SENSE DISAMBIGUATION TECHNIQUE FOR PROVIDING MORE ACCURATE RESULTS IN WEB SEARCH

Rekha Jain[1] and G. N. Purohit[2]

[1]Department of Computer Science, Banasthali University, Rajasthan, India
rekha_leo2003@rediffmail.com
[2]Department of Computer Science, Banasthali University, Rajasthan, India
gn_purohitjaipur@yahoo.co.in

ABSTRACT

As the web is increasing exponentially, so it is very much difficult to provide relevant information to the information seekers. While searching some information on the web, users can easily fade out in rich hypertext. The existing techniques provide the results that are not up to the mark. This paper focuses on the technique that helps in offering more accurate results, especially in case of Homographs. Homograph is a word that shares the same written form but has different meanings. The technique that shows how senses of words can play an important role in offering accurate search results, is described in following sections. While adopting this technique user can receive only relevant pages on the top of the search result.

KEYWORDS

Information Retrieval, Sense Disambiguation Technique, Homographs

1. INTRODUCTION

Sometimes a single word can have different senses. These words are called as polysemous words e.g. bass can be a type of fish or it can be a musical instrument. Word Sense Disambiguation is a process that selects a sense from a set of predefined word senses to an instance of a polysemous word in a particular context and assigns that sense to the word. This technique considers following two properties of a word i.e. polysemy and homonymy. Polysemy and Homonymy are two well known semantic problems. *Bank* in *river bank* and *Bank of England* are homonymous. *River bed* and *hospital bed* describe the case of polysemy property. Word Sense Disambiguation technique is useful to find semantic understanding of the text. It is an important as well as challenging technique in the area of NLP (Natural Language Processing), MT (Machine Translation), Semantic Mapping, IR (Information Retrieval), IE (Information Extraction), Speech Recognition etc.

One of the problems with Information Retrieval (IR), in case of Homographs, is to decide the correct sense of the word because dictionary based word senses definitions are ambiguous. If trained linguists manually tag the word sense then there are the chances that different annotations may assign different senses to same word, so some technique is required to disambiguate a word. Word knowledge is difficult to verbalize in dictionaries [1].

DOI : 10.5121/ijwsc.2012.3303

To disambiguate a polysemous word, two resources are necessary- 1) the context to which the word is linked and 2) some kind of knowledge related to that word. There are four parts-of-speech that need disambiguation- nouns, verbs, adjectives and adverbs. This paper focuses on the technique that will resolve the ambiguity between noun polysemous words.

The remainder of paper is organized as follows- in section 2 we discuss various approaches for resolving the sense of the word. In section 3 some knowledge resources are introduced. Section 4 discusses the applicability of Sense Disambiguation Technique, section 5 gives the brief overview of problem and our proposed approach is discussed in section 6. Section 7 provides the results of our developed algorithm and at last section 8 analyses the result. Finally conclusion and future work finishes the article.

2. APPROACHES

Word Sense Disambiguation algorithms can be roughly classified into Unsupervised Approach and Supervised Approach on the basis of training corpora.

2.1. Unsupervised Approach

In this approach training corpus is not required. This approach needs less time and power. Major use of this approach is in MT (Machine Translation) and IR (Information Retrieval), but this approach has worst performance as compare to supervised approach because less knowledge is required in this approach. It has various following sub approaches-

A. Simple Approach (SA): It refers to the algorithms that consider only one type of lexical knowledge. This approach is easy to implement but it do not have good precision and recall. Precision is the portion of correctly classified samples among classified samples. Recall is the portion of correctly classified samples among total samples [2, 3]. Generally the value of recall is less than the value of precision unless all the samples are tagged.

B. Combination of Simple Approaches (CSA): It is a combination of simple approaches that are created by simply summing up the normalized weights of individual simple approaches [4]. As multiple resources offer more confidence on a sense than a single resource does, so it usually performs better than a single approach.

C. Iterative Approach (IA): This approach only tags the words that have high confidence on the basis of information for sense tagged words from previous step and other lexical knowledge [5]. This approach disambiguates the nouns with 55% precision and verbs with 92.2 % precision.

D. Recursive Filtering (RF): This approach follows the same principle as IA but with some differences like it assumes that correct sense of a target word should have stronger semantic relationship with other words than the remaining senses. This approach does not disambiguate the sense of all words until final step. This algorithm gradually reduces the irrelevant senses and leaves only relevant ones within a finite number of cycles. It had been reported that this algorithm had 68.79% precision and 68.80 % recall [6].

