ISSN:0975-9646

An Approach to Cluster Web Pages for Personalization of Web Search

Amol Rajmane[1#], Pradeep Patil[2$]

[#] *Computer Science and Engineering Dept., Ashokrao Mane Group of Institutions, Vathar, Dist.- Kolhapur,India*

[$] *KJ's Educational Institutes, Kondhwa-Saswad Road, Pune, India*

***Abstract*— The traditional search engines organize search results into clusters for ambiguous queries, representing each cluster for each meaning of the query. The clusters are obtained according to the topical similarity of the retrieved search results, but it is possible for results to be totally dissimilar and still correspond to the same meaning of the query. People search is also one of the most common tasks on the Web nowadays, but when a particular person's name is queried the search engines return web pages which are related to different persons who have the same queried name. It is the quality of clustering algorithms that are used to disambiguate different web pages of the namesake. The clustered approach is not a better option all the way. The quality of clustering algorithms decides the approach is a better solution or not. If clusters identified by the search engines corresponded to a single person, then the cluster based approach would be a good choice, but if clusters contained pages of multiple persons which may merge into same cluster or pages of same person spread over multiple clusters. In such cases the advantages of cluster based approach are not obvious.**

By placing the burden on the user of disambiguating and collecting pages relevant to a particular person, in this paper, we have developed an approach that clusters web pages based on the association of the web pages to the different people and clusters that are based on generic entity search.

***Keywords*—Entity resolution, information retrieval, graph based disambiguation, web people search, clustering.**

I. INTRODUCTION

IT is highly needed to represent web search results effectively as it is an open problem in the information retrieval community. For each meaning of the query, traditional search engines organize search results of ambiguous queries into groups or clusters. Existing search engines like Yahoo, Google and Bing often display a long list of search results, ranked by their relevancies to the given query. It is the job of web user to go through the results and check the titles and snippets serially to identify their required results. This becomes worse when user wants to search for entities as a common activity on Internet to search for people i.e. people search based on related information like location, organization or other types of entities.

Searching information related to a person has gained substantial growth nowadays. Over 5% of the web pages are related to a person in the current Web searches [1]. A search for a person will return pages relevant to any person with the name mentioned in the query.

By using entity search we are able to browse and analyze the returned information in a more structured way by which we can enhance web search capabilities and user experience. Consider searching for the web pages for an ordinary person. This can take extra efforts to find out exact results as search engine like Google returns first few pages only of famous persons. By using clustering approach, the famous person pages can be combined into a single cluster.

The main important issue is the quality of clustering algorithms that are used to disambiguate different web pages of the namesake. The clustered approach is not a better option all the way. It is the quality of clustering algorithms which decides this approach is a better solution or not. If clusters identified by the search engines corresponded to a single person, then the cluster based approach would be a good choice, but if clusters contained pages of multiple persons which may merge into same cluster or pages of same person spread over multiple clusters. In such cases the advantages of cluster based approach are not obvious.

The algorithm that we have used extracts significant entities such as other persons, organizations and locations on each webpage. This is useful to form relationships between persons associated with the webpage and the entities extracted. The algorithm then analyzes the relationship along with features such as TF/IDF, and other useful content like hyperlink information which is used to disambiguate the pages.

II. OVERVIEW OF THE APPROACH

In this section, we provide an overview of the proposed system for web search system. We have used middleware-based approach to implement our system. The proposed approach accepts user's query by specialized web interface. After this, the middleware forwards this query to the search engine via the search engine API and a fixed number of relevant web pages are retrieved. The retrieved web pages are then preprocessed by computing TF/IDF which includes stemming, stop word removal, noun phrase identification, and inverted index computation. After this is done, the web pages named entities and web related information is extracted. The data extracted on the preprocessing step is then used to generate entity-relationship graph, this graph is used by clustering algorithm along with TF/IDF values and model parameters. The clustering algorithm disambiguates the set of (K) web pages. The output generated from this algorithm is a set of

