ISSN:0975-9646

Resolving Referential Ambiguity on the Web Using Higher Order Co-occurrences in Anchor-Texts

Rama.K, Sridevi.M, Vishnu Murthy.G

Department of Computer Science and Engineering
CVSR college of Engineering, Anurag group of Institutions
Venkatapur(V),Ghatkesar(M),Andhra Pradesh, India

Abstract— **Retrieving information about famous personalities is a common task among internet users. Finding information from web search engines becomes difficult when those people are referred by other names on the web because information about people in the web pages exist using their alias names. So by just giving the real name in search won't retrieve all the alias related information. This is the referential ambiguity problem. For this reason precise identification of aliases of a given person is important in many tasks such as information retrieval, identification of relations among entities, sentiment analysis, name disambiguation and semantic annotation related to web. The previous approaches extracted aliases for a given person which resulted in achieving a high mean reciprocal rank (MRR) and an improvement in recall. In order to achieve a good improvement in the MRR and recall compared to the previous approach we propose a system which extracts aliases by not only considering the first order co-occurrences but also the higher order co-occurrences among the anchor texts for a given name and alias which will help in the expansion of a query for retrieval of relevant results. This method will rank the aliases retrieved based on the different statistics scores calculated for a name and its corresponding alias in the anchor texts retrieved. The co-occurrences order will be known by constructing and mining an anchor text graph for a particular name and its associated aliases. We use two data sets person names and location names and for ranking the aliases we use a ranking support vector machine.**

Keywords—**Information Retrieval, Referential Ambiguity, Anchor Text Co-occurrence Graph, Web crawler, Graph Mining**

I. Introduction

Browsing for information about people is a day to day activity among internet users. According to a survey most of the searches to the web search engines include queries about person names or location names [1] [2]. The problem comes when the person they are searching about has got various aliases. Internet users, bloggers etc refer famous personalities like film stars, politicians, sports person, musicians, and famous authors using different names depending on their popularity. Alias names of a person can be anything related to their profession, title, role, pen name or the movie name in which they acted in etc. This problem of a single person being called by different names is known as referential ambiguity [5]. This is a big problem in information retrieval and name disambiguation tasks. For example if we consider film actor Sharukh Khan he is referred by different names on the web by bloggers. For example news paper tabloids refer him as Sharukh khan but when it comes to the web he is referred as “King khan”, “King of Bollywood”, ”Bollywood Badhshah”, ”Srk”, ”Don” or simply as “Mr.Khan”. Now when a user searches for Sharukh khan on the internet he will retrieve information web pages related to the search term Sharukh Khan. But the problem here is the user won't be able to retrieve the documents in which no reference is made of his aliases .Improving the retrieval of information is important to provide the user with more relevant information. There might be a chance where a user already knows information about a particular person .Providing the aliases might help the user in retrieving more important information which is not known about the person. This can be an additional advantage to the user. The aliases extracted will also help the semantic web to annotate its content with more information where the aliases will be added as meta-data [3] to the already present information. Moreover the extracted aliases will also help in disambiguating people easily in the search process.

Contributions to the proposed method can be described as follows:

We propose a method which extracts the aliases from the web in an automatic way. 1) The proposed method extracts the aliases from the anchor texts retrieved by using the web crawler. It extracts all the inbound anchor texts of an URL containing the information of a given name. 2) To extract the correct aliases from the anchor texts and to measure the association between the anchor texts for a given name and its aliases we use co-occurrences statistics. 3) The higher order co-occurrence between a name and an alias can be known by constructing the anchor text co-occurrence graph. This graph represents the words contained in the anchor texts 4)Next measuring of the hop distances for a name and its associated aliases is done by mining the graph.

II. Related Work

The previously used methods used for extracting the aliases include the Japanese alias extraction method specific to japanese names proposed by Hokama and Kitagawa (HK) [4].This method used manually crafted patterns which has information about aliases. But the patterns used were itself highly ambiguous. For a particular name manually created patterns were passed as queries to the web search engine to retrieve the aliases. For example “*koto name” is a pattern which was passed as a query to the search engine for a name to retrieve the alias names. The word “koto” is a Japanese word which means also known as in English but it has many meanings like task, incident, instrument, thing etc. This will create many unrelated aliases for a given name. And also manually creating patterns to pass it as queries to the search engine is a time consuming process.