E. Bootstrapping (BS): This approach follows a recursive optimization algorithm which requires few seed values instead of having a large number of training samples. This approach recursively processes the trained model to predict the sense of new cases and returns a model of new predicted cases. A list of 12 words is applied on this algorithm and 96.5% precision is achieved [7]. This approach truly achieves very high precision but it is limited to disambiguate a few words from the text.

2.2. Supervised Approach

This approach uses the train model of sense tagged corpora that links world knowledge to word sense. Most recently developed WSD algorithms are supervised because of availability of training corpora, but it does not mean that unsupervised approach is out of mode. It has the following sub approaches-

A. Log Linear Model (LLM): It is based on the assumption that each feature is conditionally independent of others. The probability of each sense is computed with Bayes' Rule [8]

$$p(s_i \mid c_1,\ldots,c_k) = \frac{p(c_1,\ldots,c_k \mid s_i)\, p(s_i)}{p(c_1,\ldots,c_k)} \tag{1}$$

Because $p(c_1,\ldots,c_k)$ is same for all senses of target word, we can simply ignore it. According to independence assumption:

$$p(c_1,\ldots,c_k \mid s_i) = \prod_{j=1}^{k} p(c_j / s_i) \tag{2}$$

$$s = \underset{s_i}{ARGMAX} \log p(s_i) + \prod_{j=1}^{k} \log p(c_j / s_i) \tag{3}$$

But this approach has two disadvantages 1) The concept of assumption independence is not clear 2) It needs some good techniques to smooth the terms [9].

B. Decomposable Probabilistic Models (DPM): This model fixes the false assumption of LLM's by setting the interdependence features of training data [10, 11]. This approach could achieve better results if the size of training data is large enough to compute the interdependence settings.

C. Memory Based Learning (MBL): This approach supports both numeric features as well as symbolic features so it can be used to integrate various features into one model [12]. This approach classifies the new cases by calculating the similarity matrix as follows-

$$\Delta(X,Y) = \prod_{i=1}^{n} w_i \delta(x_i, y_i) \tag{4}$$

Where

$$\delta(x_i, y_i) = \frac{x_i - y_i}{\max_i - \min_i} \quad \text{if numeric, else}$$

$$\delta(x_i, y_i) = 1 \text{ if } x_i \neq y_i$$

$$\delta(x_i, y_i) = 0 \text{ if } x_i = y_i$$

If there is no information about feature relevance the feature weight is 1, otherwise domain knowledge bias is added to weight.

D. Maximum Entropy (ME): It is constraint based approach where the algorithm maximizes the entropy of $p_\lambda(y \mid x)$. This is the conditional probability of sense Y under facts X, given a collection of facts computed from data [13, 14].

$f_i(x, y) = 1$ if sense y is under condition x, otherwise

$f_i(x, y) = 0$

$$p_\lambda(y \mid x) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_{i=1} \lambda_i f_i(x, y) \right) \quad (5)$$

Parameter λ can be computed by numeric algorithm called as Improve Iterative Scaling algorithm.

E. Expectation Maximum (EM): This approach solves the maximization problem that contains incomplete information by applying an iterative approach. Incomplete information means the contextual features are not directly associated with word senses. Expectation Maximum is a climbing algorithm where its achievement of global maximum depends on initial values of parameters [15]. We should be careful to initialize the parameters. This Expectation Maximum does not require the corpus to be sense tagged as it can learn conditional probability between hidden sense and aligned word pairs from bilingual corpora.

Table 1. Summarization of all WSD algorithms

Group	Tasks	Knowledge Sources	Computing Complexity	Performance	Other Characteristics
SA	all-word	single lexical source	low	low	
CSA	all-word	multiple lexical sources	low	better than SA	
IA	all-word	multiple lexical sources	low	high precision average recall	
RF	all-word	single lexical source	average	average	flexible semantic relation
BS	some-word	sense-tagged seeds	average	high precision	sense model converges
LLM	some-word	contextual sources	average	above average	independence assumption
DPM	some-word	contextual sources	very high	above average	need sufficient training data
MBL	all-word	lexical and contextual sources	high	high	
ME	some-word	lexical and contextual sources	very high	above average	feature selection
EM	all-word	bilingual texts	very high	above average	local maximization problem

Table-1 gives a brief summarization of all the Word Sense Disambiguation algorithms discussed above [16]. Computing complexity is one of the major issues that must be considered whenever there is a choice of Word Sense Disambiguation algorithm.

