ISSN: 2321-7782 (Online)
Volume 3, Issue 6, June 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Document Retrieval Using Person Name and Alias Based Bipolarisation Based on PCA

Rachna Sable
Department of Computer Engineering
G H Raisoni Institute of Engineering & Technology
Pune - India

Abstract: ***A topic within any document is associated usually with the time, place and person. Generally rather than simple topics, topics that involve bipolar or competing views are continuously reported now-a-days and are inattention. In order to help users understand the topics more easily it is very important to identify the relation or association between persons mentioned in the topics. Even the respective aliases of the person name should be identified to get the complete information of the person. An unsupervised approach is used, in which PCA (Principal Component Analysis) is used to partition the person names along with their aliases into bipolar groups from the set of given documents. Two techniques weighted correlation coefficient and off topic block elimination are used to reduce the unwanted blocks and data sparseness. Topic timeline system is used to arrange the topic according to its activeness chronologically to show its developments in time.***

Keywords: Bipolar relation, Bipolar timeline, Topic time lining.

I. INTRODUCTION

Searching information about the people is the most common task generally carried on the internet or Web. If two people share the same name or if the person has many nick names or aliases, the task of searching information about the person becomes very difficult. Knowledge sharing is one of the aims of Web and therefore day by day topic documents on the Web are increasing. Even online news, weblogs, discussion forums forms a major portion of information on the internet today. Topics involving bipolar or competing viewpoints are now a day's in attention. The users who are not familiar with the topic may spend a large amount of time to fully comprehend the topic and find out the association between two people [1].

A statistical technique called Principal Component Analysis (PCA) is used that automatically identifies the bipolar groups of person names along with their aliases from a set of topic documents. Identifying aliases of names is very important and can be done using lexical pattern-based approach using snippets returned by a web search engine. Out of various aliases received for a particular person numerous ranking scores methods like lexical pattern frequency, word co-occurrences and page count can be used. When PCA process the textual data, major problem is sparseness of text features. So the techniques like Weighted Correlation Coefficient and Off topic block elimination is used. Weighted Correlation Coefficient is used to denote the block in which the two persons co-occur or does not occur. Off topic block elimination is used to remove the unwanted block which does not contain the related person's names in the document. The activeness trend for each bipolar group can be measured by using Activeness Timeline of bipolar groups to find whether the group is active or inactive with time.

II. RELATED WORK

Various online news, weblogs, discussion forums are available and there is a need for automatic techniques for analyzing and presenting the news or the topic to the users in a meaningful and efficient manner. Extracting important themes in the documents of interest is the work done by existing topic mining approaches. Time, place and person are the three attributes

associated with the topic. So to identify the persons uniquely and form the clusters of unique persons the related work done is as follows:

"Person resolution in person search results: Web hawk" in this Wan, Gao, Ding [2] developed Web Hawk system in order to provide person searches on the internet or web. The various information extraction techniques are used to obtain email-id, names, organization titles, job from the web pages, which are then combined with lexical features in order to disambiguate person names. Web hawk does the work in four steps:

1. The pages that contain no information about the related person are removed by *filter*.

2. The pages are then grouped into different clusters, each for one specific person. This task is done by *cluster*.

3. Query oriented personal information can be extracted from each page by *extractor*.

4. For users to find specific information /person easily, descriptive information is generated by *Namer*.

The performance of the system is increased by using Web hawk and even we get the benefit as noisy data is removed. The system focuses only on English person names other language person names were not identified. Unwanted block removal was not done and even timeline method for showing the activeness development of topic was not used.

Using the person name to search the information on web is the most common activity. As more than one people share the same name, the person name ambiguity arises due to which, the search engine returns web pages containing information about more than one person. The task of getting the desired information related to person becomes a difficult task in this case. In the paper "Towards breaking the quality curse: a web- quering approach to web people search, ", [3] to get the required information user is then forced to go through each and every individual document or add some extra terms to query so as to disambiguate the person and get the required information. The paper addresses the challenges of Heterogeneity, where same entity is represented in multiple ways and Ambiguity where same type of representation of multiple entities is done. Precision is improved. As context for different entities are different on the basis of which they can be differentiated, but the system may not work well if contexts of different entities significantly overlap. The person name clustering described above is differentiated by the proposed approach as it generates clusters that posses bipolar orientations.