The next method which was proposed was by Bollegala, Matsuo and Ishizuka [5] which extracted aliases using the 1) lexical pattern approach from snippets retrieved for a query from the search engine and also the 2) anchor texts and URL's containing the aliases for a given name. This approach applied ranking to the extracting aliases to rank them in an order using the frequency of lexical patterns and the co-occurrences in anchor texts, and the page count based measures to eliminate irrelevant and incorrect aliases. This method considered aliases of first order co-occurrences but did not consider the higher order co-occurrences which contain information about aliases. Adding this feature will be helpful in obtaining a high increase in the MRR and recall of the search process.

III. Proposed Method

The outline of the proposed method which is shown in Fig 1 consists of retrieving anchor texts from the web search engine using a web crawler for a given name. Next is the calculation of co-occurrences measures and ranking the anchor texts for a given name and alias, construction of anchor text co-occurrence graph and calculating the co-occurrence orders. To calculate the co-occurrence measures between anchor texts for a name and alias we use nine co-occurrence statistics [5] of the previous method from which we consider the aliases of not only the first order but also of higher order co-occurrence which occur with the given name. They are Co-occurrence Frequency(CF) measure, Term frequency-Inverse document frequency (tf-idf) measure, Chi-Square (CS) measure, Log Likelihood Ratio (LLR),Point wise Mutual Information (PMI),Hyper Geometric distributions (HG), Cosine measure, Overlap measure, and Dice coefficient. Training data is given to the rank support vector machine to rank the anchor texts and identify their ranking to make co-occurrence orders among the anchor texts for a name and alias.

A. *Retrieval of Anchor Texts*

The proposed method will retrieve the anchor texts and URLs for a given name in which the name and alias appear from the web search engine using a web crawler [6]. We have many web crawlers available which extracts all the inbound anchor texts [5] of an URL. This is done using a hash table [6] which consists of a key [6] and a value associated to it. The key consists of an URL and its respective value consists of a set of anchor texts which point to the respective URL key. From the crawled data a set of links are extracted which consists of anchor texts and the URLs pointed by the anchor texts.

Next cross tabulation will be created as shown in Table 1 for a given name and alias anchor text pair to calculate the co-occurrence statistics. p and x are the name and alias anchor text pairs. C is the set of anchor texts excluding p, V is the set of all words that appear in anchor texts, C-{x} and V-{p} are all the anchor texts excluding x and p respectively, k is the co-occurrence frequency between p and x, co-occurrence frequencies sum between p and all anchor texts in C is n. Co-occurrence frequencies sum between all words in V and x is K, Co-occurrence frequencies sum between all words in V and all anchor texts in C is N.

TABLE I
Cross Tabulation For Anchor Texts Containing Name 'p' and alias 'x'

Anchor Texts	**x**	**C – {x}**	**C**
p	k	n – k	n
V – {p}	K - k	N – n – K + k	N – n
V	K	N – K	N

B. *Anchor Texts and Co-occurrences*

Anchor texts are the clickable text of a link which links to another page or resource in a web page.The words in the anchor text gives a description of the target page which helps the user whether to move to another page or not. It acts as a great tool in search engines for optimization as the text used in anchor text describes the relevant content of the target page. This will be helpful and used in search engine algorithms as they look for relevant results for a query in search. Anchor texts are used in name entity translation, web search, ranking of web pages, web query translation [7], disambiguation [2] and synonym extraction.

Co-occurrences means the occurrence of words frequently alongside each other in a text corpus in a particular order. All the words which co-occur with a given word is called as co-occurrence. The words which occur together with the co-occurring words of a given word is called as second order co-occurrence. This can be extended to next higher order co-occurrences (third, fourth or nth order). Co-occurrence information in text corpus is used to improve the performance in information retrieval [8], text data mining. These co-occurrences are used in association or similarity measures between two words and the association or similarity decreases with the occurrence of first word without second word or second word without first word, and slightly increases with high-order co-occurrences. The higher order co-occurrences provide more specific and concrete information than the first order co-occurrences.

Anchor texts of two different web pages pointing to the same URL is called are called inbound anchor texts [5]. The name p and the alias x are co-occurring if they appear in two different inbound anchor texts of a URL and their frequency as the number of different URL's in which they co-occur. For example in the Fig 2, the web page of Sharukh Khan is shown which is linked by four different anchor texts. Here the anchor texts "*Sharukh Khan*" and "*Srk*" are co-occurring and also "*Bollywood Badhshah*" and "*Don*" are also co-occurring.