3. KNOWLEDGE RESOURCES

There are two categories of knowledge Resources 1) Lexical Knowledge that is released for public use and 2) World Knowledge that is learned from training corpora [16].

3.1 Lexical Knowledge

It is the base for unsupervised WSD approaches. It has the following components-

i) Sense Frequency is the occurrence or frequency of each sense of word.
ii) Sense Gloss provides the sense of a word by definitions and examples. The word sense can be tagged by counting common words between the gloss and context of the word.
iii) Concept Trees describes the relationships between synonym, hypernym, homonym etc. A WSD algorithm can be derived from this hierarchical concept tree.
iv) Selection Restrictions are semantic restrictions that can be placed on word sense. LDOCE (Longman Dictionary Of Contemporary English) provides this kind of information.
iv) Subject Code refers to the category the sense of target word belongs to. Some weighted indicative words are also used with subject code. These indicative words are fetched from training corpus.

3.2 Learned World Knowledge

It is very much difficult to verbalize the World Knowledge. So some technique is required that can automatically fetch world knowledge from contextual knowledge by machine learning techniques. Components of Learned Knowledge are as follows-

i) Indicative Words are the words that surround the target word and help to sense the target word. The word that is more close to the target word is more indicative word to sense.
ii) Syntactic features refer to sentence structure. They check position of the specific word. It may be subject, direct object, indirect object etc [13].
iii) Domain Specific Knowledge is about some semantic restrictions that can be applied on each sense of the target word. This knowledge can only be retrieved from a training corpora and it can be attached to WSD algorithm for better learning of world knowledge [17].
iv) Parallel Corpora is based on the concept of translation process. This process implies that major words like nouns, verbs etc. share the same sense or concept in different languages. These types of corpora contain two languages one is primary language and other one is secondary language. The major words of language are aligned using third party software [18].

4. APPLICABILITY OF WSD

Word Sense Disambiguation does not play a direct role in human language technology instead it gives its participation into other applications like Information Retrieval (IR), Machine Translation (MT), Word Processing etc. Another field, where WSD plays a major role is Semantic Web [16]. Here WSD participates in Ontology Learning, Building Taxonomies etc. The Information Retrieval (IR) is open research area that needs to distinguish the senses of word that are searched by the user and returns only pages that contain needed senses.

5. STATEMENT OF PROBLEM

To disambiguate a word two issues must be considered 1) context in which the word has been used and 2) some kind of world knowledge. A human being contains the world knowledge that helps to disambiguate the words easily. For example the word “bass” appears in a text, it needs to be disambiguated because of its multiple senses. It may refer to the musical instrument “bass” or it may also refer to the kind of fish “bass”. Since computers do not have world knowledge used by human beings to disambiguate a word, they need some other resources for fulfilling this task. Some technique is required that can resolve the ambiguity between these polysemous words. Precision and recall are two important factors for measuring the performance of WSD. Precision is the proportion of correctly classified instances of those classified. Recall is proportion of correctly classified instances of total instances. In general the recall value is less than precision value. WSD is applied whenever a semantic understanding of text is needed.

6. OUR APPROACH

There are four parts-of-speech that allow polysemy: nouns, verbs, adverbs and adjectives. Our approach is based on supervised technique that is used to disambiguate noun polysemous words. To disambiguate the sense of a word we need sense knowledge and contextual knowledge. Sense knowledge comprises of lexical knowledge and world knowledge. There is no separation line between lexical knowledge and world knowledge, usually unsupervised approaches use lexical knowledge and supervised approached use learned world knowledge. Our approach is based on supervised approach that uses domain specific knowledge to resolve the ambiguities between polysemous words. Contextual knowledge contains word to be sensed and its features.

The proposed algorithm disambiguates the word sense of polysemous words when the user performs search on Web. The approach is based on domain specific knowledge. This knowledge can be attached with WSD algorithm by empirical methods. Proposed algorithm has two subsections. In the first part we have applied pre-processing before sending the query to Search Engine. In the second part or next module we would apply some mechanism that would rearrange the pages retrieved from Search Engine according to user’s needs. This module would first rearrange the pages according to users’ needs then on the basis of their ranks. Mostly the users explore top 6-7 pages that are included in their search result. This module would provide the relevant pages on the top of search result.

