] proposed a two stage clustering method using features like named entities, compound keywords, and URLs for disambiguating personal names in the web. They first used Hierarchical agglomerative clustering to cluster similar web documents, extracted compound keywords from the clustered results and finally performed soft clustering. Rabia Nuray et al, [25] proposed a new people search technique by issuing auxiliary queries to the web. Based on the co-occurrence statistics gathered using web and a new skyline based classifier is then used to decide whether to merge a document or not. Bekkerman et al, [8] proposed two unsupervised frameworks for disambiguating peoples name in social network. First is based on link structure of web pages and the other uses Agglomerative/Conglomerative Double Clustering. Their method requires social network (friends circle) of an individual whose name has to be disambiguated. Han, Xianpei [26] professional names for personal name disambiguation. They first extracted professional information for each ambiguous personal names and then a trained classifier for each profession is used to classify ambiguous names into their respective professional categories. The method is suited well for well-known persons whose professional categories are available, which is not always the case. Zhengzhong Liu et al, [4] used standard vector space model to represent document for clustering. They extracted seven tokens namely Web page tiltle, url of webpage, web page meta data, snippet, words within a window containing query name and sentence containing the query name and bag-of-words. They then used Hierarchical agglomerative clustering (HAC) to cluster same topic documents of each person. Chong Long et al, [21] used Wikipedia concepts, bag-of-words and named entities as features and two feature weighing models namely feature relevance to the ambiguous person and feature relevance to the text content. Razvan Bunescu et al, [19] used knowledge from Wikipedia and with trained SVM kernel disambiguated entity names. However the problem with Wikipedia is that all the real world people does not have a page in Wikipedia.

Sugiyama, Kazunari et al [27], used semi supervised clustering for disambiguating peoples name in the web. They used Wikipedia page or top ranked web page in the web search results as seed pages and performed agglomerative clustering which outperformed conventional agglomerative clustering. It is also observed that usage of sentences in the web pages containing ambiguous personal name for clustering improves significant clustering results. Ergin Elmacioglu et al, [28] used a number of features like tokens, named entities, hostnames and domains, page URL's for name disambiguation process. Using Hierarchical agglomerative clustering with these features they disambiguated personal names across web pages. Duo Zhang, [29] proposed a constraint-based probabilistic name disambiguation model using semi supervised learning. They have defined six types of constraints and using constraint functions they disambiguated author names in bibliographic databases. The use of large number of constraints is difficult given the diverse and sheer scale of the web. Guha [30], proposed a ranking algorithm where the user selected page is used to re-rank the search results. Einat Minkov et al, [31] used lazy graph walk method that exploits the link between emails to disambiguate names in emails. Quang Minh Vu et al, [22] used web directories as external knowledgebase for name disambiguation process. Their method first finds common contexts first and then using this it finds the document similarity. Lee Ingyu [32] used linear algebraic approaches Singular valued decomposition and Nonnegative Matrix Factorization for solving name disambiguation problem. Lu, Yiming [33], used web connection for disambiguating personal name. They used distinctive information of ambiguous persons as query to search engine, and resultant web pages are then used to construct web connection. Gideon Mann [16], method of unsupervised personal name disambiguation involves extracting biographic information like birthday, year, occupation etc by bootstrapping process. They used both words and nouns from the web pages individually, and used standard cosine measure for finding similarity between feature vectors.

1.2 Motivation and Justification

Despite a number of methods proposed for name disambiguation in the web, the problem still persists because of addition of new names as the time go by and the highly unstructured nature of the web content. The search engines efficiency can be improved if results of personal names are clustered according to the corresponding sense rather than presenting as flat [19].Motivated by this, a novel method for web based name disambiguation is proposed in this paper.

Hierarchical Agglomerative Clustering algorithm provides better results for name disambiguation process [34] [4] [24]. The use of Semi supervised learning suits well for solving name disambiguation process in the web [27] [29]. The use of nouns alone clusters well the web pages according to the namesakes to which it represents [16]. It is quite uncommon that a single web page containing ambiguous names that refers to more than one person sharing that name [34] [35]. Most of the previous works in name disambiguation views it as a hard clustering problem and involves unsupervised clustering [8] [16] [34] [36] [28]. This is because most of the web pages belong to maximum one of the several personal

name sharing the same name. Following the same trend, in this paper also the problem of name disambiguation is viewed as hard clustering problem. Justified by this, a semi-supervised learning based name disambiguation process is proposed to disambiguate personal names in the web pages.

