TOPIC TRACKING FOR PUNJABI LANGUAGE

Kamaldeep Kaur[1] and Vishal Gupta[2]

[1]University Institute of Engineering & Technology, Panjab University, Chandigarh, India
kamal.gndec@gmail.com
[2]University Institute of Engineering & Technology, Panjab University, Chandigarh, India
vishal@pu.ac.in

ABSTRACT

This paper introduces Topic Tracking for Punjabi language. Text mining is a field that automatically extracts previously unknown and useful information from unstructured textual data. It has strong connections with natural language processing. NLP has produced technologies that teach computers natural language so that they may analyze, understand and even generate text. Topic tracking is one of the technologies that has been developed and can be used in the text mining process. The main purpose of topic tracking is to identify and follow events presented in multiple news sources, including newswires, radio and TV broadcasts. It collects dispersed information together and makes it easy for user to get a general understanding. Not much work has been done in Topic tracking for Indian Languages in general and Punjabi in particular. First we survey various approaches available for Topic Tracking, then represent our approach for Punjabi. The experimental results are shown.

KEYWORDS

Text mining, NLP, Topic tracking, NER, Keyword extraction

1. INTRODUCTION

Text mining is a new area of computer science which fosters strong connections with natural language processing, data mining, machine learning, information retrieval and knowledge management. It seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns. The techniques employed usually do not involve deep linguistic analysis or parsing, but rely on simple 'bag-of-words' text representations based on vector space. Several approaches exist for the identification of patterns including dimensionality reduction, automated classification and clustering [1]. The field of text mining has received a lot of attention due to the always increasing need for managing the information that resides in the vast amount of available documents [2]. The goal is to discover unknown information, something that no one yet knows.

[3] The problem introduced by text mining is obvious: natural language was developed for humans to communicate with one another and to record information, and computers are a long way from comprehending natural language. Humans have the ability to distinguish and apply linguistic patterns to text and humans can easily overcome obstacles that computers cannot easily handle.

[2]A typical text mining process can be shown as:

DOI : 10.5121/cseij.2011.1304

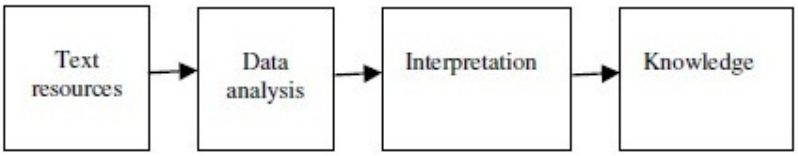

Figure 1. Typical text mining process

Taking a collection of text resources, a text mining tool would proceed with the data analysis. During this analysis many sub processes could take place such as parsing, pattern recognition, syntactic and semantic analysis, clustering, tokenization and application of various other algorithms.

2. TOPIC TRACKING

A topic tracking system works by keeping user profiles and based on the documents the user views, predicts other documents of interest to the user [4].

The task of topic tracking is to monitor a stream of news stories and find out what discuss the same topic described by a few positive samples [5]. The main purpose is to identify and follow events presented in multiple news sources, including newswires, radio and TV broadcasts [6]. With the fast development of internet, topic related information is often isolated and scattered in different time periods and places. TDT techniques are used to organize news pages from a lot of news websites into topics [7]. It collects dispersed information together and makes it easy for user to get a general understanding [8].

[4] There are many areas where topic tracking can be applied in industry. It can be used to alert companies anytime a competitor is in the news. This allows them to keep up with competitive products or changes in the market. Similarly businesses might want to track news on their own company and products. It could also be used in the medical industry by doctors and other people looking for new treatments. Individuals in the field of education could also use topic tracking to be sure they have the latest references for research in their area of interest.

A typical topic tracking system can be illustrated as: [9]

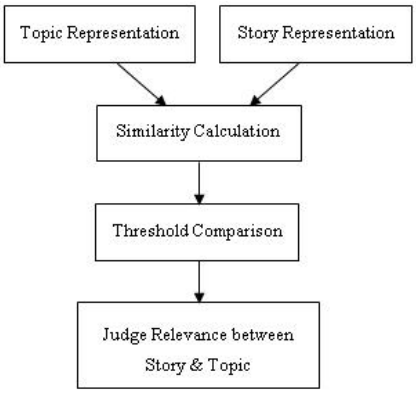

Figure 2. Architecture of a topic tracking system

When performing topic tracking, the topic tracker needs to represent topic/ story by some inner format, select similarity function for measuring topic-story similarity, and finally do threshold comparison. If the similarity is higher than predefined threshold, then the story is judged to be on-topic; otherwise, off-topic.

