Volume 3, Issue 8, August 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Role of Text Mining for Lexical Monolingual Punjabi Dictionary

Gurdeep Singh
Assistant Professor, M.C.A Dept.
G.N.D.E.C, Ludhiana, India

R.S. Uppal
Associate Professor, C.S.E Dept
B.B.S.B.E.C, Fathegarh Sahib, India

Manpreet Singh
Assistant Professor ,I.T Dept.
G.N.D.E.C, Ludhiana, India

Abstract— This document gives brief description about various Punjabi dictionaries that have been developed and text mining techniques. The various text mining techniques are briefly studied and are compared according to different parameters. The best technique is then proposed for developing a text mining model for Punjabi data dictionary. This purposed model can be used to create a dynamic ever updating dictionary model.

Keywords— Text Mining, Punjabi Dictionary, PLSA, LSA, LDA, CTM

I. INTRODUCTION

There is huge amount of Punjabi data and information is available at web. Moreover thousands of new words are created and changed every day across the world. The amount of information is increasing exponentially. So there is a need of special tool that can update online dictionaries from time to time and present dictionaries as a dynamic tools.

Dictionary[1] is a book, optical disc, mobile device, or online lexical resource (such as Dictionary.com) containing a selection of the words of a language, giving information about their meanings, pronunciations, etymologies, inflected forms, derived forms, etc., expressed in either the same or another language; lexicon; glossary. Print dictionaries of various sizes, ranging from small pocket dictionaries to multivolume books, usually sort entries alphabetically, as do typical CD or DVD dictionary applications, allowing one to browse through the terms in sequence. All electronic dictionaries, whether online or installed on a device, can provide immediate, direct access to a search term, its meanings, and ancillary information: an unabridged dictionary of English; a Punjabi-English dictionary.

There are number of dictionary tools[2] that are available in Punjabi. Some of the paper dictionaries have been converted into electronic dictionaries while some are specially being developed as electronic dictionaries. Many online dictionaries have also been developed. Some of the developed online Punjabi dictionaries are as given below.

A. Punjabi Kosh

Punjabi Kosh is an English-to-Punjabi and Punjabi-to-English dictionary designed by Noah Hart. It allows keyboard entry and dictionary use in both languages. There are also a number of learning games and lessons for Punjabi students. It is a very useful tool for Punjabi learners. The dictionary is freely available at[3].

B. Punjabi Shabdkosh

Punjabi Shabdkosh is a Punjabi to English dictionary developed by Harwinder Singh Tiwana. The dictionary is freely available at[4].

C. Punjabi Dictionary by CDAC

An ISCII based Punjabi-English dictionary developed by CDAC is made available on the language CD freely made available by MCIT. The GIST typing tools have to be used for typing and searching for words in Punjabi.

D. Gur Shabad Ratanakar Mahankosh

Gur Shabad Ratanakar Mahankosh, popularly called as Mahankosh, is the first dictionary of Sikh Scripture and books on Sikh Religion. It is also a classical reference book of Sikh history, philosophy and contemporary Sikh states. The complete Mahankosh has been digitized by Bhai Baljinder Singh Rarewal and the pdf file is available for download from[5].

Text Mining[6] referrers to the process of extraction of high quality information from the bulk data and the information extracted should be applicable to the requirement of the users. Text mining is a variation of a field called data mining, which tends to find patterns from large databases. Text mining is also known Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting information and knowledge from unstructured text. Text mining is a new interdisciplinary field which focuses on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information is stored as text, text mining is believed to have a high commercial value. Information may be discovered from many sources, yet unstructured texts remain the largest readily available source of knowledge. The aim of text mining is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process. KDT plays an increasingly significant role in growing applications, such as Text Understanding. Text mining and data mining are same, except that data mining tools are designed to handle structured data from databases , but text mining can work with unstructured or semi-structured documents sets such as emails, full-text documents and HTML files etc. As a result, text mining is a much better solution

for corporate. However, most research and development efforts have centred on data mining efforts using structured documents. There are number of Text Mining techniques that can be used to make dynamic Punjabi dictionary[15] but their performance may vary. The performance of these techniques can be visualized by factors such as performance, accuracy, synonymy's extraction, time complexity, space complexity etc.