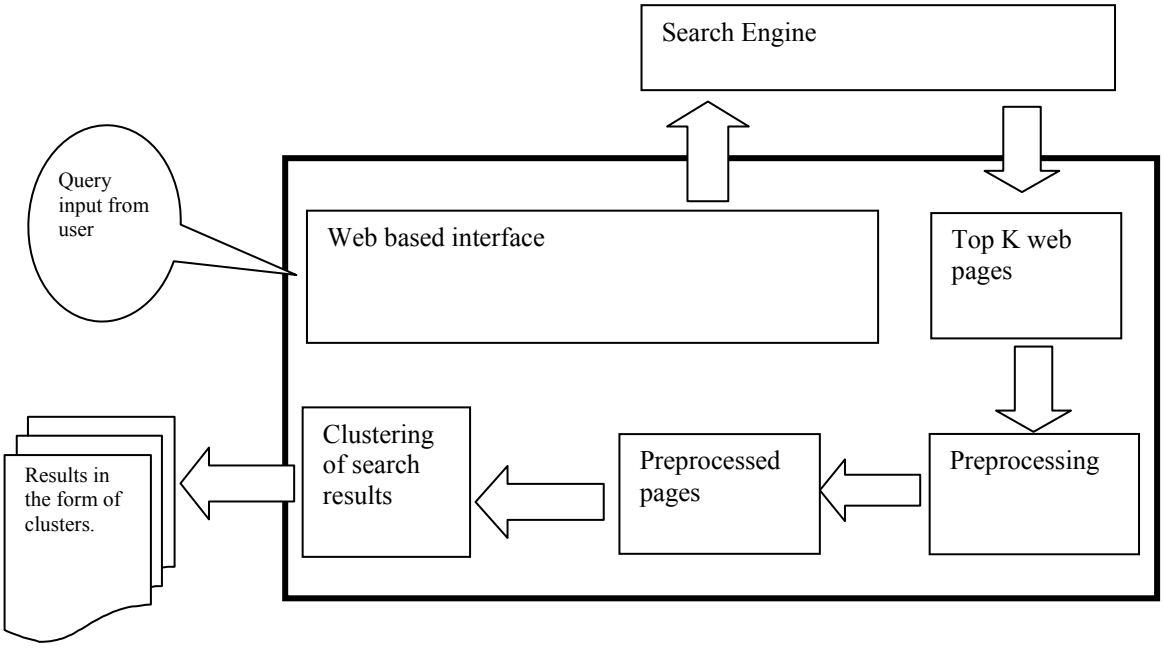

Fig.1 The middleware approach of the query processing

clusters of these pages. Once the clusters are obtained, each cluster is then processed by generating sketches, the set of keywords that represent the web pages within a cluster. After that, based on chosen criterion, all clusters are ranked. If the user gets satisfied with particular cluster, the web pages in this cluster are presented to the user.

III. GRAPH CREATION

The main goal of our work is to analyze entities relationships and features present in the dataset for disambiguation. We need to extract entities and the relationships between the entities. For that purpose, we have used information extraction software. The hyperlinks and email addresses also are extracted from the web pages. The graph is generated whose nodes correspond to the relationships between entities and web pages. After the extraction of named entity, a node gets created for that named entity which represents all related named entities which have same name. One node represents the collection of persons that have the same name. For example, the person "Pradeep Patil" might be extracted from two or more different web pages. A single node for "Pradeep Patil" may have the two pages referring to the same person or to two or more different people. The node is used to represent the group of persons that share the same name. For the locations and organizations also nodes get created. For each of the web page in top *K* web pages a node is created. After this, relationship edge is added between a node representing a web page and a node corresponding to each named entity extracted from that webpage. The relationship edges are type distinct in nature. The relationship edge between a webpage node and a person node will have type distinct from relationship edge between a webpage node and an organization node or location node.

IV. CORRELATION CLUSTERING ALGORITHM

The Correlation Clustering (CC) algorithm [11] is used to group the nodes that represent the web pages that belong to the same person. CC assumes that there is similarity function *s(u,v)* for any objects. If *u* and *v* are similar to each other then they return similarity function *s(u,v)*. By using past data the similarity function is learned.