The strength of the proposed method includes

- Two stage semi-supervised method that require little prior background knowledge.
- The use of little natural language processing tasks like named entity recognition, information extraction tasks etc

In the second phase, the actual web page collection is subjected to preprocessing, feature extraction, and stop word removal. It then is subjected again to bottom up hierarchical agglomerative clustering along with the clusters formed already with the seed pages using the same Group average agglomerative clustering. The disambiguation accuracy increases since the second phase makes use of clusters generated at the first phase. The method uses well known Elbow method to find the optimum number of clusters needed to be formed. Finally each cluster so formed is considered representing a namesake.

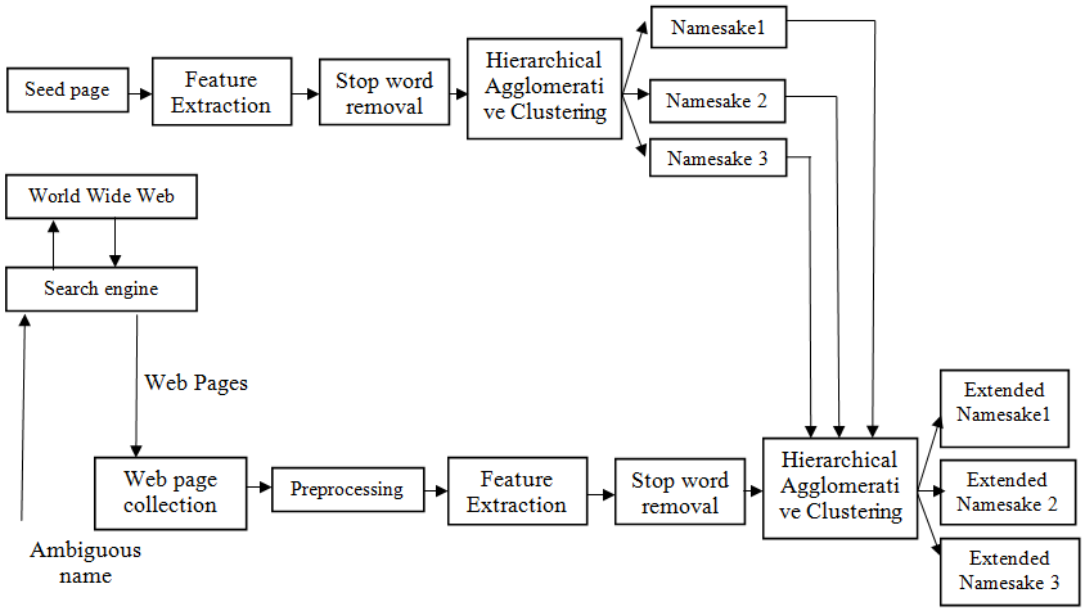

Fig.1. The Process Involved In The Proposed Name Disambiguation Process

II. Method

2.1 Outline of the Proposed work

Fig 1.0 shows the process involved in the proposed name disambiguation system. The process involves two phases namely clustering the seed page and the actual name disambiguation process. A set of seed pages that contains seed pages for each namesakes in the web page cluster are input manually. The feature are extracted from the seed pages and are clustered using bottom up hierarchical agglomerative clustering. The features considered includes nouns (place names, location names, organizational name etc), noun phrases and frequent keywords.

2.2 Preprocessing

First, junk pages are removed i.e pages containing ambiguous names, but that does not refer to real persons, instead referring to either place names or building names etc. (Example Ford, Bloomberg, and, Disney etc). This is followed by removal of Web pages of namesakes that are not considered for disambiguation. For example, there may be 'n' number of namesakes sharing the ambiguous names but for experimental purpose a subset of these 'n' number are considered. This is because it is often impossible to known how many namesakes share a personal name in the web. Thus removing other web pages will reduce misplacing the web pages in other namesake categories.