In order to identify the feature pattern in datasets of high dimensions a well known statistical technique called Principal Component Analysis is used. The relation matrix Cij is calculated by using correlation coefficient. Single eigenvector called Principal Eigen vector is used to cluster persons into bipolar groups. To partition person names into bipolar groups signs in PCA's Principal Eigenvector are effective. The major problem with PCA while processing textual data is the sparseness of text features & to avoid these problem two techniques weighted correlation coefficient & Off topic block elimination is used.

Statistical text mining & language models suffer from data sparseness."Introduction to Information Retrieval*,"* [4] with same polarity person names data sparseness could lead to underestimation and for opposite polarities data sparseness could lead to overestimation. So to alleviate language models data sparseness problem Laplace law was developed but it was not appropriate for PCA, as PCA correlation coefficient measures divergence of person names from their means and the divergence will not change if we add one to each person's vector entry. So weighted correlation coefficient is used which separates blocks into occurring and non co-occurring parts. A parameter α is set between 0 & 1 to differentiate the blocks in which two persons occur or do not occur & accordingly calculate persons correlation. To eliminate the unwanted or off topic blocks, a centroid of all decomposed blocks are taken by averaging bi's and a predefined threshold β is taken. The blocks whose cosine similarity to the centroid is less than β are excluded and more than β are included.

To extract feature patterns in terms of eigenvectors PCA is not the only method [1].

Other Eigen Vector based methods

The relation matrix Cij is calculated by using Inner product or Cosine similarity. Use of more than one eigen vector is made to identify feature cluster. Signs of individual eigenvector are considered.

To describe the development of topic included in a number of topic documents a timeline is constructed by topic timeline mining. "Bursty and hierarchical structure in streams," [5], *in* this a mining technique is proposed in which topic documents are constructed as a series of topic timeline. Hidden Markov Model (HMM) is used in case documents contain bursty information. Bursts with sharp boundaries can be identified and it helps to get clear beginning and end. To model the activeness status of the topic HMM is used which splits the status into active themes & accordingly trees nodes & branches are modeled. Chen & Chen "ISCAN: a novel method for topic summarization and content anatomy," [6] proposed an Eigen vector based approach in order to identify themes which are important in topic documents. To determine the degree of correlation between topic block, amplitude of an entry in an Eigen vector is used.

A person is represented by various name aliases on the web. In order to find all references of a single entity we need to go for alias extraction "Approximate Personal Name-Matching through Finite- State Graphs" [7]. For finding or extracting variants or abbreviations of person names approximate string matching algorithms are used. To compare names, methods like edit-distance based and rules in the form of regular expressions are used "Adaptive Duplicate Detection using Learnable String Similarity Measurable," [8]. But such approaches of string matching did not identified aliases if they didn't share any word or letters with original name. So a method is proposed to automatically generate such lexical patterns using names & aliases training data set.

After finding the aliases the ranking of aliases is done. In the paper "Automatic Discovery of Personal Name Aliases from the Web" [9], three different approaches are used to define ranking scores and to measure the association between person name and his aliases.

Lexical pattern frequency: If person name and its aliases occur in many lexical patterns then it is a good alias for that person. Candidate aliases are ranked in descending order of the number of different lexical patterns in which they appear with name.

Co-occurrences in Anchor text: Along with the concise text, they also provide links which can be considered as expressing a citation.

*Page count based association measures***:** Word association measures that not only considers co-occurrences in anchor text but in the web overall.