2.1. Background

[10][11][12]Topic Detection and tracking is fairly a new area of research in IR, developed over the past 8 years. Began during 1996 and 1997 with a Pilot study conducted to explore various approaches and establish performance baseline. The research began in 1996 with DARPA funded pilot study.
Quite soon the traditional methods were found more or less inadequate for online detection purposes. In recent years, TDT techniques have been developed to identify the issues discussed in a large collection text.
Very brief descriptions of the existing TDT tasks are given:

- TDT-1: deals with the three major tasks: (1) segmenting a stream of data, especially recognized speech, into distinct stories; (2) identifying those news stories that are the first to discuss a new even occurring in the news; (3) given a small sample news stories about the event. The TDT corpus includes approximately 16,000 stories about half collected from Reuters newswire and half from CNN broadcast news transcripts during the period July 1, 1994 to June 30, 1995. An integral and key part of the corpus is the annotation in terms of news events discussed in the stories. Twenty-five events were defined that span a variety of event types and that cover a subset of the events discussed in the corpus stories. Annotation data for these events are included in the corpus and provide a basis for training TDT systems.
- TDT-2: they ran tracking experiments on just the Mandarin stories in the development corpus. The TDT2 English Corpus has been designed to include six months of material drawn on a daily basis from six English news sources. The period of time covered is from January 4 to June 30, 1998. The six sources are the New York Times News Service, the Associated Press World stream News Service, CNN "Headline News", ABC "World News Tonight", Public Radio International's "The World", and the Voice of America.
- TDT-3: the tracking part of the corpus consists of 71,388 news stories from multiple sources from English and Mandarin (AP, NYT, CNN, ABC, NBC, MSNBC, Xinhua, Zaobao, Voice of America and PRI the World) in the period of October to December 1998. Machine-translated versions of the non-English stories (Xinhua, Zaobao and VOA Mandarin) are provided as well.
- TDT-4: In TDT-4, LDC defined 60 topics based upon a stratified, random sample of the eight English and seven Chinese news sources collected from October 2000 through January 2001. The seed stories that generated the final 60 topics are equally divided between English and Mandarin.
- TDT-5: Corpus was newly collected from English, Chinese and Arabic from April-September 2003. Unlike previous TDT corpora, TDT-5 does not contain any broadcast news data; all sources are newswire.

  To solve the TDT challenges, researchers are looking for robust, accurate, fully automatic algorithms that are source, medium, domain, and language independent.

2.2. Approaches

Many techniques have been proposed for topic/ event tracking by researchers. The main focus in this paper is on the latest research being done in this field.

2.2.1. Vector Space Model

This approach has been performed over Chinese news document corpus. The training and test news documents are both represented as vectors in vector space model, similarity is calculated and K- most similar documents are selected.[13][14]

2.2.2. Improved KNN classification

This approach is an improvement over traditional KNN by adding time window strategy to the process, that reduces time complexity.[15]

2.2.3. Hierarchical clustering

This method obtains the set of candidate topics by agglomerative clustering, calculates the similarity between *ct* and each single previous topic within the latest period of time Δt , considers them similar if it is higher than the threshold.[16]

2.2.4. Related topic network

Related topics are determined by finding pair- wise similarities between stories, after representing the story with vector having weight components for its title and the main content.[17]

2.2.5. Proposed Incremental Algorithm

This method has been applied for Egyptian news website. It identifies readers trends to dynamically display the most expected news by using PI algorithm, which uses association rules over the user data and their IP's and their navigation of various topics.[18]

2.2.6. Dual center event model

This method modifies hierarchical clustering by considering only two event centers, recording the center and intensity of each event, calculating the similarity between new story and the two centers, then readjusting centers.[19]