2.3 Feature Extraction

There are many methods like TF-IDF (term frequency and inverse document frequency), biographic features etc to extract features. Features can be extracted within a window size say within few words or in the same sentence or in the same paragraph or in the whole document in which the ambiguous person name occurs. For this experiment, the entire web page containing the namesake is considered as window, even though there are methods of name disambiguation which considers few words or sentences are context window and extract features within that window. In this proposed method, Nouns, Noun phrases and frequent key words contained in the web pages are extracted and serves as feature vector of the web pages and are subsequently used for the clustering purpose.

2.4 Stop word removal

Stop words that would not contribute to the clustering process are to be removed to improve accuracy of clustering. The presence of stop words overwhelms many other significant words since they frequently occur in the documents. Default stop words like "of", "is" etc are removed from the features list because they does not signify any information.

2.5 Hierarchical Agglomerative Clustering

Hierarchical agglomerative clustering, [37] is widely used clustering in information retrieval tasks. Hierarchical clustering can be top down or bottom up. In bottom up Hierarchical clustering, initially each individual document is assigned to a cluster and then at each step most similar clusters are merged until desired number of clusters are obtained. In hierarchical agglomerative clustering, each document is initially considered as leaf clusters or single cluster and in each clustering iteration the most similar documents are clustered together to form larger cluster. The centroid vector of the cluster is altered after each iteration. In Hierarchical agglomerative clustering (HAC) there is no need to specify the number of clusters in the beforehand. This makes it more suitable for personal name disambiguation in the web pages.

Assign each instance to a separate cluster
Until desired number of clusters
Evaluate pairwise distance between clusters using distance matrix
Look for clusters with shortest pair-wise distance
Merge the shortest pair clusters and delete them in distance matrix
Evaluate the distance between the new cluster with all the other clusters and
Update the distance matrix

Four different types of similarity measures can be employed in agglomerative clustering algorithms namely single link, complete link, group average, and centroid similarity. The proposed method employs group average agglomerative clustering, which is the average distance between web pages in the first cluster and the second cluster.

2.6 Group average Hierarchical Agglomerative Clustering

The objective of the clustering process is to optimize the clustering purity (instances of same class should be placed in the same cluster). Its main idea is to enhance the performance of clustering process by with only a small set of training samples. Group average agglomerative clustering avoids the pitfalls of single link and complete link by averaging the similarities between two clusters.

It is obtained by

$$d(G,H) = \frac{1}{kl}\sum_{i=1}^{k}\sum_{j=1}^{l} d(G_i H_j)$$

Where G and H are two clusters whose similarity distance has to be calculated, k and l are the number of web pages in the cluster G and H.

2.7 Semi-supervised Group average Hierarchical Agglomerative Clustering

Input: A set of web page collection W= {p1, p2…} and set of seed pages S= {s1, s2...} for each namesake in the collection and 'n' the number of clusters (namesakes).

Output: Clusters c1, c2…cn//n is the number of namesakes the collection C represents.

Step 1. Set seed pages as initial clusters of their appropriate namesake it represents
Step 2. For each web page in the collection W

Calculate the similarity between the each web page in C and the cluster W.

If similarity is high then cluster C with the web page, and recomputed the centroid.

Step 3. Repeat the step 2 until each web page is assigned to a cluster and no more changes in the clusters are observed.

2.8 Elbow Method

It is one of the well-known method to determine the optimum number of clusters in the dataset. Its simplicity and ease to use makes it obvious choice for determining the number of clusters in data mining.

For K=1to n
Calculate sum of square errors $SSE = \sum_{i=1}^{k}\sum_{x\in c_i}^{n} (dist(x,c_i)^2)$ between the centroid vector and other vectors.

End of For
Find the elbow value by plotting between SSE and K
As the number of clusters increases (K), the SSE decreases. The elbow value is the value of k, where the SSE value plummets abruptly.