II. MAJOR TECHNIQUES IN TEXT MINING

Text Mining has been developed from distinct models such as Associative text mining, Latent semantic analysis method to generative methods such as Probabilistic latent semantic analysis, latent dirichlet allocation and correlated topic model. The latter two techniques are part of the topic model.

A. Latent Semantic Analysis

Latent Semantic Analysis (LSA)[7] represents the words used in it document and any set of these words—such as a sentence, paragraph, or essay—either taken from the original corpus or new, as points in a very high (e.g. 50-1,500) dimensional "semantic space". LSA is closely related to neural net models, but is based on singular value decomposition, a mathematical matrix decomposition technique closely akin to factor analysis that is applicable to text corpora approaching the volume of relevant language experienced by people. A passage or paragraph is a linear equation and its meaning is defined by the sum of all the words in the passage i.e Paragraph(pg) = $m(word_1) + m(word_2) + m(word_n)$. A matrix is conducted containing word count per paragraph is constructed containing large number of texts. A mathematical technique called SVD (Singular Value Decomposition) is used to reduce the number of columns while preserving the similarity structure among rows. The Words are then compared by taking the cosine of the angle between the two vectors formed by any two rows. Values near to 1 represent very similar words while nearly close to 0 represent very dissimilar words. SVD returns the vector of the singular values. In LSA Singular value decomposition technique divides the original matrix into three matrix a document, eigenvector matrix, an eigen value matrix, and a term eigenvector matrix. The original matrix can be reformed from these matrix by multiplying them. The values that are near 1 can be regarded as Synonyms and in dictionary this technique can be used to extract the words from the paragraph as synonyms. LSA basically has three limitations. It works on the operation of reduction of matrix and not on the robust probability theorem. The number of factors cannot be judged as such and mainly depends on the human factor. The problem of antonyms is partially dealt with LSA which makes it less accurate to be used as a major technique for dictionary meaning extraction.

B. Probabilistic Latent Semantic Analysis

Probabilistic latent semantic analysis (PLSA) (PLSI),[8] especially in information retrieval circles) is a statistical technique for the analysis of co-occurrence data. PLSA technique is almost similar to Latent Semantic Analysis but in semantic analysis singular value decomposition technique is used to decode the current matrix into three different matrix and in probabilistic latent semantic analysis Exception maximization algorithm is used. The probabilistic latent semantic analysis model assumes that documents are generated throughout three steps. First, a document d is generated with probability P(d). Second, topic z is picked with probability P(z|d). Third, each word w in a topic is generated with probability P(w|z). The solution to the problem is found by EM algorithm and unknown parameters are also calculated, PLSA estimates topic probabilities *P(z)*, document probabilities given topics *P(d|z)*, and word probabilities given topics *P(w|z)*.

C. Latent Dirichlet allocation

Topic models are a group of algorithms that uncover the hidden structures in document collections. These algorithms help us to develop new methods to search, browse and summarize large archives of texts. An LDA[9] model is a generative model originally proposed for doing topic modelling. The process of Latent Direchlet allocation method is carried out in Three steps. First, Poisson distribution is used to determine the number of words by sampling. Second, a distribution over topics for a document is drawn out from the Dirichlet distribution. At last based on the document-specific distribution, topics are generated, and then words for each topic are generated. LDA is effective for modelling long length documents that contain multiple topics. LDA has been used in vast areas such as topic detection, emotion detection, and word sense disambiguation.

D. Correlated topic model

The correlated topic model (CTM) is also part of topic modelling. CTM has been developed to overcome the limitations of LDA. CTM follows the same process of LDA, but the difference is in the use of logistic normal distribution rather than Dirichlet distribution to capture topic relations. The components in the Dirichlet distribution are logically independent and thus they don't have a relationship with other topic. Due to this reason the occurrence of words in other topic is minimized. CTM employs more flexible logistic normal distribution by considering covariance structure among the components of proportions. It is easy to find out synonyms within the same topic. For words with different meaning can be searched in other topics.