C. *Co-occurrence Statistics*

1) *CF*

Co-occurrence frequency [5] is the measurement value among a name p and an alias x denoted by the value k in Table II.

2) *tf-idf*

The co-occurrence frequency is biased towards highly frequent words. Stop words present in the anchor texts have high frequency. The tf-idf measure [5] [9] resolves this by normalizing the bias towards high frequency words by reducing the weight assigned to the words of anchor texts. From Table III the tf-idf (p, x) [5] for the anchor texts containing p and x is calculated as

$$\mathrm{tfidf}(p, x) = k \log\left(\frac{N}{K+1}\right)$$

3) CS

The Chi Square measure [5] is a test of dependence between two words used in many tasks such as collocation detection, word similarity measures and identification of translation pairs in aligned corpora. The chi-square measure compares the observed frequencies in Table IV with the expected frequencies for test of independence.

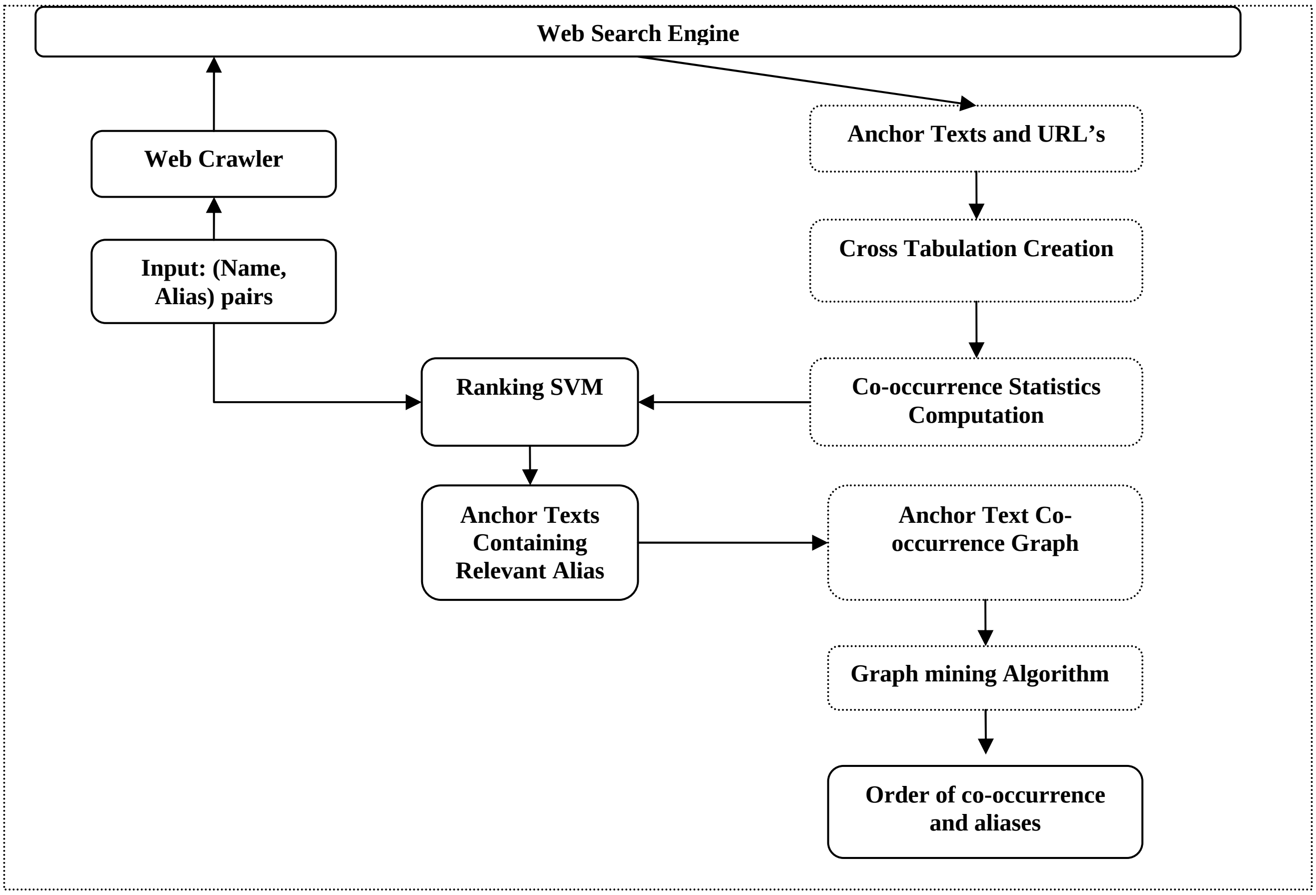
Web Search Engine
Web Crawler
Anchor Texts and URL's
Cross Tabulation Creation
Input: (Name, Alias) pairs
Ranking SVM
Co-occurrence Statistics Computation
Anchor Texts Containing Relevant Alias
Anchor Text Co-occurrence Graph
Graph mining Algorithm
Order of co-occurrence and aliases