III. Implementation Detail

A. *Proposed system*

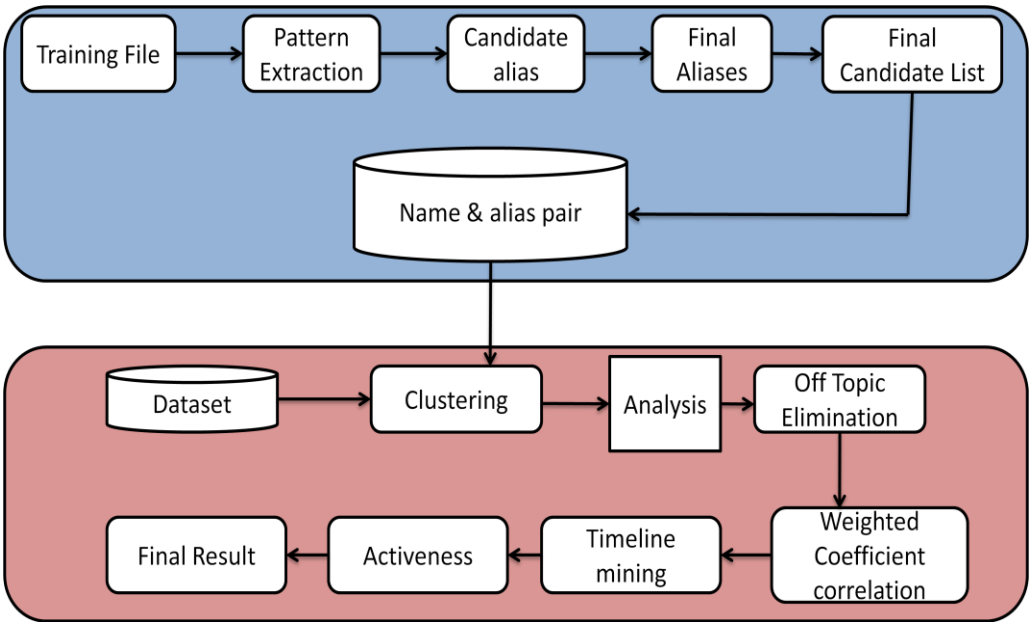
A topic is usually associated with attributes such as time, place and person. Searching for information about people in the web is one of the most common activities of internet users. However, when a person has nick name or name aliases or even if people have more than one aliases then retrieving information from web search engines about people and associating him or her to proper documents can become difficult. Generally, topics that involve bipolar or competing viewpoints are attention getting and are thus reported in a large number of documents.

The proposed system tries to

- Identify the association between important persons mentioned in numerous topics along with their aliases by using PCA.
- Elimination of unwanted blocks by off-topic block elimination.

- Reduction of data sparseness.
- Sequencing the events by using Time lining.

B. Block Diagram



IV. METHODOLOGY

The Proposed system can be categorized by the following modules:

A. Pattern Extraction

To extract aliases of a given name using snippets, Lexical pattern based approach is used.

Algorithm: Pattern Extraction Method.

S1 = {N, A, D}

Where N - Name of the person

A - Aliases of the person

D - Web documents

ALGORITHM: EXTRACT PATTERNS(S)

1. Comment: S is a set of (Name, Alias) It is a pair of (N, A)
2. C ← null
3. For each (pair i.e. name, Alias) ε S
4. $do \begin{cases} D \leftarrow \text{Get Snippets ("NAME*ALIAS")} \\ \quad \text{for each snippet } d \in D \\ \text{do } P \leftarrow P + \text{Create Pattern (d)} \end{cases}$
5. Return (P)

B. Candidate Alias

Using extracted patterns for getting the list of candidate aliases.

Algorithm: Extract Candidates Aliases.

S 2 = {N, P, Ngram}

Where P- Extracted Patterns

N- Name of the Person

Ngram – Length of the alias name by default.

ALGORITHM: EXTRACT CANDIDATES (Name, P)

1. Comment: P is a set of Patterns returned from S1.
2. C ← null
3. For each pattern p ε P
4. $do \begin{cases} D \leftarrow \text{Get Snippets ("NAME p * ")} \\ \quad \text{for each snippet d} \in D \\ \text{do C} \leftarrow C + \text{GetMgrams (d, NAME, p)} \end{cases}$
5. Return (C)

C. Alias Ranking

To find the correct aliases among the extracted candidate aliases. The strength of association between the name and the alias can be measured by the following co-occurrences statistics:

ALGORITHMS: CF, Cosine and Dice, Even Hub Discounting is used.

CF is the Co-occurrence frequency

S3 = {P, C, V, x}

Where P – Set of person names

C - Set of candidate list

V – Set of all words that appear in the anchor text

x – It is the Set of specific aliases

The output parameters of the above modules are

k- It is the co-occurrence frequency between x and p

K–It is the Sum of co occurrence frequencies between x and all words in V

n – It is the value same between p and all candidates in C

N – It is the total co-occurrences between all words pairs taken from C and V.

C – {x} = denotes all candidates except x

V-{P} = denotes all words except the given name p.