2.2.7. Name Entity Recognition (NER) model

This method has been applied to Topic tracking for Bengali language, by forming an event vector of various NER features such as first name, middle name, prefix, suffix, designation, organization, location, date, time, for the two stories which are to be compared, then measuring similarity between two event vectors and checking it against a pre defined threshold value.[7][10]

2.2.8. Time adaptive boosting model

The method has been applied for Chinese news, the training set is enlarged dynamically, by adding the test stories which match with the training set and weighting each training example according to the time character, which decays exponentially, as a function of the difference in time between the test story and the first on-topic story.[20]

2.2.9. Lexical chains

In this model, an appropriate Lexical chain is chosen for each candidate word depending upon its relation with other words, calculating the similarity score between two sets of chains of the stories and checking it against threshold to declare two stories as related or not.[21]

2.2.10. Emerging topic tracking system

In this model, Topic Tracking is implemented by using three components, Area view system that derives the most relevant domains as per the keywords entered, Web Spider scans all these domains to collect all the modified and newly added HTML pages, Change Summarizer analyzes the updated and new pages and sentences with the highest average weight will be extracted to construct a summary for the user.[22]

2.2.11. Hidden markov model

This topic tracking system makes transition from the start state to the first topic on the basis of a transition probability.[23]

2.2.11. Based on keyword extraction

Keywords are index terms that contain most important information about the document. This method for topic tracking extracts keywords or keyterms from a news document and compares the news on the basis of these keywords, to decide whether they are tracking the same topic or not.[24]

3. TOPIC TRACKING FOR PUNJABI

Punjabi is the language of Punjab, spoken mainly in Northern parts of India. Punjabi is highly inflectional and agglutinating language providing one of the richest and most challenging sets of linguistic and statistical features resulting in long and complex word forms. Each word in Punjabi is inflected for a large number of word forms. It is primarily a suffixing language. An inflected word starts with a root and may have several suffixes added to the right. It is a free word order language.

Punjabi, like other Indian languages, is a resource poor language- annotated corpora, name dictionaries, good morphological analyzers, POS taggers are not yet available in the required measure. Although Indian languages have a very old and rich literary history, technological developments are of recent origin. Web sources for name lists are available in English, but such lists are not available much in Punjabi.

3.1. Approach

The topic tracking for Punjabi language has been experimented with two approaches. NER based approach and keyword extraction approaches have been implemented. The system determines whether two Punjabi news documents track the same topic or event or not. A number of features are extracted out of the text using the two approaches. The NER and keyword features of initial news document are compared with the respective features of target news document. The percentage of match or tracking same topic is evaluated.

3.2. Features extracted

3.2.1. NER Features

The name entities being extracted for Punjabi language in our experiment include Name, Time/ Date, Location, Designation, Organization. Name includes the prefix, first name, middle name, last name and then forming the complete name of a person. Time/date include date, month, week day, and year. Location refers to any location name within the document. Designation includes various roles or designation names. Organization refers to name of an organization. The NE's extracted are language dependent features and language independent features.

a. Language independent features: The language independent features include

- Context word feature: preceding and following words of a particular word.
- Presence of digits: may represent date, time, month, and year.
- Complete word: if a word is a complete word or it is a part of another word.

b. Language dependent features

A number of rules specific for Punjabi language are formed to extract the language dependent features.

i. Date/time rules

- Any format of the form dd/mm/yyyy, dd-mm-yyyy or dd.mm.yyyy, is extracted as date.
- Any format of the form yyyy-yy is extracted as year.
- When month name is found, it is extracted as month.
  - The previous word is checked, if it is of the form dd(<=31), then extracted as date
  - The next word is checked, if it is of the form yyyy, then extracted as year.
- When week day is found, it is extracted.
- If Punjabi word ਸੰਨ (sann) or ਸਾਲ (sāl) is found followed by three ([1-9][0-9][0-9]) or four([1-2][0-9][0-9][0-9]) digits, then it represents year.
- If Punjabi word ਈ. (ī.) is found, its previous word checked for three ([1-9][0-9][0-9]) or four([1-2][0-9][0-9][0-9]) digits, then it represents year.