There are varieties of disambiguation approaches exist which can be applied for various applications. The traditional approach analyzes object features to decide two objects descriptions corefer [15], [19]. The second approach is relational approach, in which dependencies among co reference decisions get analyze [18], [21]. The disambiguation algorithm that we have proposed is based on object features and the ER graph for the data set. We can observe improved quality of disambiguation by analyzing the ER graph with the analysis of object features [14], [16], [17].

The clustering problem is represented as fully connected graph in which each object becomes a node in the graph. According to the similarity function *s(u,v)* similar "+" or different "-" label is assigned to each edge *(u,v)*. We have to find out the partition of the graph into clusters which agrees the most with the assigned labels. If *s(u,v)* is assigned "+" and "-" labels perfectly to the edges inside the clusters as well as outside the clusters, the right can be obtained by the removal of all the negative edges in the graph, all remaining components of the graph will represent the right clusters. When *s(u,v)* is not perfect it mislabel some of the edges, CC is used in these cases.

In order to define similarity function *s(u,v)* we need to use notion of Connection Strength *c(u,v)* between two objects *u* and *v*. We need to analyze features of two objects *u* and *v* because the similarity of the features of *u* and *v*

defines certain affinity/attraction between those objects *f(u,v)*and if this attraction is large, then the objects are said to be same. The same idea is applied to analyze paths. It is assumed that each path between two objects carries in itself a certain degree of attraction. A path between *u* and *v* semantically captures interactions between them via intermediate entities. If the combined attractions of all these paths are sufficiently large, the objects are likely to be same.

The connection strength *c(u,v)* is used to measure attraction between two nodes *u* and *v* via paths which is defined as the sum of attractions contributed by each path[1].

$$c(u,v) = \sum_{p \in P_{uv}} w_p \quad (1)$$

P_{uv} denotes the set of all *L-short* simple paths between *u* and *v* and *Wp* denotes the weight contributed by path *p*. If length of the path does not exceed *L* then path is *L-short* and if it does not contain duplicate nodes then it is simple. From each type of path, the weight path *p* contributes is derived. Two paths having the nodes of the same type connected via edges of the same type are considered to be of the same path type. There is possibility that number of possible path types, for *L-short* simple path is limited for each domain. Let W_k be the attributes associated with a path of type *k*. Let P_{uv} consist of c_1 paths of type 1, c_2 paths of type 2, ……, c_n paths of type *n.* Then, (1) can be written as

$$c(u,v)=c_1w_1 +c_2w_{2+}...+c_nw_n \quad (2)$$

In order to minimize mislabeling of data the similarity function *s(u,v)* should be powerful. The similarity function is designed such that it will be able to automatically self-tune itself to the particular domain being processed. Function *s(u,v)*is constructed as a combination of the connection strength *c(u,v)* and feature similarity *f(u,v)*

$$s(u,v)=c(u,v) + \gamma f(u,v) \quad (3)$$

The labeling of data is done by comparing *s(u,v)*value against the threshold τ, where τ is a non negative real number.

V. Related Work

There is plenty of work done on disambiguation, entity resolution. A review of the main work is presented here, but the review is not exhaustive. In [10], it is proposed that if clusters contained errors (multiple people merged into same cluster) the advantages of cluster based approach are not obvious. A novel algorithm is developed for the disambiguating people who have same name. The algorithm establishes relationships in between the person who is associated with the web page and entities extracted like name of the persons, organizations and locations on web page. The authors have compared clustered entity search implemented by using disambiguation algorithm with traditional people search supported by current search engines. Whereas [2] developed a disambiguation algorithm & then studied its impact on people search. The proposed algorithm extracts significant entities such as names, organizations & locations on each web page by using extraction techniques. It extracts & parses HTML & web related data on each web page. The algorithm uses Entity-Relationship Graph where entities are interconnected via relationships. The algorithm does analysis of several types of information like attributes, interconnections which exist among entities in the ER Graph.