Cosine

S4 = {X, Y, D}

Where X – Denote the occurrences of candidate alias

Y – It is the occurrences of name p

D – Set of documents

Cosine(p,x) = k/(√n √k) Denotes measure of association between a name and a candidate alias

Dice

S5= {X, Y, D}

Where X- Denote the occurrences of candidate alias

Y – Denotes the occurrences of name p

D – Set of documents

Dice(p,x) = 2k/(n+k) Denotes ranking scores

Hub Discounting

Hub is the frequently observed phenomena in web; where in many pages with diverse topics are linked to. To reduce its adverse effect while measuring co-occurrences measures, co occurrences of words linked with h is multiplied with α where

S6 = {h, t, d}

Where h is the hubs

t – It is the inbound anchor texts of h that contain the real name p

d – Denotes the total number of inbound anchor texts of h.

$$\alpha(h, p) = \frac{t}{d}$$

D. Final Candidate list

Final subset of ranked aliases.

S7 = {Sum and Average of S3, S4, S5, S6}

E. Cosine Similarity

Extract features from text data, Correlation Coefficient Relationship between one document to other in form of 0 or 1.

S8 = {Set of documents}

F. Weighted Correlation Coefficient

Relation between topics of documents in the form of float value that is weight.

S9 = {P, D}

Where, P is the set of person names

D is the set of documents

$Corr_w$ (Pi, Pj) represents the weighted Correlation coefficient between the persons pi and pj.

Co (i,j) Denotes the block set in which persons i and j co-occur.

G. Lagrange Multiplier

For Matrix reduction.

H. Eigen values & Eigen vectors

To extract features using PCA from text document.

I. Off topic block elimination

To remove unwanted blocks from the document. 0 remove the block or paragraph and 1 to include the block.

J. Activeness Timeline of bipolar groups

Shows the intensity and activeness development of the identified bipolar person groups with time.

S = {g, t}

Where, g represents activeness of a bipolar group at time t.

Activeness $_{g,t}$ Means number of person name Occurrences bipolarized to g at t.

Large Activeness $_{g,t}$ means group is active and small Activeness $_{g,t}$ means group is inactive and does not attract many reports.

V. RESULTS

Data Set

The Text file containing the pair of Name and Alias is given as input to first module and patterns extracted from this module are provided as input to other modules and the associated documents for these modules are taken from web.