6.1 Algorithm

1. Receive the string entered by user to search
2. Divide the string in tokens
3. for each token
4. search its root word from dictionary
5. check the root word in the list of polysemous words
6. if found
7. retrieve the world knowledge of specific token from dictionary
8. retrieve the contextual information from the domain specified
9. create the sense disambiguation knowledge from world knowledge and contextual information of token
10. attach the sense of word with string
11. otherwise
12. retain the token as it is
13. if more tokens available
14. go to step 4
15. pass the resultant string to Search Engine

6.2 System Architecture

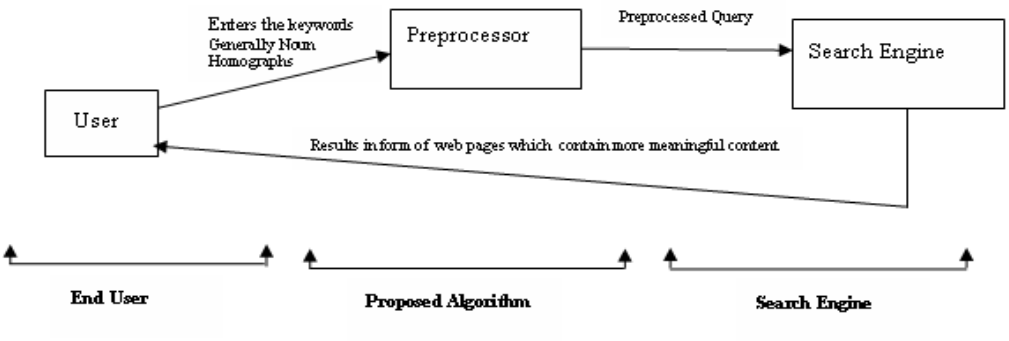

Figure 1. System Architecture

This algorithm shows the result in form of URLs which are ranked according to the user's domain and their importance.

6.3 Methodology

Two users were considered in this experiment. Each user was asked to specify his/her domain of interest. It had been reported that generally the users were interested to explore only 6-7 pages of search result, so the query result should be relevant according to users' interest. First user was an Ichthyologist whose domain was to study the fishes, and second user was a Musician. This user was interested in searching the information about various musical instruments.

7. EXPERIMENTAL EVALUATION

The disambiguation algorithm remembers the primary domain of interest and retrieves more meaningful contents to the users.

An Ichthyologist searched the word bass via Google Search Engine and entered the word bass on search engine interface as shown in Figure 2.

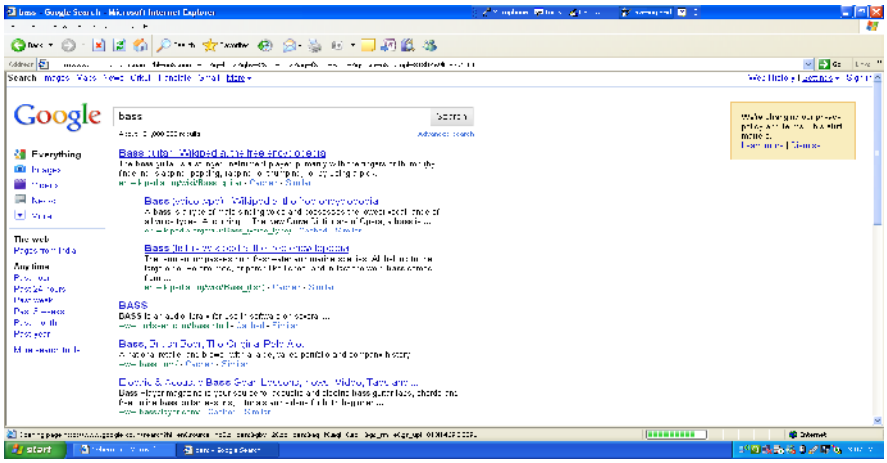

Figure 2. Results retrieved by Google Search Engine Directly

The results received were not up to the mark because he/she was expected the details about the fish “bass” not about a musical instrument or anything else.

The proposed algorithm resolved the ambiguities between Noun Homographs. At the time of searching users never bothered about the multiple meanings of the word; their only requirement is that their relevant content must appear at top of result.