III. COMPARISON OF MAJOR TEXT MINING TECHNIQUES

Our main focus is to discover the best technique that is best for extracting synonyms for the Punjabi dictionary from the existing dictionary documents to create a ever updating dynamic dictionary. The problem is similar to the extracting the word from the English documents with certain assumptions. Stop words in Punjabi language are different from that which are used in English. The next step is to evaluate the nouns from the document. The words with higher occurring Term Frequency-Inverse Sentence Frequency[10]. Apart from these two problems the rest technique is similar to the extraction of English words from the document. Thus a common problem can be taken to compare these techniques and

evaluate their performance. Spam mail[9] is such a problem in information technology because it wastes the time of user for deleting spam mails. For solving this problem, various spam filtering methods such as machine learning, Bayesian classification, and statistical learning have been developed. Spam filtering can be regarded as text categorization problem for classifying spam and legitimate mail and also as information retrieval problem for selecting the highest matched mail between two types of mails. Because many text mining methods are used for spam filtering, and the performance measurement is nearly objective, spam filtering could be another for performance comparison. For the comparison of four methods, we specifically pay an attention to single type of an application rather than a mixed method combined with text mining methods.

LSA has been applied to spam filter from a long time [11, 12]. The spam filtering in LSA has two phases. In the training phase, singular value decomposition(SVD) is applied to the term-document matrix that is obtained from a training data set. The classification phase is consisted of the query indexing with test mails and the retrieval of the closest mail from the training set. If the closest mail from the training set is a spam mail, it is classified as a spam mail; otherwise it is classified as a genuine mail. Table 1 shows the performance of four text mining methods in terms of various evaluation metrics such as precision, recall, accuracy, and weighted accuracy (WA), and total cost ratio (TC). There is a major risk in detecting a genuine mail as a spam mail because it would lead to loss of important data. For incorporating asymmetric risk, we set the cost ratio with 9 for WA and TC with reasonable cases. The performance of different text mining techniques are shown in table 1.

TABLE 1

PERFORMANCE[9] OF FOUR TEXT MINING METHODS IN SPAM FILTERING

Parameters	**LSA**	**PLSA**	**LDA**	**CTM**
Spam Precision	0.881	**0.895**	0.892	0.891
Spam Recall	0.942	**0.955**	0.949	0.949
F_1	0.911	**0.924**	0.920	0.920
LS/Total	0.021	**0.018**	**0.018**	0.019
LS/SL	2.178	**2.524**	2.250	2.292
Accuracy	0.969	**0.974**	0.973	0.972
Weighted Accuracy	0.974	**0.978**	0.977	0.976
Total cost ratio	4.123	**4.793**	4.688	4.597

PLSA can be applied to spam filtering using the same approach as in LSA using information retrieval and text categorization method. By applying PLSA to the training set, we estimate parameters of the model and construct a lower-dimensional representation in the factor space. Then, test mails which were not part of a training set is folded in by fixing the $p(w|z)$ and calculating weights $p(z|q)$. The third column of Table 1 shows the performance of PLSA. Compared to LSA, PLSA shows better performance. The fourth column of shows the performance of LDA. Compared to PLSA, the precision, recall, and F1 are decreased, but those performances are higher than LSA. The rate of false positive to total is the same with LDA, but false negative is increased. Overall, the performance of LDA is almost the same with PLSA, and there is only 2% decreasing of TC. CTM is applied to spam filtering with the same way of LDA. The recall, precision, and F1 are very close to those of LDA. Compared to LDA, the number of SS is increased but the number of LL is decreased, where SS denotes that a spam mail is classified as a spam mail. However, false positive and false negative are almost the same with LDA. Because of the reduced number of SS, TC is decreased about 2%. In terms of TC, among four text mining methods, PLSA shows the highest performance, and next to LDA, CTM, and LSA in order. Also, between discriminative models and probabilistic models, probabilistic models show better performance in terms of precision, false positive, and TC. The experiment suggests that generative models may be promising for spam filtering.