III. Experiments, Results and Discussion

3.1 Performance Metrics

The most commonly used metrics to evaluate clustering algorithms are Purity, Inverse Purity and their harmonic mean (F measure) [38]. WePS systems' performances are evaluated using the standard clustering metrics Purity and Inverse Purity, and the Manual annotation as gold standard. The following metrics are used to measure the performance of the proposed name disambiguation method.

$$Purity = \sum_{i}\frac{|C_i|}{n}\{\max(\frac{|C_i \cap L_j|}{|C_i|})\}$$

Where C the set of clusters, L is the list of classes and n is the total number of documents clustered. Purity penalizes noise in the cluster.

Inverse purity index, [19] rewards grouping items together.

$$InversePurity = \sum_i \frac{|L_i|}{n}\{\max(\frac{|L_i \cap C_j|}{|L_i|})\}$$

F-Score is the harmonic mean of purity and inverse purity.

$$F - Score = \frac{2 * \text{Purity} * InversePurity}{Purity + InversePurity}$$

These three measure can take value from 0 to 1, where 1 represents the optimal value.

3.2 Dataset

The proposed method is tested using both benchmark and synthetic dataset. Bekkerman and McCallum dataset [3], is used as Benchmark dataset and is hereafter referred as dataset – I. The synthetic dataset constructed for the experiments containing web page collection for two ambiguous names "Henry smith" and "Jim Clark" is hereafter referred as dataset –II. The details of these dataset used in the experiments are tabulated in table 2 and table 3.Both the datasets I and II uses gold standard method (manually annotated) for preparing the datasets.

Dataset - I

Bekkerman and McCallum dataset, has for each name 100 web page collection that belongs to 'n' number of namesakes. The number of namesakes sharing the name is in table 1.

Table 1. Statistical Information of Dataset – I.

Name	Namesakes	Web page collection
Adam Cheyer	2	100
William Cohen	10	100
Steve Hardt	6	100
David Israel	19	100
Leslie Pack Kaelbling	2	100
Bill Mark	8	100
Andrew McCallum	16	100
Tom Mitchell	37	100
David Mulford	13	100
Andrew Ng	29	100
Fernando Pereira	19	100
Lynn Voss	26	100

Dataset – II

Two ambiguous personal names "Henry smith" and "Jim Clark" are prevalent in the web and hence they are selected as synthetic datasets. For collecting the dataset, ambiguous personal names were given and web pages were collected from the web. Care is taken to ensure the collection consists of harmonic mixture of web pages of all the names sakes that shares the name. This can be achieved by query of the form "Name" AND "Object", where object is any closely associated word with the name. For each name sake, twenty relevant web pages are collected from the web and are used for the experiments. Table 2 shows list of people sharing the name "Henry smith" in the web and their respective affiliations. Table 4 shows the statistical information of dataset – II.

Table 2. A List of Different Persons Sharing the Name "Henry Smith" in the Web.

Name	Association
Henry Smith	Lynch victim
Henry Smith	British Politician
Henry Smith	Australian footballer
Henry Smith	American footballer
Henry Smith	Harvard Law School
Henry smith	English clergyman
Henry A. Smith	American physician and poet
Henry Smith	Royal Navy officer
Henry L Smith	police force of Ireland
Henry Smith	Mennonite historian

Table 3. Statistical Information of Dataset II.

Name	Namesakes	Web page collection
Henry smith	10	200
Jim Clark	10	200

3.3 Experiments

For each namesake, five web pages are used as seed pages for the clustering purpose. These web pages are first subjected to feature extraction which extracts nouns, noun phrases and frequent keywords and are then clustered using Group average hierarchical agglomerative clustering. Then features are extracted from the web pages in the actual collection. For this experiment, the entire web page containing the namesake is considered as window, even though there are name disambiguation methods which considers few words or sentences are context window and extract features within that window. It is then followed by stop words removal, if any present as features. The stop word free features are then used for clustering purpose.

Bottom up hierarchical agglomerative clustering that uses group average method for finding similarity between clusters is used for clustering. The clustering algorithm at each stage, merges the existing cluster with the most similar cluster present. Hard clustering is followed, since each feature instances belongs to only one cluster. Euclidean distance is used for measuring the similarity between clusters. Since the existing clusters in the first phase is used along with the web page collection, the accuracy of second clustering that includes the actual web

page collection to be clustered is improved. The process of clustering is stopped when the required number of clusters is obtained as found out using elbow method. In the experiments, we did not encounter any web page representing ambiguous person containing more than one namesakes. Thus it vindicates the stance that name disambiguation is a hard clustering problem.