ii. Designation rule

- When designation name found, it is extracted.
- For a two word designation such as ਡਿਪਟੀ (ḍipṭī) as first word, next word is checked for designation. If it is, then the word and the next word are collectively taken as designation. e.g. ਡਿਪਟੀ ਕਮਿਸ਼ਨਰ (ḍipṭī kamishanar)

iii. Organization rule

- When an organization suffix such as ਕੰਪਨੀ (kampnī), ਕਮੇਟੀ (kamēṭī), ਕਲੱਬ (kalabb), ਦਲ (dal), ਬੋਰਡ (bōraḍ), ਵਿਭਾਗ (vibhāg), ਆਰਗੇਨਾਈਜੇਸ਼ਨ (ārgēnāījēshan), ਐਸੋਸੀਏਸ਼ਨ (aisōsīēshan), ਯੂਨੀਅਨ (yūnīan) etc is found, then its previous word and the suffix are collectively extracted as organization.

iv. Name rules

- Prefix rule
  - When some prefix is found, its next word is taken as first name.
  - Word next to first name is checked for middle name or last name. If it is, then both are concatenated.
  - Word next to middle name is checked for last name. If it is, again complete name is formed by concatenation.
- Middle name rule
  - When Middle Name is found the previous word is taken as First Name.
  - Also the word after the Middle Name is checked whether it is Last Name or not. If it is Last Name, the First Name, Middle Name and Last Name are concatenated and complete name is formed. Otherwise just First Name and Middle Name is concatenated as name.
- Last name rule
  - When Last Name is found the previous word is checked if it is Middle Name or not. If it is not Middle Name then it is taken as First Name.
  - If it is Middle Name then the word before the Middle Name is taken as First Name. The Founded First Name, Middle Name and Last Name are concatenated to form name.

v. Location rule

- ਵਿਖੇ (vikhē) rule: when Punjabi word ਵਿਖੇ is found, its previous word as extracted as a location name.
- Location suffixes ਪੁਰ (pur) or ਗੜ੍ਹ (gaṛh) found, then complete word is extracted as location name.

3.2.2. Keyword extraction Features

There are number of approaches that exist for keyword extraction. Since Punjabi language has certain different characteristic features such as different grammatical features, sentence structure, morphology etc. So keywords are extracted based on mixed approach. Different functions produce keywords first individually and then merging those features together to get better output in form of effective keywords that represent the document well.

- So, Keywords are extracted based on mixed approach
- Candidate keywords extracted by each approach are weighted
- The highly weighted keywords are merged as final keywords.

Various features used for keyword extraction include:

i. Term frequency: It describes the count or the number of times a word occurs in a given file and its ranking with respect to other words in the file.

ii. Noun frequency: Since term frequency cannot alone be considered as the complete measure to find keywords. Those words having part of speech as noun are also important ones to be considered as keywords. So high frequency nouns are considered as keywords, which are taken to be a small percentage of the total number of terms in the document.

iii. Title: title of the text file also plays a very crucial role in deciding the final keywords since from title itself one could judge the complete content of the file. Hence title feature provides the important keywords after eliminating the stop words.

iv. Cue phrases: Cue phrases such as summary phrase or transition phrase are considered as most important for depicting the main theme of whole text. So words in the cue phrase are considered as keywords. Some Punjabi cue phrases include words such as ਨਤੀਜਾ (natījā), ਅੰਤ (ant), ਸਿੱਟਾ (siṭṭā), ਨਿਚੋੜ (nicōṛ), ਸੋ (sō), ਅਖੀਰ (akhīr) etc.

3.3. Gazetteer lists

- Cue phrase
- Names of months
- Days of a week
- Prefix for name
- Designation names
- Organization suffix
- Middle names
- Last names
- Stop words
- Location names
- Nouns

3.4. Methodology

First, all stop words are removed from the document, because they slow down the process as they are large in number and are of no use. Then NER rules are implemented which use the gazetteers lists to extract Names, time/ date, location, designation, organization from the document. Then the keyword extraction rules are implemented which use the gazetteers lists to extract keywords from title, cue phrase and nouns which are highly frequent. The features extracted from both the news documents are compared to get the match percentage between them. The similarity has been measured using overlap coefficient.

3.5. Evaluation metrics

The performance of system is measured using precision (P), recall (R) and F-measure (F) for NER and keyword extraction modules. The topic tracking system is evaluated by calculating the match percentage between the documents.