But when the same user (Ichthyologist) performed the same search through our developed module, the result varies. Those results were more relevant as compared to earlier results as shown in Figure 2, because the pages appear at the top of result provided the details regarding the bass fish.

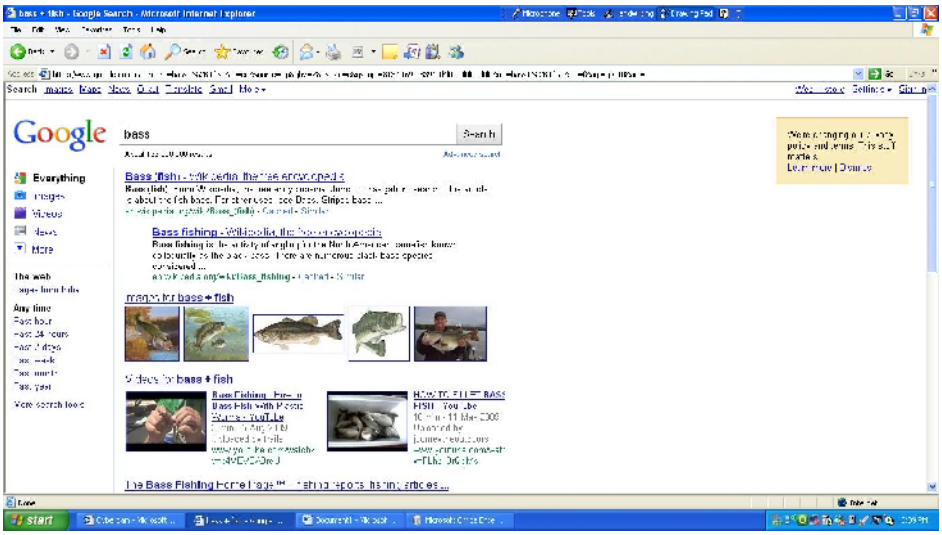

Figure 3. Results retrieved by new Algorithm-1

If the user is a musician then it is obvious that he/she is interested in searching the details for bass, a musical instrument Figure 4 shows the results in following manner such that if a musician searched the details for word Bass. Here the top of result provided the details for the Bass, a musical instrument.

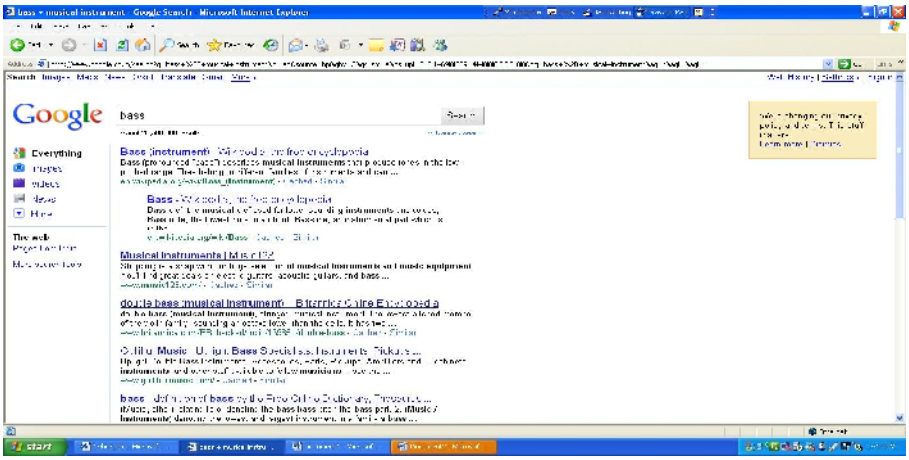

Figure 4. Results retrieved by new Algorithm-2

8. ANALYSIS OF RESULT

Figure 2 shows the result when the user directly enters bass keyword on to Google interface. Here Google searches all the possible pages having word bass in them and then arranges them in the descending order of their page ranks. It includes pages from all the possible domains. In new developed algorithm user never enters search keywords on to Google interface instead he/she performs the search via our algorithm's search interface. The algorithm provides the result in different manner as it can be seen in Figure 3 and Figure 4 that both the users (Ichthyologist and Musician) enter the same word to search and disambiguation algorithm performs some preprocessing and then passes the resultant query to Search Engine and as a result the Ichthyologist and Musician receive respective web pages.