Fig 1 Outline of the Proposed Method

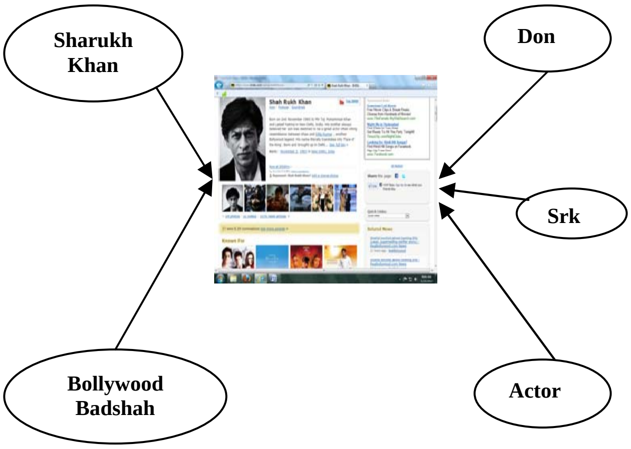

Fig.2 A web page of Sharukh Khan being linked by different anchor texts on the web

The anchor texts containing name p and alias x are dependent there is a large difference between the observed and expected frequencies. The *chi square* X^2[5] calculated from Table V sums the difference between the observed and expected frequencies by order of magnitude of expected values given as

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Here O_{ij} is the observed and E_{ij} is the expected frequency. From equation (2) the chi square CS (p, x) measure [5] for anchor texts containing p and x from the Table VI is calculated as follows

$$CS(p,x) = \frac{N\{k(N \quad K \quad n + k) \quad (n \quad k)(K \quad k)\}^2}{nK(N-K)(N-n)}$$

4) LLR

Log likelihood ratio [5] [10] is calculated between two alternative hypotheses: that the name p and the alias x are independent or they are dependent of each other. LLR (p, x) [5] is calculated from Table VII as follows

$$LLR(p,x) = k\log\frac{kN}{nK} + (n-k)\log\frac{N(n-k)}{n(N-K)} +$$
$$(K-k)\log\frac{N(K-k)}{K(N-n)} +$$
$$(N-K-n+k)\log\frac{N(N-K-n+k)}{(N-k)(N-n)}$$

5) PMI

Point wise mutual information [5] [11] is a measure of association which reflects the dependence between two probabilistic events. The PMI [5] is defined for random variables y and z events as

$$PMI(y,z) = \log_2\left(\frac{P(y,z)}{P(y)P(z)}\right)$$

Where P(y) and P(z), respectively, represent the probability of events y and z. Whereas P(y, z) is the joint probability of y and z. The PMI(y, z) [5] is calculated from Table VIII as

$$PMI(y,z) = \log_2\left(\frac{kN}{Kn}\right)$$

6) HG

Hyper Geometric distribution [5] [12] is a discrete probability distribution which gives the number of successes in a sequence of draws from a finite population without any replacement. For example, the probability of the event that "k red balls are contained among n balls, which are arbitrarily selected from among N balls containing K red balls" is calculated using hyper geometric distribution *hg*(N, K, n, k)[5] as

$$hg(N,K,n,k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

The values in Table 1 is used in definition (7) and the probability of observing more than k number of co-occurrences of p and x is given by HG (p, x) [5] as

$$HG(p,x) = -\log_2\left(\sum_{l\geq k} hg(N,K,n,l)\right)$$

$$\max\{0, N+K-n\} \geq l \geq \min\{n,K\}$$

7) Cosine

Cosine [5] computes the association between words in anchor texts. The association between elements in two sets X and Y is cosine (p, x) [5] and is computed as