The precision measures the number of correct NEs/keywords, obtained by system, over the total number of NEs/ keywords extracted by the system.
P= number of correct NEs or keywords/total number of NEs or keywords.

The recall measures the number of correct NEs/ keywords, obtained by the system over the total number of NEs/ keywords in a text that have been used for testing.
R= number of correct NEs or keywords /total number of NEs or keywords in a text
The F-measure represents harmonic mean of precision and recall.
F= 2RP/R+P

3.6. Implementation details

The system has been implemented using vb.net platform and gazetteers lists are stored as tables in the database. The experiment required news documents as input, which are to be compared to check if they track same topic or not. Such test documents are taken from Punjabi news web sources such as likhari.org, jagbani.com, ajitweekly.com, punjabispectrum.com, europevichpunjabi.com, quamiekta.com, sahitkar.com, onlineindian.com, europesamachar.com, parvasi.com etc.

3.7. Experimental results

Four experiments have been carried out to implement topic tracking for Punjabi. In the first experiment, NER module has been tested. In the second experiment, the keyword extraction module has been tested. The third experiment tested the topic tracking system by evaluating using NER technique alone and keyword extraction technique alone. After that, topic tracking is implemented by combining both the techniques. In the last experiment, a number of similarity measures have been analyzed to evaluate which similarity measure finds the best results for topic tracking.

3.7.1. Experiment 1

In our first experiment, the NER module has been tested.

Table 1. NER results.

NE Class	Precision(%)	Recall(%)	F measure(%)
Person	74.52	62.86	65.67
Location	91.52	92.89	91.25
Organization	90.27	90.10	88.77
Designation	98.84	87.09	91.98
Date/Time	94.79	89.79	91.75
Total	89.98	84.55	85.88

This experiment shows good results for all the features, except the Name feature for which the values are comparatively low. It is because many first names in Punjabi are also common nouns which the system is not able to recognize.

3.7.2. Experiment 2

In our second experiment, keyword extraction module has been tested.

Table 2. Keyword extraction results.

Features	Precision(%)	Recall(%)	F measure(%)
Title	97.04	97.58	97.30
Cue Phrase	100	97.56	98.41
Noun	97.81	75.48	83.39

This experiment shows good results for all the features used to extract keywords. The little lack in the percentage is due to the reason that some common nouns in Punjabi are also names due to which wrong words are found sometimes. And after the removal of stop words, the system can not find important abbreviations that contain that stop word. But results can be improved by incorporating more features to the existing and considering the issues that lower the results.

3.7.3. Experiment 3

In the third experiment, topic tracking is implemented by combining the features from both techniques. That is, the system evaluates the NER and keyword features from the initial and target news documents, and compares the features of initial document with the target document and evaluates the percentage match between the two news documents, so that it can be concluded whether the news are related to each other or not, that is, whether they are tracking the same topic or not.

It shows good percentage match for similar news and low percentage match for the news of different topics.

For similar news:

Table 3. match percentage for similar news.

Feature	Match percentage
Time	80
Location	100
Name	100
Organization	100
Designation	66.67
Title	75
Cue phrase	100
Noun	62.5
Total NER	89.32
Total keyword	79.16
Total	85.52

The system shows high percentage match for similar news documents, and with this high percentage, the two news documents can be said to be tracking the same topic.

For news of different topics:

Table 4. match percentage for dissimilar news.

Feature	Match percentage
Time	25
Location	0
Name	0
Organization	0
Designation	0
Title	25
Cue phrase	0
Noun	0
Total NER	12.5
Total keyword	28.57
Total	15

The system shows low percentage for the news documents of different topics. With such low percentage, it can be concluded that they do not track same topic.

3.7.4. Experiment 4

This experiment evaluates topic tracking results with different similarity measures. The similarity measures chosen for the experiment include jaccard coefficient, overlap coefficient, dice coefficient, cosine coefficient. It has been evaluated that overlap coefficient shows high match percentage for the similar documents which track the same topic.