3.4 Experimental results and Discussion

Experiments are conducted for both dataset –I and dataset – II. The performance of the proposed method in comparison with the existing techniques is shown in table 4. The performance of the proposed system is evaluated against noteworthy similar techniques including works that uses two stage clustering [24], using Lexical, linguistic and personal features [36], using unsupervised method [16], using tokens, Named entities, Urls [28]. Along with related works, the performance of the proposed method is compared with baseline clustering algorithms like simple agglomerative clustering and k means clustering using keywords and nouns as features. The performance of these techniques on dataset – I is tabulated in table 4. Agglomerative clustering and simple k-means clustering followed by extracting frequent terms as features performs well.

Table 4. Performance of Various Methods Using Dataset –I.

Features	Clustering	Purity Index	Inverse Purity Index	F-Score
Nouns, Noun phrases and keywords (Proposed method)	Hierarchical agglomerative clustering	0.66	0.49	0.56
Two stage clustering	Hierarchical agglomerative clustering	0.63	0.50	0.55
Tokens, Named entities, Urls	Hierarchical agglomerative clustering	0.61	0.47	0.53
Unsupervised method	Bottom up agglomerative clustering	0.57	0.50	0.53
lexical linguistic and personal features	Bottom up agglomerative clustering	0.53	0.42	0.46
Frequent terms and nouns	Agglomerative clustering	0.47	0.39	0.42
Frequent terms and nouns	K-Means clustering	0.43	0.41	0.41

It is evident from the table 4 that proposed method outperforms other previous works in name disambiguation process. Among the other methods, two stage clustering algorithm that uses two subsequent clustering procedure to improve the accuracy of clustering performs clustering with more accuracy than other methods. Most of the other methods uses nouns and hence their accuracies are not disappointing.

By considering the same related techniques, the experiment is repeated just merely by changing the dataset. The accuracy of the proposed method on dataset – II is shown in the table 5.

Table 5. Performance of Various Methods Using Dataset –II.

Features	Clustering	Purity Index	Inverse Purity Index	F-Score
Nouns, Noun phrases and keywords	Hierarchical agglomerative clustering	0.64	0.67	0.65
Two stage clustering	Hierarchical agglomerative clustering	0.61	0.64	0.62
Unsupervised method	Bottom up agglomerative clustering	0.59	0.63	0.60
Tokens, Named entitites, Urls	Hierarchical agglomerative clustering	0.59	0.57	0.57
Lexical, linguistic and personal features	Bottom up agglomerative clustering	0.52	0.51	0.51
Frequent terms and nouns	Agglomerative clustering	0.45	0.47	0.45
Frequent terms and nouns	K-Means clustering	0.39	0.43	0.40

The proposed methods' purity and inverse purity index values are higher compared to other name disambiguation techniques. Apart from the proposed method, two stage clustering and unsupervised name disambiguation methods performs well on dataset- II.

IV. CONCLUSION

In this paper, a semi-supervised two stage clustering based name disambiguation process is proposed that uses rich set of informative features like Nouns, Noun-phrases and Keywords as features from webpages. Using a few web pages of the interested namesake containing ambiguous name as seed pages, the proposed method uses Group average Hierarchical agglomerative clustering to cluster web pages into different namesakes. To find the number of clusters to be formed, well known elbow method is used. The proposed methods' efficiency is validated on synthetic and organic dataset containing ambiguous names along with the other well-known existing methods. Results show that the proposed method outperforms other similar methods in terms of purity, inverse purity and f-score.

Interesting future work includes improving the disambiguation accuracies of place names, organizational names, etc. Usage of disambiguation in information retrieval, information extraction, and question and answering etc will certainly increase the accuracy of these systems. Exploiting Web structure, web usage pattern against that person and social network of that person may provide useful to disambiguate ambiguous personal name. It should also be noted that key words alone cannot distinguish a persons, number of persons are there who are having same name, belonging to same fields residing the same country. Use of potent features will help in disambiguation such ambiguous names. We hope that this work will open more avenue on personal name disambiguation research.