9. CONCLUSION AND FUTURE WORK

As specified earlier we have developed an algorithm for pre processing of query that we want to send to Search Engine to retrieve some relevant contents from WWW. The future work related to this area will revolve around second part of the research. Here our proposed algorithm would rearrange the pages so that user can get more meaningful contents at the top. This rearrangement of pages would be based on some mathematical formula which takes the value of PageRank as one of the parameter.

REFERENCES

[1] Veronis, J.,Sense Tagging: Don't Look for the Meaning But for the Use, Workshop on Computational Lexicography and Multimedia Dictionaries, Patras, Greece, pp. 1-9 (2000)

[2] Lesk, M. Automatic Sense Disambiguation: How to Tell a Pine Cone from and Ice Cream Cone. Proceedings of the SIGDOC'86 Conference, ACM (1986)

[3] Galley, M., & McKeown, K., Improving Word Sense Disambiguation in Lexical Chaining, International Joint Conferences on Artificial Intelligence (2003)

[4] Agirre, E. et al., Combining supervised and unsupervised lexical knowledge methods for word sense disambiguation. Computer and the Humanities, Vol.34, P103-108 (2000)

[5] Mihalcea, R. & Moldovan, D., An Iterative Approach to Word Sense Disambiguation. Proceedings of Flairs, Orlando, FL, pp. 219-223 (2000)

[6] Kwong, O.Y., Word Sense Selection in Texts:An Integrated Model, Doctoral Dissertation, University of Cambridge (2000)

[7] Yarowsky, D., Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. Meeting of the Association for Computational Linguistics, pp. 189-196 (1995)

[8] Yarowsky, D., Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. Proceedings of COLING-92, Nantes, France, July 1992, pp. 454-460 (1992)

[9] Chodorow, M., Leacock, C., and Miller G., 2000. A Topical/Local Classifier for Word Sense Identification Computers and the Humanities Vol. 34, pp.115-120 (2000)

[10] Bruce, R. & Wiebe, J., Decomposable modeling in natural language processing. Computational Linguistics, Vol. 25, No 2 (1999)

[11] O'Hara, T, Wiebe, J., & Bruce, R., Selecting Decomposable Models for Word Sense disambiguation: The Grling-Sdm System. Computers and the Humanities, Vol. 34, pp. 159-164 (2000)

[12] Daelemans, W. et al., 1999. TiMBL: Tilburg Memory Based Learner V2.0 Reference Guide, ILK Technical Report- ILK 99-01 (1999)

[13] Fellbaum, C. & Palmer, M., Manual and Automatic Semantic Annotation with WordNet. Proceedings of NAACL Workshop (2001)

[14] Berger, A. et al., A maximum entropy approach to natural language processing. Computational Linguistics, Vol. 22, No 1 (1996)

[15] Dempster A. et al., Maximum Likelihood from Incomplete Data via the EM Algorithm. J Royal Statist Soc Series B Vol. 39, pp. 1-38 (1977)

[16] Xiaohua Zhou, Hyoil Han, Survey of Word Sense Disambiguation Approaches. 18th FLAIRS Conference, Clearwater Beach, Florida (2005)
[17] Hastings, P. et al., Inferring the meaning of verbs from context Proceedings of the Twentieth Annual Conference of the Cognitive Science Society (CogSci-98), Wisconsin, Madison (1998)
[18] Bhattacharya, I., Getoor, L., and Bengio, Y., Unsupervised sense disambiguation using bilingual probabilistic models. Proceedings of the Annual Meeting of ACL (2004)

Authors

Rekha Jain completed her Master Degree in Computer Science from Kurukshetra University in 2004. Now she is working as Assistant Professor in Department of "Apaji Institute of Mathematics & Applied Computer Technology" at Banasthali University, Rajasthan and pursuing Ph.D. under the supervision of Prof. G. N. Purohit. Her current research interest includes Web Mining, Semantic Web and Data Mining. She has various National and International publications and conferences.

Prof. G. N. Purohit is a Professor in Department of Mathematics & Statistics at Banasthali University (Rajasthan). Before joining Banasthali University, he was Professor and Head of the Department of Mathematics, University of Rajasthan, Jaipur. He had been Chief-editor of a research journal and regular reviewer of many journals. His present interest is in O.R., Discrete Mathematics and Communication networks. He has published around 40 research papers in various journals.