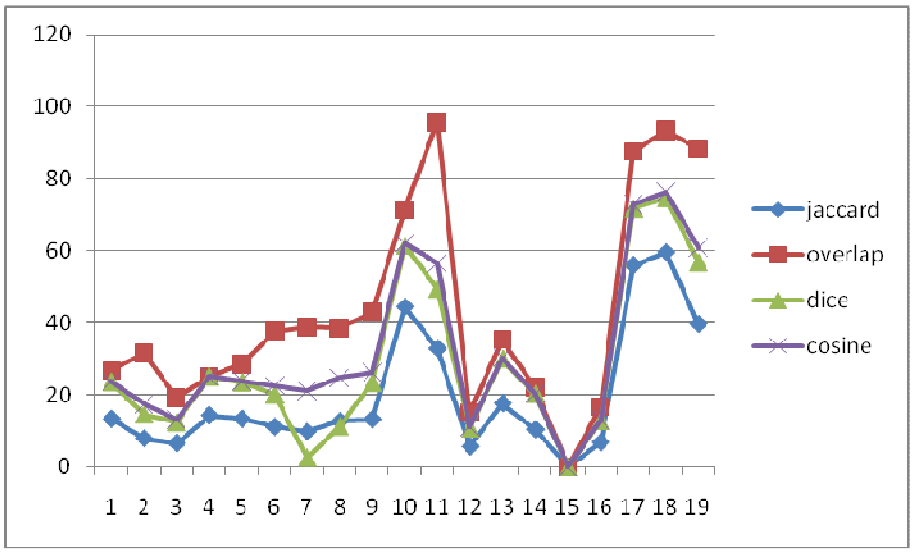

Figure 3. Comparison of different similarity measures

4. CONCLUSION & FUTURE SCOPE

Not much work has been done in Topic Tracking for Punjabi and other Indian languages. In this paper, we have reported our work on NER, keyword extraction and topic tracking for Punjabi. We have prepared the system with the combination of two approaches, i.e. combining a number

of features from NER and keyword extraction to generate effective topic tracking system. The language dependent features and language independent features are formed and analyzed. The approach uses the gazetteer lists created from the dictionary with part of speech tagging, morphological analyzer, Punjabi vocabulary. Hence, NEs such as date/ time, location, person name, organization, designation and keywords from title, cue phrase and high frequency noun are extracted. The system shows good results for all features independently and the total results for the system are improved with the combination of these features resulting in effective topic tracking system.

Future works include incorporating more features to improve the existing results. The features that can be included are position weight algorithm which weighs words according to their position of occurrence in the document, length of the word by giving more importance to the lengthy words as compared to the short words, informative feature selection such as bold, italic, underlined words in keyword extraction. As many first names in Punjabi are also common nouns, this limitation lowers the performance of the NER system. This issue can be considered to improve the system. More NE's can be extracted such as monetary expressions, measurement expressions etc to improve the performance of the system.

REFERENCES

[1] Milos Radovanovic, Mirjana Ivanovic, (2008), "*Text Mining: Approaches and Applications*", Novi Sad J. Math, Vol. 38, No. 3: 227-234.

[2] Anna Stavrianou, Periklis Andritsos, Nicolas Nicoloyannis, (2007), "*Overview and Semantic Issues of Text Mining*", Sigmod Record, Vol. 36, No. 3

[3] Navathe, Shamkant B. and Elmasri Ramez, (2000), "*Data Warehousing and Data Mining*", in "*Fundamentals of Database Systems*", Pearson Education pvt Inc, Singapore, 841-872

[4] Vishal Gupta, G.S. Lehal, (2009), "*A Survey of Text Mining Techniques and Applications*",in Journal of Emerging Technologies in Web Intelligence

[5] Xianfei Zhang, Zhigang Guo, Bicheng Li, (2009), " *An Effective Algorithm of News Topic Tracking*", Global Congress on Intelligent Systems,IEEE

[6] JingQiu, LeJian Liao, XiuJie Dong, (2008), " *Topic Detection and Tracking for Chinese News Web Pages*", International conference on Advanced Language Processing and Wen Information Technology, IEEE

[7] Wang Xiaowei, JiangLongbin, MaJialin, Jiangyan, (2008), "*Use of NER Information for Improved Topic Tracking*", Eighth International Conference on Intelligent Systems Design and Applications, IEEE

[8] Yan Liu, Nan Lv, Junyong Luo, Huijie Yang, (2009), "*Subtopic Based Topic Evolution Analysis*", International Conference on Web Information Systems and Mining, IEEE