A Web People Search (WePS) approach is proposed in [3]. It is based on collecting co-occurrence information from the web. In order to make clustering decisions, a sky-line based classification technique is developed which classifies the collected co-occurrence information. The dominance in the data is handled by this technique and adopts quality measure of clustering.

In [4], the goal is to group search engine return citations, in a manner citation in each group relates to the same person. This is based on the three facets: attribute, link & page similarity. In this, the confidence matrix is constructed for all facts then a grouping algorithm is applied on final confidence matrix for all facets. The citations belong to each person get collected in each group. Such groups are output of this technique. The authors in [5] have proposed a framework which tackle various information retrieval and Web mining tasks. The proposed framework makes the heuristic search viable in the vast domain of the WWW and applicable to clustering of Web search results and to Web appearance disambiguation. In [6], the authors have presented a graphical approach for entity resolution. The overall idea behind this is to use relationships & to look at the direct and indirect (long) relationships that exist between specific pairs of entity representations in order to make a disambiguation decision. In terms of the entity-relationship graph that means analyzing paths that exist between various pairs of nodes.

The problem of entity disambiguation in the Web setting has been explored in some research efforts [7], [8], [9], [10]. Applications for Web people search can be implemented by using two different approaches. In server side approach disambiguation mechanism is integrated into the search engine directly. The other approach is middleware approach in which we use existing search engine to implement web people search by wrapping the existing search engine. The middleware uses the search engine API to retrieve the top *K* relevant web pages for user query. Then it clusters the web pages based on their associations to real entities.

In this paper, we have employed clustering approach whereas in [16], [17] and [20] completely different solutions are proposed for different types of problems. A Fuzzy Lookup disambiguation challenge has been solved in [16], [17] and [20] whereas we need to address different type of disambiguation challenge known as Fuzzy Grouping. In Fuzzy Lookup, the list of objects is provided to the algorithm, whereas for grouping there is no list available and the references that co-refer are grouped [12], [13].

There are some of the search engines available that return results in clusters. Search engines Clusty (http://www.clusty.com), Grokker (http://www.grokker.com) return clustered results, but the clusters are obtained on the intersection of broad topics like family pages form one cluster, job related pages could form another cluster. These all search engines cluster the results based on entire web page content. Search engine ZoomInfo(http://www.zoominfo.com) is similar to the one proposed by us in this paper. In order to identify different people on the web, it extracts named entities by applying machine learning and data mining algorithms.

There is exploitation of relationships for disambiguation in [24], is somewhat similar to our approach. In this a sketch of each web page which represents a person name is constructed by using variety of external data sources such as DBLP and TAP are used. This approach is restricted to only person searches and our approach does not depend on any such precompiled information and thus we can scale to person search for any person on the web. Similar kind of effort is found in [1], [4], [8], [22], [23] and [25]

In [8] the link structure of pages on the Web is exploited, assuming that web pages belonging to the same real person are tightly linked together. For disambiguation three algorithms are presented, the first exploits the link structure of the pages, the second algorithm is based on word similarities between documents, and the third approach combines link analysis with A/DC clustering.

VI. Experimental Results

A. Experimental Setup

We have chosen middleware approach to access the results from the Web. The system is implemented in Java and all experiments were conducted on run on IBM xenon server with 2.4 GHz processor and 4GB RAM.

The system first downloads and preprocesses pages before applying the actual clustering algorithm. This takes 4.21 seconds per web page. The preprocessing cost can be minimized if we apply server side approach instead of middleware approach. It is somewhat difficult to compare effectiveness of cluster- based people search to the traditional search engine, as so many factors are unknown like background information the user knows and their intention to use that information in the query.