REFERENCES

[1] Spink, Amanda, Bernard J. Jansen, and Jan Pedersen. "Searching for people on Web search engines." *Journal of Documentation* 60, no. 3 (2004): 266-278.

[2] Bollegala, Danushka, Yutaka Matsuo, and Mitsuru Ishizuka. "Automatic discovery of personal name aliases from the web." *IEEE Transactions on Knowledge and Data Engineering* 23, no. 6 (2011): 831-844.

[3] Bekkerman, R. "Name data set." (2005).

[4] Liu, Zhengzhong, Qin Lu, and Jian Xu. "High performance clustering for web person name disambiguation using topic capturing." *Ratio* (2011).

[5] J. Artiles, J. Gonzalo, and F. Verdejo, "A Testbed for People Searching Strategies in the WWW," Proc. SIGIR, 2005.

[6] Artiles, Javier, Julio Gonzalo, and Satoshi Sekine. "The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task." In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 64-69. Association for Computational Linguistics, 2007.

[7] V. Hatzivassiloglou, P. Duboue, and A. Rzhetsky. Disambiguating proteins, genes,and rna in text: A machine learning approach. In *Proceedings of the 9th International Conference on Intelligent Systems for Molecular Biology*, Tivoli Gardens, Denmark, July 2001.

[8] Bekkerman, Ron, and Andrew McCallum. "Disambiguating web appearances of people in a social network." In *Proceedings of the 14th international conference on World Wide Web*, pp. 463-470. ACM, 2005.

[9] Fleischman, Michael B., and Eduard Hovy. "Multi-document person name resolution." In *Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics (ACL), Reference Resolution Workshop*, pp. 66-82. 2004.

[10] An, Ning, Lili Jiang, Jianyong Wang, Ping Luo, Min Wang, and Bing Nan Li. "Toward detection of aliases without string similarity." *Information Sciences* 261 (2014): 89-100.

[11] http://nlp.uned.es/weps

[12] http://start.csail.mit.edu/index.php

[13] Matsuo, Yutaka, Junichiro Mori, Masahiro Hamasaki, Takuichi Nishimura, Hideaki Takeda, Koiti Hasida, and Mitsuru Ishizuka. "POLYPHONET: an advanced social network extraction system from the web." *Web Semantics: Science, Services and Agents on the World Wide Web* 5, no. 4 (2007): 262-278.

[14] Chen, Ying, S. Yat Mei Lee, and Chu-Ren Huang. "Polyuhk: A robust information extraction system for web personal names." In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*. 2009.

[15] Bagga, A., & Baldwin, B. 1998. Entity-based cross-document coreferencing using the vector space model. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, 79–85. Association for Computational Linguistics.

[16] Mann, Gideon S., and David Yarowsky. "Unsupervised personal name disambiguation." In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 33-40. Association for Computational Linguistics, 2003.

[17] Popescu, Octavian, and Bernardo Magnini. "Irst-bp: Web people search using name entities." In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 195-198. Association for Computational Linguistics, 2007.

[18] Cucerzan, Silviu. "Large-Scale Named Entity Disambiguation Based on Wikipedia Data." In *EMNLP-CoNLL*, vol. 7, pp. 708-716. 2007.

[19] Bunescu, Razvan C., and Marius Pasca. "Using Encyclopedic Knowledge for Named entity Disambiguation." In *EACL*, vol. 6, pp. 9-16. 2006.

[20] Ted Pedersen, Amruta Purandare, Anagha Kulkarni. Name Discrimination by Clustering Similar Contexts. In Proceedings of CICLing, 2005.

[21] Chong Long, and Lei Shi. "Web Person Name Disambiguation by Relevance Weighting of Extended Feature Sets." In *CLEF (Notebook Papers/LABs/Workshops)*. 2010.

[22] Vu, Quang Minh, Atsuhiro Takasu, and Jun Adachi. "Improving the performance of personal name disambiguation using web directories." *Information Processing & Management* 44, no. 4 (2008): 1546-1561.

[23] Lu, Zhao, Zhixian Yan, and Liang He. "OnPerDis: Ontology-Based Personal Name Disambiguation on the Web." In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, vol. 1, pp. 185-192. IEEE, 2013.