[9] Xiangju Qin, Yang Zhang, (2008), "*Improving the performance of Topic Tracking System by Ensemble*", International Conference on Computer Science and Software Engineering, IEEE

[10] Anup Kumar Kolya, Asif Ekbal, Sivaji Bandyopadhyay, (2009), "*A Simple Approach for Monolingual Event Tracking System in Bengali*", 8th International Symposium on Natural Language Processing, IEEE

[11] Omid Dadgar, "*Topic Detection and tracking*", Available: www.tcnj.edu/~mmmartin/.../TDT/TopicDetectionTracking04.ppt

[12] Topic Detection and Tracking, Available: www.projects.ldu.upenn.edu

[13] Shengdong Li, Xueqiang Lv, Yuqin Li, Shuicai Shi,(2009), " *Study on Feature Selection Algorithm in Topic Tracking*"

[14] Shengdong Li, Xueqiang Lv, Qiang Zhou, Shuicai Shi,(2010), " *Study on Key Technology of Topic Tracking Based on VSM*", Proceedings of the 2010 IEEE International Conference on Information and Automation, June 20-23, Harbin, China

[15] Hongxiang Diao, Zhansheng Bai and Xilin Yu, (2010), "*The Application of Improved K-Nearest Neighbor Classification in topic tracking*", International Conference on Educational and Information Technology, IEEE

[16] Xiang Ying Dai, Qing Cai Chen, Xiao Long Wang and Jun Xu, (2010), "*Online Topic Detection and Tracking of financial News based on Hierarchical Clustering*", Proceedings of the 9th International Conference on Machine Learning and Cybernetics, IEEE

[17] Chao Chang, Daniel Zeng, Huimin Zhao,(2010), " *Related Topic Network*", IEEE

[18] Laila Mohamed ElFangray, (2009), " *Applying an Enhanced Algorithm for Mining Incremental Updates on an Egyptian Newspaper Website*", 5th International Joint Conference on INC, IMS and IDC, IEEE

[19] Wei Wang, JunZheng, Wu Yang and Yongtian Yang, (2008), "*A Dual Center Event Description Model Used in Event Tracking*", 2009 World Congress on Computer Science and Information Engineering,IEEE

[20] Huizhen Wang, Jingbo Zhu, Duo ji, Na Ye and Bin Zhang, (2005), "*Time Adaptive Boosting Model for Topic Tracking*", Proceeding of NLP-KE'05,IEEE

[21] Paula Hatch, Nicola Stokes and Joe Carthy, "*Topic detection, a new application for Lexical Chaining*"

[22] Khoo Khyou Bun and Mitsuru Ishizuka, (2001), "*Emerging Topic Tracking System*",IEEE

[23] JP.Yamron, Carp L. Gillick, S.Lowe, P. Van Mulbregt, (1997), "*Event Tracking and Text Segmentation via Hidden Markov Models*",IEEE

[24] Canhui Wang, Min Zhang, Liyun Ru, Shaoping Ma,(2008), " *An Automatic Online News Topic Keyphrase Extraction System*",IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

Authors

Kamaldeep Kaur is pursuing M.Tech in Computer Science & Engineering at University Institute of Engineering & Technology, Panjab University Chandigarh. She has done B.Tech. in computer science & engineering from Guru Nanak Dev Engineering College, Ludhiana in 2008. She is among university toppers. She secured 82% Marks in B.Tech. Kamaldeep is devoting her research work in field of Natural Language processing. She is also a merit holder in 10th and 12th classes of Punjab School education board.

Vishal Gupta is Lecturer in Computer Science & Engineering Department at University Institute of Engineering & Technology, Panjab University Chandigarh. He has done M.Tech. in computer science & engineering from Punjabi University Patiala in 2005. He secured 82% Marks in M.Tech. He did his B.Tech. in CSE from Govt. Engineering College Ferozepur in 2003. He is also pursuing his PhD in Computer Sc & Engg. Vishal is devoting his research work in field of Natural Language processing. He has developed a number of research projects in field of NLP including synonyms detection, automatic question answering and text summarization etc. One of his research papers on Punjabi language text processing was awarded as best research paper by Dr. V. Raja Raman at an International Conference at Panipat. He is also a merit holder in 10th and 12th classes of Punjab School education board.