TABLE I
STATISTICS ON 2005 DATASET

Personal name	Position	No. of pages	No. of Categories	No. of relevant pages
Adam Cheyer	SRI Manag	97	2	96
William Cohen	CMU Prof	88	10	6
Steve Hardt	SRI Eng	92	19	20
David Israel	SRI Manag	92	19	20
Leslie Pack Kaelbling	MIT Prof	89	2	88
Bill Mark	SRI Manag	94	8	11
Andrew McCallum	UMass Prof	94	16	54
Tom Mitchell	CMU Prof	92	37	15
David Mulford	Standford Undergrad	94	13	1
Andrew Ng	Stanf Prof	87	29	32
Fernando Pereira	UPenn Prof	88	19	32
Lynn Voss	SRI Eng	89	26	1
	OVERALL:	1085	187	420

Dataset:

We have collected dataset available from www.cs.umass.edu/~ronb[8]. This dataset consists of 1085 Web pages which are collected from Melinda Gervasio's social network which consists of 12 names of people. According to the person's occupation, the dataset is labeled. Out of 12 people, two people are unique on the Web, while rest have relatively common names. There are some names that appear extremely ambiguous, e.g. given a query "Tom Mitchell", 37 different Tom Mitchells were found within the first 100 Google hits. There are web pages of 187 unique people in the dataset, while only 12 of them were relevant. The detailed statistics and overall process is given in [8]. These 12 names are then issued as queries to the Google and first 100 pages were retrieved for each query issued. We did manual filtration of the pages by removing pages in non-textual formats, HTTPD error pages and empty pages. After that we labeled the remaining pages by the occupation of the individuals whose name appeared in the query. TableI shows statistics of the dataset.

Apple Dataset

For clustering web search results, we built a new dataset. We issued query "apple" to the Google search engine and labeled first 100 hits obtained from Google. In the obtained 100 pages we found 23 different categories, out of those categories the largest ones are the apple watch, the music and the fruit. Table II presents the statistics on this dataset.

Clustering of Web search results

The problem of clustering Web search results is not considered as one class problem as compared with Web appearance disambiguation problem. To evaluate our system, we have chosen k largest classes of the data. We have chosen k= 3 for our apple dataset and we have built three clusters namely apple watch, fruit, and music. Let CCi be one of these clusters and Cl_i is its corresponding

TABLE II
CONTEXT DATA SET RESULTS

				Baseline		Our algorithm	
Query	Query Text	# Pages	#	B-cubed	Fp	B-cubed	Fp
Q1	"Andrew McCallum" music	100	29	53.5	69.6	73.8	81.7
Q2	"Andrew McCallum" poster	100	4	87.6	93.5	76.5	86.6
Q3	"Andrew McCallum" dance	100	30	57.7	68.7	65.0	75.1
Q4	"Andrew McCallum" uci	100	1	78.3	88.9	95.9	98.0
Q5	"George Bush" bible scholar	98	13	61.7	77.5	74.1	84.6
Q6	"William Cohen" cmu	100	7	89.4	94.5	91.1	95.4
Q7	"William Cohen" uci	100	17	71.7	82.4	55.1	73.8
Q8	"Tom Mitchell" psychology	98	17	63.2	78.4	76.0	85.5
Q9	"Tom Mitchell" soccer	97	40	56.6	69.3	69.6	77.9
Mean		99	18	68.9	80.3	75.2	84.3

class. Let $Corr_i$ be a set of pages from Cl_i which is correctly assigned into *CCi* by the system we have implemented. Then the precision and recall of the system are defined as:

$$Prec = \frac{\sum_{i=1}^{k}|Corri|}{\sum_{i=1}^{k}|CCi|}; \qquad Rec = \frac{\sum_{i=1}^{k}|Corri|}{\sum_{i=1}^{k}|Cli|}$$

We have compared our work with two other searches namely sequential search [8] and incremental search algorithm [9]. After three iterations the sequential exhaustive search fails; out of 100 pages 70 pages are all connected together.

The incremental algorithm shows better results; the precision drops after obtaining 82.4% precision. When we apply our algorithm, we find better result, especially after the first iteration. The resulting system shows better performance up to 93.3%, while the F-measure consistently goes on improving from 56% to 59% and then to 62 % at the third hops from the source pages.