Result Set

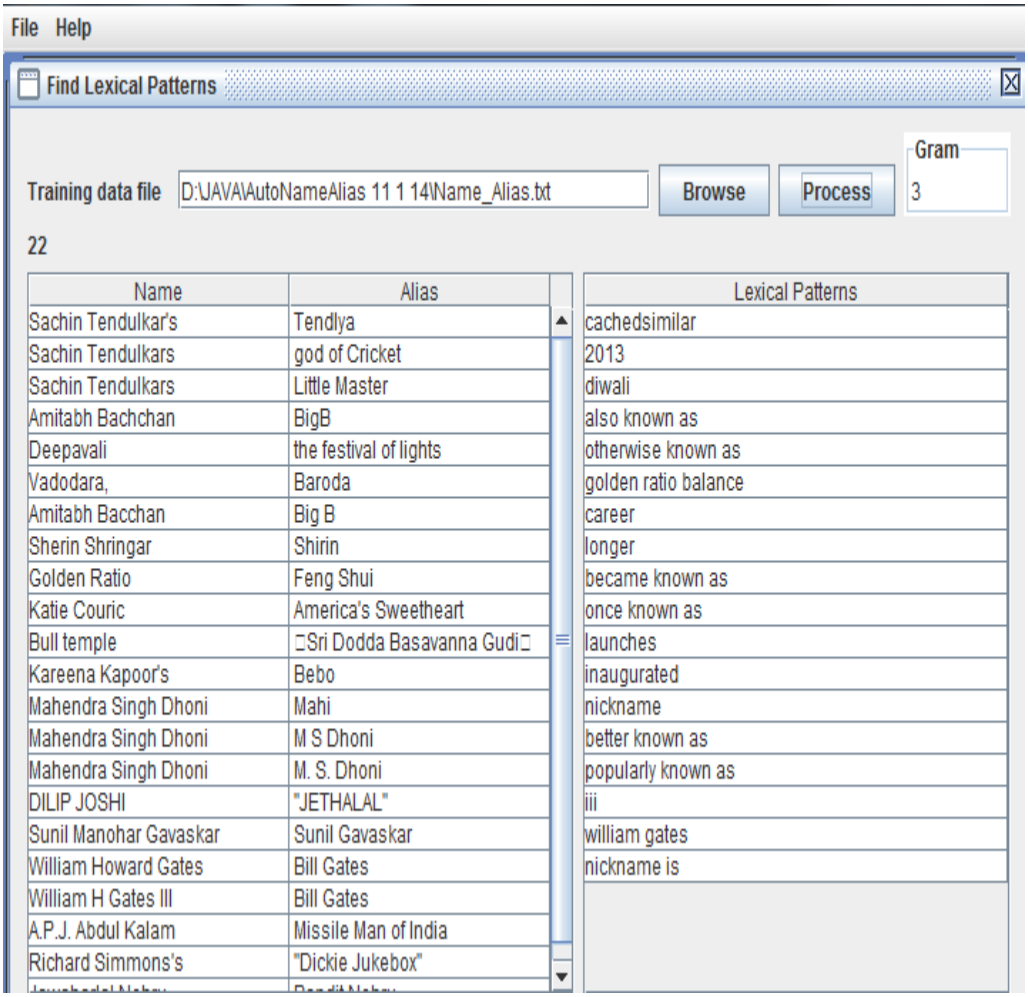


Fig 2 Output of pattern extraction

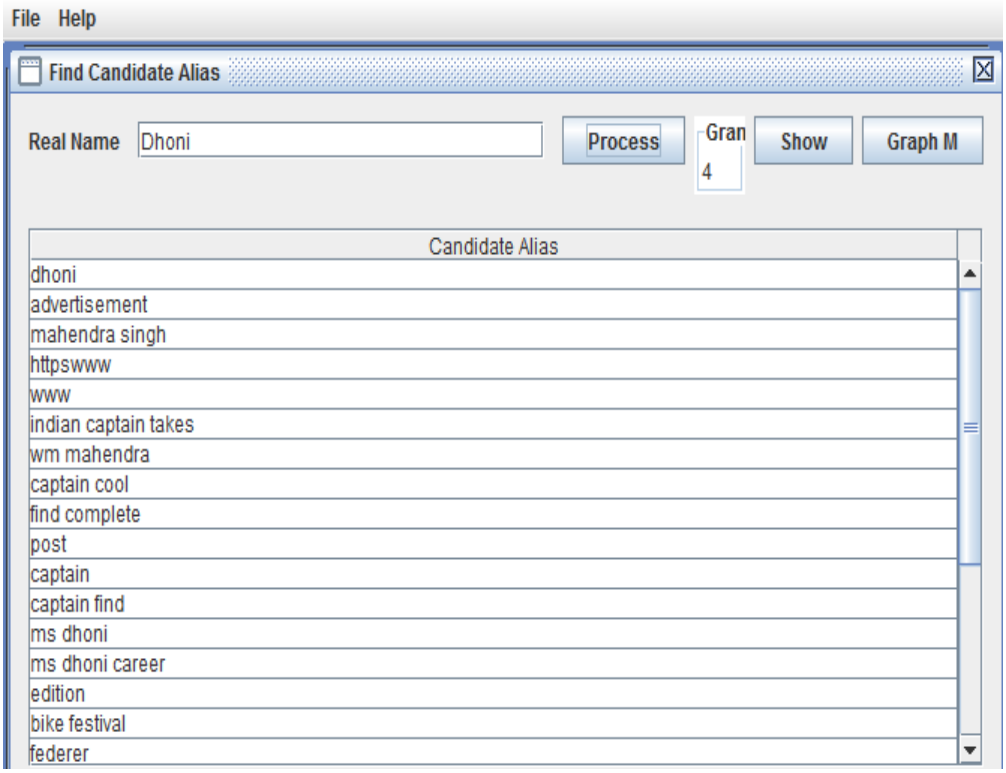


Fig 3 Output of candidate extraction

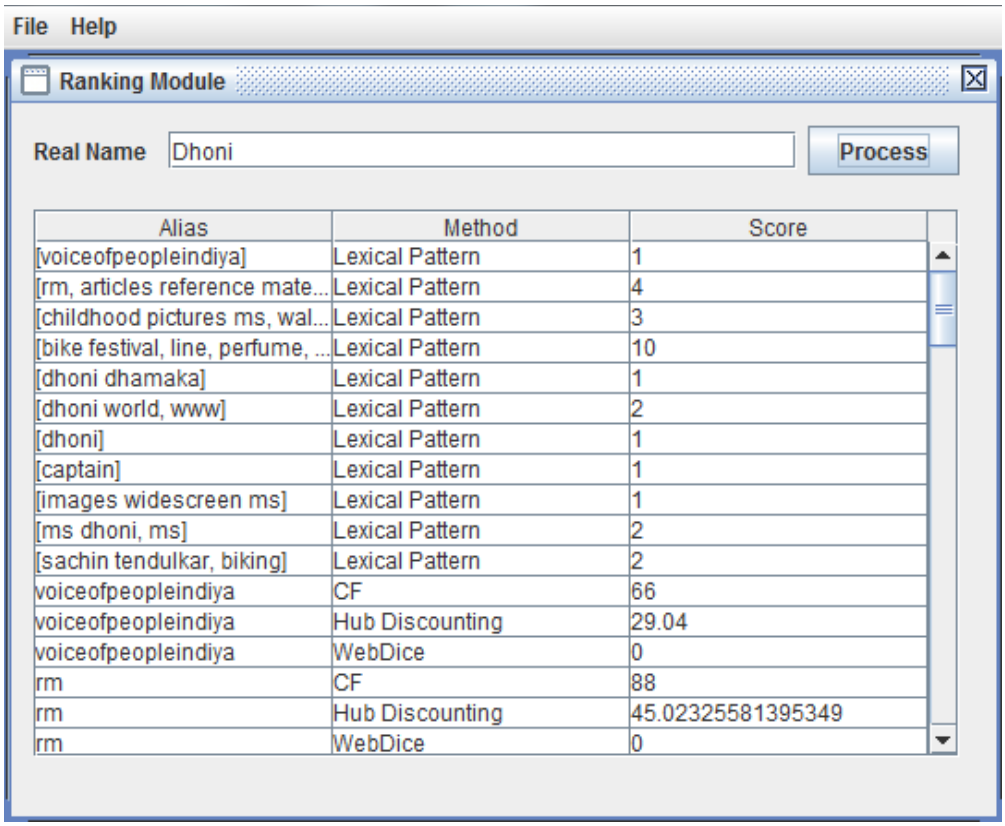


Fig 4: Output of Alias Ranking

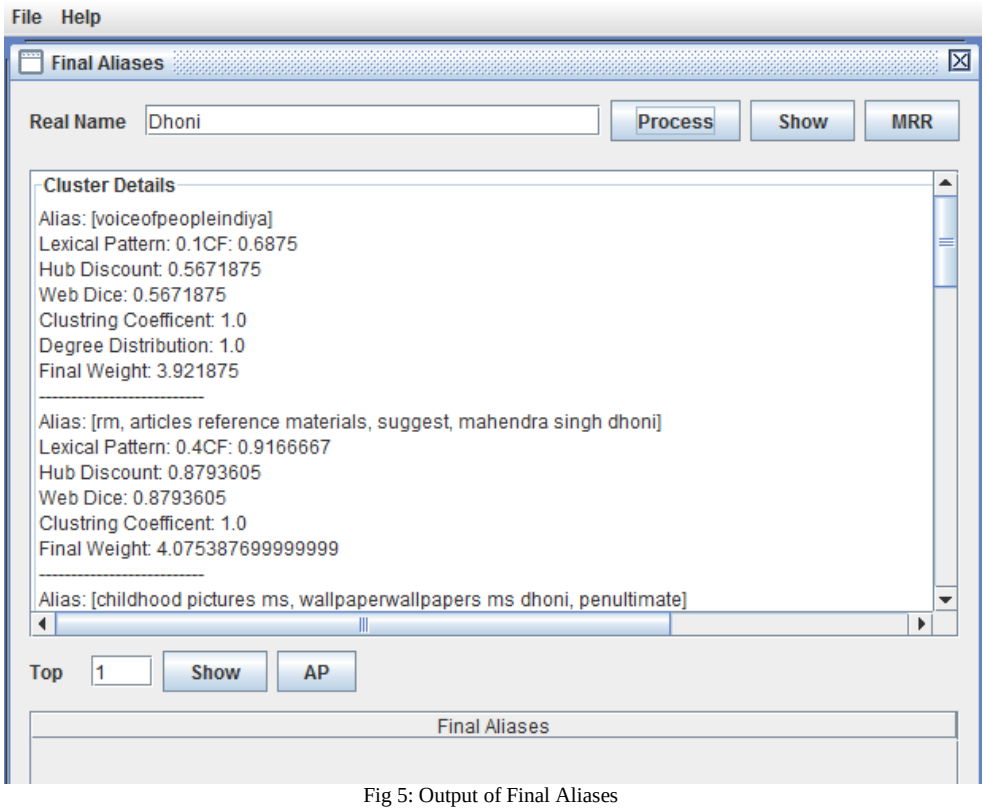


Fig 5: Output of Final Aliases

VI. Conclusion and Future Work

In order to help users understand the topics more easily it is very important to identify the relation or association between persons mentioned in the topics. The system provides the efficient way for finding the person names along with their aliases in the documents or blocks for fast retrieval of information and better understanding of topic. It has provided a practical framework for person name alias based bipolarization which can be used in a real environment. Topic timeline system is used to arrange the topic according to its activeness chronologically to show its developments in time. Also, we believe that the methodology developed here tells us that using sophisticated techniques can be quite useful in person name bipolarization and finding their positive or negative orientation and finally some of the evaluated person names possessed neutral orientations. Developing an effective method to identify neutral persons in topics would be the area for future work.

Acknowledgement