$$cosine(p,x) = \frac{|X \cap Y|}{\sqrt{|X|}\sqrt{|Y|}}$$

Where |X| and |Y| represent the number of elements in set X and Y respectively.The cosine (p, x) measure calculated from Table 1 taking X as the co-occurrences of anchor text containing alias x and Y as the co-occurrences of anchor texts containing name p is given as

$$cosine(p,x) = \frac{k}{\sqrt{nK}}$$

8) Overlap

The overlap [5] between two sets X and Y is defined as

$$overlap(X,Y) = \frac{|X \cap Y|}{\min(|X|,|Y|)}$$

X and Y represent occurrences of name p and alias x in anchor texts. The appropriateness of an alias x is given by overlap of (p, x) [5] as

$$overlap(p,x) = \frac{k}{\min(n,K)}$$

9) Dice

Dice [5] [13] is used to retrieve collocations from large textual corpora. The Dice(x, y) [5] is defined over two sets X and Y as

$$Dice(x,y) = \frac{2|X \cap Y|}{|X|+|Y|}$$

The co-occurrence values from Table IX are used for defining a rank score based on Dice (p, x) [5] as

$$Dice(p,x) = \frac{2k}{n+K}$$

D. *Training Data and Rank Support Vector Machines*

The training data sets of name-alias pairs are given to the rank support vector machines [5] [14] for ranking the aliases. The data sets that are considered for training the rank support vector machine (SVM) are the person names data set and the location names data set. The person names data set will include people from different fields such as film stars, politicians, writers, musicians, sports person. The location names contain different locations names and its aliases. The values which are calculated from the co-occurrence measures [5] are normalized to a range suitable for considering the first and higher order co-occurrences. These measures along with the data sets are used for training the rank support vector machine to rank the anchor texts having a name and aliases. The SVM [5] will rank each alias contained in the anchor text with its corresponding anchor text. The anchor text which gets a highest ranking score will be chosen with its corresponding anchor text for which ranking was performed to make first order co-occurrence. The number of irrelevant pairs will be ignored in the training data by the SVM. The co-occurrence graph will be constructed as per the ranking given for the measure of association between a name and aliases.

E. *Creation of Anchor Text Co-occurrence Graph*

A co-occurrence graph is an undirected graph which models the co-occurrences for words that appear in the anchor texts for a name and alias. The graph is created with nodes for each word in the anchor text and as per the definition two words are said to be co-occurring if two different anchor texts containing these words link to the same URL. If the words representing the nodes co-occur then an edge is formed between them. A co-occurrence graph is created for a name and alias as per the co-occurrence orders between them and the same is shown for *Sharukh khan* in the Fig 3.The nodes represent the name and aliases and the edges between them gives us the order of co-occurrences. The hop distance gives us the order of occurrences for a name and aliases referred by nodes by using graph mining algorithm.

F. *Measuring Hop Distances*

The co-occurrence graph which is constructed with the nodes represented by name and aliases and the edges representing the association is then mined to calculate the hop distances between the nodes using graph mining algorithm [15] [16].The count related to the number of edges between the nodes gives us the hop distance which will give us the order of association. The nodes which lies n hops away from p has an n order co-occurrence with p and these co-occurrence orders can be known by measuring the hop distances between a node containing the name with its corresponding nodes containing aliases.

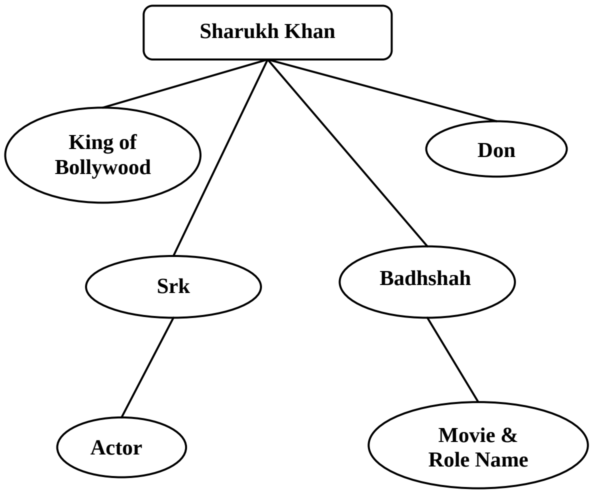

Fig 3 Anchor text co-occurrence graph for a personal name *"Sharukh Khan"*

IV. CONCLUSION

The proposed method retrieves anchor texts and URL's for a name and uses co-occurrence statistics to measure the association of anchor text for a name and its aliases. The co-occurrence graph is constructed consisting of nodes representing names and its aliases. The order of co-occurrence is measured taking the hop distances into account using the graph mining algorithms. The order of co-occurrences including the first and higher order will give us the relevant aliases for a name and also gives us semantic information about a given name. These aliases will improve the recall of a web search task and is helpful in obtaining a high MRR compared to the previous methods.