B. Testing Disambiguation Quality

To assess the quality of disambiguation, we have used the B-cubed and *Fp* measures. B-cubed is more fine grained and is better measure than *Fp* and many other measures. We have used baseline method which is Agglomerative Vector Space clustering algorithm with TF/IDF. This method is widely used as benchmark to evaluate such kind of tasks.

Disambiguation Quality: Queries with Context

We have used context keywords that are related with person and generated dataset by querying that person with context from Google. Nine different person queries are used. The details are given in Table II. This table shows that for query Q_5= "Georg Bush" bible scholar, 98 meaningful pages were found and these pages contain 13 namesakes for Georg Bush. The results of disambiguation quality for the proposed approach and baseline algorithms are shown in Table II. There is significant improvement 6.3 percent can be seen over B-cubed measure.

Quality of Generating Cluster Sketches

Our algorithm represents keywords to summarize each cluster. Table III shows the output of the algorithm for the "Andrew McCallum" query on WWW 2005 dataset. Stemmed versions are used to show the keywords and phrases. In the table only top 10 keywords for each cluster are shown. As we can see each cluster has different set of keywords, therefore each cluster can be easily differentiated from the other cluster for the same person name.

User Observations for Web People Search

For the web people search user queries the search engine with the name of the person the user is interested in retrieving theweb pages for that person e.g. Sharad Pawar, and in order to satisfy the objective of finding all the web pages of that person among the top *K* pages, he scans through the *K* pages. The user is able to decide whether the page is relevant to the person he is looking for or irrelevant at each observation *i*, where *i*= 1, 2,…, *K*. usually this scenario is used in traditional search. In the cluster based people search, the user first looks the interface, then user sequentially reads cluster information, until on the m^{th} observation he is able to find the cluster which he is interested in. when user opens that cluster, system shows the original set of *K* web pages returned by the search engine, only in this case the web pages are taken from set of pages *S* that our system has identified for that namesake. User decides on the sketches for the cluster are relevant or irrelevant in a similar fashion as with the standard interface. It is possible that the user can make mistake in deciding the relevant/ irrelevant based on cluster sketches, this happens because none of the cluster matches the person he is interested in. In such cases the user is able to retrieve original *K* web pages that are returned by the search engine. The reported quality measures will be optimistic for both the new and standard interfaces.

The quality of the interface that we have implemented and the standard interface are compared using Precision, Recall and F-measure. The precision shows the fraction of relevant pages among all the web pages examined on the i^{th} observation.

TABLE III
FOR "ANDREW MCCALLUM" QUERY IN THE WWW2005 DATASET

Group Name	Cluster Summary
UMASS Professor	Learn,artificeintellig, machin, proceed, machine learn, extract, model, classif, comput, data
ACOSS President3	Student, incom, univers, educ, fee, famili, low,east, timor, timor, cent
ACOSS President11	Acos, childcar, welfare, Australian, council, service, social, service, income, family, president
Teacher	Aclandburghlei, Camden, burghlei, sch, acland, school, English, month, biographi, uk
Writer	Competit, stori, read, toowrit, winner,author, mouse, nep, toowritepoetri, children
Artist3	Philosophi, mentalfoss, festiv, science, blog, pietersen, utim, coburg, guthri, flyer
Artist1	Rockbox, jukebox, archo, studio, tedford, stuart, bod, boru, dobaghi, melih
Photographer	Amico, imag, collect, library, conspir, davidrumsei, penitentiary, trial, Williamston, court
Kid	Theatre, shakspear, tempst, grouch, Juliet, romeo, crew, dream, festiv, night
Medical Professor1	Ccfp, kari, kgh, leroyv, med puddi, queensu, jennif, em, md
Customer Support	Initil,opensr,domain, tucow, loui, dn, chronology, protect, sent, client
Humanist	Dreambook, humanist, color, human, ge, homepage, plz, secular, vacat, individu
Painter	Height, imag, larger, price, Sherwin, keith, cub, leopard, sefton, Richard
ACOSS President1	Hospit, nurs, health, Australian, service, treatement, Kingston, acid, local, care
Medical Professor1	Inquest, Ontario, coron, ministri, death, eastern, ontario, union
Poll Analyst	Declan, Zealand, fc, jul, video, game, horn, censorship, fitug, office
Poll Analyst	Regul, electron, transact, swain, act, notic, paper, disclosur, discuss
Economist	Acidif, soil, cost, farm, acid, agriculture, land, econom, lime, research
Technician	Chemistri, dapart, otago, chemic, comput, laboratory, univers, calm, comput support cooper work