[24] Ikeda, Masaki, Shingo Ono, Issei Sato, Minoru Yoshida, and Hiroshi Nakagawa. "Person name disambiguation on the web by two-stage clustering." In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*. 2009.

[25] Nuray-Turan, Rabia, Dmitri V. Kalashnikov, and Sharad Mehrotra. "Exploiting web querying for web people search." *ACM Transactions on Database Systems (TODS)* 37, no. 1 (2012): 7.

[26] Han, Xianpei, and Jun Zhao. "CASIANED: web personal name disambiguation based on professional categorization." In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, pp. 2-5. 2009.

[27] Sugiyama, Kazunari, and Manabu Okumura. "Titpi: Web people search task using semi-supervised clustering approach." In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 318-321. Association for Computational Linguistics, 2007.

[28] Elmacioglu, Ergin, Yee Fan Tan, Su Yan, Min-Yen Kan, and Dongwon Lee. "Psnus: Web people name disambiguation by simple clustering with rich features." In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 268-271. Association for Computational Linguistics, 2007.

[29] Zhang, Duo, Jie Tang, Juanzi Li, and Kehong Wang. "A constraint-based probabilistic framework for name disambiguation." In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 1019-1022. ACM, 2007.

[30] GUHA, R., and A. GARG. 2004. Disambiguating people in search. *In* 13thWorldWideWeb Conference, Stanford University Stanford, CA.

[31] Minkov, Einat, William W. Cohen, and Andrew Y. Ng. "Contextual search and name disambiguation in email using graphs." In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 27-34. ACM, 2006.

[32] Lee Ingyu, Byung-Won On, and Seong No Yoon. "Algebraic Algorithms to Solve Name Disambiguation Problem." In *DMIN*, pp. 468-474. 2009.

[33] Lu, Yiming, Zaiqing Nie, Taoyuan Cheng, Ying Gao, and Ji-Rong Wen. "Name disambiguation using Web

connection." In *Proc. of AAAI*. 2007.
[34] Bollegala, Danushka, Yutaka Matsuo, and Mitsuru Ishizuka. "Automatic annotation of ambiguous personal names on the web." *Computational Intelligence* 28, no. 3 (2012): 398-425.
[35] Balog, K., Azzopardi, L., & Rijke, M. de. 2005. Resolving person names in web people search. *Weaving services and people on the World Wide Web*, 301–323.
[36] Wan, Xiaojun, Jianfeng Gao, Mu Li, and Binggong Ding. "Person resolution in person search results: Webhawk." In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 163-170. ACM, 2005.
[37] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Vol. 1. Cambridge: Cambridge university press, 2008.
[38] Amigó, Enrique, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. "A comparison of extrinsic clustering evaluation metrics based on formal constraints." *Information retrieval* 12, no. 4 (2009): 461-486.

Authors' Profiles

P.Selvaperumal received his Bachelor degree (2006) in Computer science from Sacred heart college, Tirupattur and Master degree in Computer science (2009) from Bannari Amman Institute of Technology, Sathyamangalam and a second masters in technology in Veltech Technical university, Chennai. He is currently a Ph.D student pursuing Computer engineering in M.S University, Tirunelveli, Tamilnadu, India. His areas of interest include Text Mining, Machine Learning, NLP, Data science, big data exploration and Information Retrieval. He is an ACM Student member and a member of Indian society of technical education (ISTE).

A.Suruliandi received his B.E(1987) in Electronics and Communication Engineering from Coimbatore Institute of Technology-Coimbatore, Bharathiyar University, Tamilnadu, India, M.E (2000) in Computer Science and Engineering from Government College of Engineering-Tirunelveli, Manonmaniam Sundaranar University, Tamilnadu, India, Ph.D (2009) from Manonmaniam Sundaranar University as well. He started his academic career in 1987 and held his various positions in the Department of Computer Science, Kamaraj College, Tuticorin, Tamilnadu, India. Currently he is working as Professor in Department of Computer Science and Engineering Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India. He has published many papers in international journals and conferences. His research interest includes Datamining, pattern recognition, image processing, and remote sensing.