I would like to express my sincere gratitude and thanks to my dear friends and husband for their help, encouragement and scholarly influence which made this paper possible. Their invaluable guidance and suggestions have been instrumental in successful completion of this paper.

References

1. Chien Chin Chen,Zhong-Yong Chen, Chen-Yuan Wu, " An Unsupervised Approach For Person Name Bipolarization Using Principal Component Analysis,"
2. X. Wan, J.Gao, M. Li, and B. Ding, "Person resolution in person search results: Web hawk," Proceedings of the 14th ACM international Conference on information and Knowledge management. Pp.163-170, 2005.
3. D.V. Kalashnikov, R. Nuray-Turan, and S. Mehrotra, "Towards breaking the quality curse: a web- quering approach to web people search, "Proceedings of the 31st Annual international ACM SIGIR conference on Research and Development in information Retrieval, pp.27-34, and 2008.
4. C. D. Manning, P. Raghavan and H. Schutze, Introduction to Information Retrieval, Cambridge University Press, 2008.
5. J. M. Kleinberg, " Bursty and hierarchical structure in streams," Proceedings of the eight ACM SIGKDD international conference on knowledge discovery and data mining, pp-101,2002
6. C. C. Chen, and M. C. Chen, "ISCAN: a novel method for topic summarization and content anatomy," Proceedings of the 31st annual international ACM SIGIR Conference on Research and development in Information retrieval, pp.579-586, 2008.