The recall shows the fraction of related web pages out of all the related pages discovered so far on the i^{th} observation. By using our interface the user starts examining the first web page only on the $(m+1)^{th}$ step, when it locates the exact cluster on the m^{th} step. In order to find out observations that are needed to discover a certain fraction of relevant pages, recall plots are useful. The user is able to find the related pages if the observations that he needs to do are fewer and that interface is better which gives fewer observations.

The system is assessed for a number of real-world queries; also we have analyzed the results obtained from our system with respect to certain characteristics of the input data. The queries are mainly categorized in four types such as ambiguous, general, compound and people name. The system is tested for all these queries & the result obtained is satisfactory. Some sample queries are taken for the testing of system. Table IV shows the name of query under each type, its clusters obtained. From the list of clusters obtained, only first 15 clusters are shown in Table IV.

TABLE IV
CLUSTERS FOR DIFFERENT TYPE OF QUERIES

Query Type	Queries	Obtained clusters
Ambiguous	Mouse	Computer Mouse, Mickey Mouse, Website, Cells from Mouse, Gaming Mouse, Common Mouse, House Mouse, Technology, Gene, Graphics, Series, Apple, Magic Mouse, Support, Windows
General Terms	Yellow pages	Search Yellow pages, Local Business Directory, White Pages, Local Business Listings, Phone numbers,MapsDirections,TelephoneDirectory,Companies,Classifieds,International,People,Sourse,City Guides, Complete, Global
Compound Query	To be or not to be	Google Buchsuche-Ergebnisseite, Games, Tobe T, Face book, Name Tobe, News, Reviews,Seiten,Service, People, Play Games, Question, Search, Tobe, Hooper, Years
People Name	Pratibha Patil	PratibhaPatilNews,Woman Behind Woman, New Delhi, Woman, Governor of Rajasthan, PratibhaPatil Photos, Patil to visit, National, PratibhaPatilPictures,rashtrapatiBhavan,Video,FemalePresident,ManmohanSingh,AbdualKalam,Presidental Candidate
	G.A Patil	G.A., PratibhaPatil, Department, Georgia GA, Internal Medicine, Journal, Indian National Congress, R.T., S.L., A.Y., Nagana Gowda, Address & Phone Number, Education, PaR, Book

VII. CONCLUSION

In this paper, we have proposed a novel approach to cluster web pages based on their association to different people. Our technique disambiguates among the namesakes referred to on the web pages by exploiting different information extracted from web pages such as named entities and hyperlinks. We have processed information by taking only top-*k* web pages. In future we can use external information like ontologies, encyclopedias and the web to disambiguation. We can also apply more advanced extraction capabilities that would bea better interpretation of extracted entities by taking role of each individual with one another. Now our algorithm relies on co-occurrence relationships, in future we will focus on extraction of relationships. We will work on other search problems that have different settings.