IV. TEXT MINING PROCESS FOR PUNJABI DICTIONARY

The process of text mining for Punjabi dictionary is same with certain assumptions. First the words that are sorted and removed are different than the normal English text. It is difficult for the computer to differentiate words depending upon noun, verbs and adverbs, because for such thing large database of language is required which is not available at the current stage. The rest technique of text mining is similar. . It has been observed and proved that the PLSA technique is efficient than the other text mining techniques. Hence we will be using PLSA as the major text mining technique in Text mining tool for Punjabi dictionary. The process of text mining begins with first removing the stop words from the documents. Example of such stop words are shown in table 2.

TABLE 2

EAMPLE OF STOP WORDS OF PUNJABI

ਹਨ	ਕਰ	ਪਰ	ਕਰਕੇ
ਦੀ	ਵੀ	ਅਤੇ	ਵਾਲੇ
ਸੀ	ਤੇ	ਹੋ	ਓਹੀ
ਨਾਂ	ਜਿਸ	ਵਾਲੇ	ਵੀ

After the extraction has been performed then the data is transformed and the clusters of the similar meaning are formed. After the extraction and cluster formation PLSA is applied on the extracted data. The probability is then calculated using the given equation

$$P(t \mid doc) = \sum_{k=1}^{K} P(t \mid k) P(k \mid doc)$$

After calculating the probability the data is sent into the temporary database from where the user can check and define according to the human intelligence. The main advantage of using the text mining technique is that it provides the user with more options to pick the desired synonyms, antonyms etc from the database.

V. CONCLUSIONS

Different text mining and topic modeling techniques can be used for extraction of data from the document .From the above comparison it is clear that the Probabilistic latent semantic analysis technique is much better to extract text from the document to list out synonyms and antinomy's. And the same technique when applied to dictionary data extraction will provide us better result than the other techniques. The extraction of text through these techniques will help the dictionary to update its content from the existing dictionaries thus providing us a dynamic model that will update dictionaries from time to time.

REFERENCES

[1] http://dictionary.reference.com/browse/dictionary?s=t&ld=1148.
[2] Lehal. G.S. (2009), "A Survey of the State of the Art in Punjabi Language Processing", Language in India Strength for Today and Bright Hope for Tomorrow Volume 9 ISSN 1930-2940.
[3] http://punjabikosh.googlepages.com/.
[4] http://www.4shared.com/file/39293942/9d333376/Punjabi_Shabdkosh__English_to_Punjabi_Dictionary_.html?s=1.
[5] http://www.ik13.com/mahan_kosh.htm.
[6] Gupta V. and Lehal G.S. (2009), "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1**.**
[7] Thomas K Landauer, Darrell Laham (2001)." An Introduction to Latent Semantic Analysis" *Discourse Processes,* 25, 259-284.
[8] Hofmann Thomas (2005)." Probabilistic Latent Semantic Indexing ", Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval.
[9] lee S., Song J. and Kim V. (2010), "An Empirical Comparison of Four Text Mining Methods".
[10] Neto, Joel al., "Document Clustering and Text Summarization", In: Proc. of 4th Int. Conf. Practical Applications of Knowledge Discovery and Data Mining, London, 2000, pp. 41-55.
[11] Bellegarda, J., Naik, D., and Silverman, K., "Automatic junk e-mail filtering based on latent content", 2003, 465-470.
[12] Gansterer, W., Janecek, A., and Neumayer, R., "Spam filtering based on latent semantic indexing", *Survey of Text Mining II: Clustering, Classification, and Retrieval*, 2008, 165-183.
[13] Gupta V. and Lehal G.S. (2011), "Automatic Keywords Extraction for Punjabi Language", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, September 2011 ISSN.
[14] Bendersky M. and Smith D. (2012), "A Dictionary of Wisdom and Wit: Learning to Extract Quotable Phrases", Workshop on Computational Linguistics for Literature, pages 69–77.
[15] H´eja E. and Tak´acs D. (2012) "Automatically Generated Customizable Online Dictionaries", Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 51–57.