REFERENCES

[1] J. Artiles,J. Gonzalo, and F. Verdejo "A Testbed for People Searching Strategies in the WWW," Proc. SIGIR '05, pp. 569-570, 2005.

[2] R.Guha and A.Garg, "Disambiguating People in Search," technical report, Stanford Univ., 2004.

[3] P. Cimano, S. Handschuh, and S. Staab, "Towards the Self-Annotating Web," Proc. Int'l World Wide Web Conf. (WWW '04), 2004.

[4] [5] T. Hokama and H. Kitagawa, "Extracting Mnemonic Names of People from the Web," Proc. Ninth Int'l Conf. Asian Digital Libraries (ICADL '06), pp. 121-130, 2006.

[5] D.Bollegala, Y. Matsuo, and M. Ishizuka, "Automatic Discovery of Personal Name Aliases from the Web," IEEE Transactions on Knowledge and Data Engineering, vol. 23, No. 6, June 2011.

[6] Gautam, Pant,Padmini Srinivasan,and Filippo Menczer "crawling the web" Web dynamics: adapting to change in content, size, topology and use - Page 153

[7] W. Lu, L. Chien and H. Lee, "Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach," ACM Transactions on Information Systems, Vol. 22, No. 2, April 2004, Pages 242-269.

[8] G. Salton and M. McGill, Introduction to Modern Information Retrieval. McGraw-Hill Inc., 1986.

[9] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information processing and Management, vol. 24, pp. 513-523, 1988.

[10] T. Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," Computational Linguistics, vol. 19, pp. 61-74, 1993.

[11] K. Church and P. Hanks, "Word Association Norms, Mutual Information and Lexicography," Computational Linguistics, Vol. 16, pp. 22-29, 1991.

[12] T. Hisamitsu and Y. Niwa,"Topic-Word Selection Based on Combinatorial Probability," Proc. Natural Language Processing Pacific-Rim Symp. (NLPRS '01), pp.289-296, 2001.

[13] F.Smadja, "Retrieving Collocations from Text: Xtract," Computational Linguistics, Vol. 19, no 1, pp. 143-177, 1993.

[14] T. Joachims," Optimizing Search Engines using Click through Data," proc. ACM SIGKDD '02, 2002.

[15] D. Chakrabarti and C. Faloutsos, "Graph Mining: Laws, Generators, and Algorithms," ACM Computing Surveys, Vol. 38, March 2006, Article 2.

[16] C.C. Agarwal and H. Wang,"Graph Data Management and Mining : A Survey of Algorithms and Applications," DOI 10.1007/978-1-4419-6045-0_2,@Springler Science+Business Media, LLC 2010

AUTHORS PROFILE

Rama.K has received B.Tech degree in Computer Science and Information Technology branch. She is now pursuing M.Tech (Computer Science and Engineering) from CVSR college of Engineering, Anurag Group of institutions.

Mrs. Sridevi.M has received B.Tech and M.Tech degrees in Computer Science and Engineering branch. She is working as an Associate Professor, Department of Computer Science and Engineering in CVSR college of Engineering, Anurag Group of Institutions. She has 8 years of teaching experience.

Mr. Vishnu Murthy G received his B.E and M.Tech degrees in Computer Science and Engineering. He is having 15 years of teaching experience. He is presently pursuing his Ph.D. in JNTU, Hyderabad and is the Head of the Computer Science and Engineering Department, CVSR college of Engineering, Anurag Group of Institutions. He has organized and attended various workshops and conferences at National and International level. He has been the resource person for Institute of Electronic Governance and BITS off campus programs. He is the Life Member of ISTE, IEEE, ACM, CRSI & CSI. He had 5 publications in international journals and presented 2 papers in conferences. His areas of interest include Software Engineering, Information Security and Image Processing.