REFERENCES

[1] R. Al-Kamha and D.W. Embley, "Grouping Search-Engine Returned Citations for Person-Name Queries," Proc. Int'l Workshop Web Information and Data Management (WIDM), 2004.
[2] D.V. Kalashnikov, S. Mehrotra, Z. Chen, R. Nuray-Turan, and N. Ashish, "Disambiguation Algorithm for People Search on the Web," Proc. IEEE Int'l Conf. Data Eng. (ICDE '07), Apr. 2007.
[3] D.V. Kalashnikov, R. Nuray-Turan, and S. Mehrotra, "Towards Breaking the Quality Curse. A Web-Querying Approach to Web People Search," Proc. SIGIR, July 2008.
[4] J. Artiles, J. Gonzalo, and F. Verdejo, "A Testbed for PeopleSearching Strategies in the WWW," Proc. SIGIR, 2005.
[5] R. Bekkerman, S. Zilberstein, and J. Allan, ―Web Page Clustering Using Heuristic Search in the Web Graph, Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), 2007.
[6] Z. Chen, D.V. Kalashnikov, and S. Mehrotra, ―Adaptive Graphical Approach to Entity Resolution, Proc. ACM IEEE Joint Conf. DigitalLibraries (JCDL), 2007.
[7] J. Artiles, J. Gonzalo, and F. Verdejo, "A Testbed for People Searching Strategies in the WWW," Proc. SIGIR, 2005.
[8] R. Bekkerman and A. McCallum, "Disambiguating Web Appearances of People in a Social Network," Proc. Int'l World Wide Web Conf. (WWW), 2005.

[9] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Extracting Key Phrases to Disambiguate Personal Names on the Web," Proc. Int'l Conf.Intelligent Text Processing and Computational Linguistics (CICLing),2006.
[10] D.V.Kalashnikov, S.Mehrotra, R.N.Turen and Z.Chen, "Web People Search via Connection Analysis" IEEE Transactions on Knowledge and data engg.Vol 20, No11, Novr 2008.
[11] N. Bansal, A. Blum, and S. Chawla, "Correlation Clustering," Foundations of Computer Science, pp. 238-247, 2002.
[12] I. Bhattacharya and L. Getoor, "Iterative Record Linkage for Cleaning and Integration," Proc. ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery (DMKD), 2004.
[13] I. Bhattacharya and L. Getoor, "A Latent Dirichlet Model for Unsupervised Entity Resolution," Proc. SIAM Data Mining Conf. (SDM), 2006.
[14] Z. Chen, D.V. Kalashnikov, and S. Mehrotra, "Exploiting Relationshipsfor Object Consolidation," Proc. Int'l ACM SIGMOD WorkshopInformation Quality in Information Systems (IQIS), 2005.
[15] I. Fellegi and A. Sunter, "A Theory for Record Linkage," J. Am.Statistical Assoc., vol. 64, no. 328, pp. 1183-1210, 1969.
[16] D.V. Kalashnikov and S. Mehrotra, "Domain-Independent Data Cleaning via Analysis of Entity-Relationship Graph," ACM Trans. Database Systems, vol. 31, no. 2, pp. 716-767, June 2006.
[17] D.V. Kalashnikov, S. Mehrotra, and Z. Chen, "Exploiting Relationships for Domain-Independent Data Cleaning," Proc. SIAM Int'lConf. Data Mining (SDM '05), Apr. 2005.
[18] A. McCallum and B. Wellner, "Conditional Models of Identity Uncertainty with Application to Noun Coreference," Proc. Ann. Conf. Neural Information Processing Systems (NIPS), 2004.
[19] H. Newcombe, J. Kennedy, S. Axford, and A. James, "Automatic Linkage of Vital Records," Science, vol. 130, pp. 954-959, 1959.
[20] R. Nuray-Turan, D.V. Kalashnikov, and S. Mehrotra, "Self-Tuning in Graph-Based Reference Disambiguation," Proc. Int'l Conf. Database Systems for Advanced Applications (DASFAA '07), Apr. 2007.
[21] P. Singla and P. Domingos, "Multi-Relational Record Linkage," Proc. Workshop Multi-Relational Data Mining (MRDM), 2004.
[22] C. Tiwari, "Entity Identification on the Web," technical report, Indian Inst. Technology, 2006.
[23] X. Wan, J. Gao, M. Li, and B. Ding, "Person Resolution in Person Search Results: Webhawk," Proc. ACM Conf. Information and Knowledge Management (CIKM), 2005.
[24] R. Guha and A. Garg, Disambiguating People in Search. Stanford Univ., 2004.
[25] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Extracting Key Phrasesto Disambiguate Personal Names on the Web," Proc. Int'l Conf. Intelligent Text Processing and Computational Linguistics (CICLing), 2006.