7. C. Galvez and F. Moya-Anegon, "Approximate Personal Name-Matching through Finite- State Graphs" J.Am. Soc.for Information Science and Technology, vol.58, pp.1-17,2007
8. M. Bilenko and R. Mooney, "Adaptive Duplicate Detection using Learnable String Similarity Measurable," Proc. SIGKDD' 03, 2003.
9. Danushka Bollegala, Yutaka Matsuo, Mitsuru Ishizu ka, "Automatic Discovery of Personal Name Aliases from the Web"2011.
10. L. I. Smith, A Tutorial on Principal Components Analysis, Cornell University 2002.

11. R. Nallapati, A. Feng, F. Peng and J. Allan, "Event Threading within new topics," Proceedings of the thirteenth ACM international conference on information and knowledge management, pp. 446-453, 2004.
12. G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. J. Miller, "Introduction to WordNet. An On-line Lexical Database," International Journal of Lexicography, Vol 3, issue 4, pp.235-244, 1990.
13. R. Guha and A. Garg, "Disambiguating People in Search," technical report, Stanford University, 2004.
14. G. Mann and D. Yarowsky, "Unsupervised Person Name Disambiguation," Proc. Conf. Computational Natural Language Learning (CoNLL '03), pp. 33-40, 2003.

AUTHOR(S) PROFILE

Rachna Sable is working as Asst. Professor in G. H. Raisoni Institute of Engineering & Technology, Pune. She has a working experience of 7 years and has been core member in the development of students.She received the M.E. degree in Computer Engineering from Savitribai Phule Pune University, Maharashtra in 2014.

