International Journal of Computer Applications (0975 - 8887)
Volume 132 - No.15, December 2015

Review on Pattern based Document Modelling Techniques

Jimsy Johnson
PG Scholar
Department Of Computer Science and Information Technology
Perumon, Kerala

Smitha C.S.
Assistant Professor
Department Of Computer Science and Engineering
Perumon, Kerala

ABSTRACT

Topic Modelling has been widely used in the fields of machine learning, text mining etc. It was proposed to generate statistical models to classify multiple topics in a collection of document, and each topic is represented by distribution of words. But many variants of topic models have been proposed and most of them are based on the concept of bag-of-words and it ignores the association of words for representing topics. Nowadays patterns are used for representing topics, since they have more discriminative power than words for representing multiple topics in a document. A detailed survey of some of the most important methods for topic modelling is presented. A brief comparison among the key techniques is also presented to complete the survey.

General Terms

Topic Modelling, Pattern mining

Keywords

Information Retrieval,Information Filtering,Topic models

1. INTRODUCTION

Recent years there was a dramatic increase in the web information.Therefore advanced techniques are needed to understand and analyse the user's information needs and to deliver the best results based on the user's information needs. Knowledge discovery and data mining have work together with a need for meaningful data intos user information and knowledge. Hence data mining has been used as an efficient step for the discovery of knowledge in databases. Data mining is the process of extracting hidden patterns or information from a large database. Data mining also known as Knowledge discovery is a computer aided process in analysing and digging large amount of data and then extracting the meaning of the data.

Text Mining is the analysis of data in a large amount of text data. It is the process of deriving high quality information from text. In the past years, several data mining technologies have been presented to perform different knowledge tasks. These techniques includes association rule mining [1], frequent item set mining[2], sequential pattern mining[3], maximum matched pattern mining, closed pattern mining[4] etc. The challenging issue towards this approach is to find the accurate knowledge or features from large amount of available information.Features can be binary,categorical or continuous.Feature extraction firstly starts with an initial set of measured data and builds some values(features),which are intended to be informative and non redundant.The main thing behind feature extraction is dimensionality reduction.

To compress large data to manageable knowledge topic modelling can be used. Topic modelling[5] is one of the prominent area that comes under text mining.Topic modelling automatically classifies documents in a collection by a number of topics and represents each and every documents with multiple topics and their distributions.Nowadays several topic modelling techniques have been used.The main motive behind in doing all these are how efficiently produce more accurate and efficient information based on the user needs.The main challenges in the field of topic modelling are the problem of polysemy and synonymy.Polysemy means the words which have multiple meaning with it,whereas synonymy means the state of being synonymous(similar).The topic based representation have the advantage of avoiding the problems of semantic confusion compared with the traditional text mining techniques.From recent years there are several types of topic modelling techniques have been evolved.Each of which have its own pros and cons.

2. LITERATURE REVIEW

This topic presents a critical review of the literature essential to addressing the literature essential to addressing and analyses current theory and methodologies that have been used in the area of document modelling.

The ultimate aim of all data mining tasks are information filtering.The word information filtering means that a system to remove redundant or unwanted information from an information stream or document using automated or computerized methods prior to presentation to a human user.The related research areas includes Information Filtering,Text mining and Topic modelling.All these areas are overlapped or in other words these areas are dependent on each other.

3. PATTERN MINING TECHNIQUES

The meaning of the word pattern is "a regular and intelligible form or sequence discernible in the way in which something happens or is done". When it comes to the area of text mining, pattern extraction have a great influence.In past years,text mining,topic modelling etc. are done with the help of individual words(terms).But it is not efficient,because it takes large time to work on the individual terms in a document.so the concept of term-based approach is out dated.Then a new and much more efficient method was discovered by data mining experts.Phrase based approach.Through researches it can be find that phrase based approaches can do well with modelling the document than the traditional term-based approach,Since it carries more semantic informations.The main advantage of using this method is that,phrases are less ambiguous and more discriminative than individual words in describing a document.Phrase-based approach also have some disadvantages.They are, phrases have statically inferior properties, low frequency of occurrence and many phrases are redundant and noisy.So there is a need for finding new concepts for effective pattern mining techniques.As year goes several techniques were introduced by several experts.Some of them are Sequential patterns,Matching patterns,Maximum matched patterns etc..The ultimate aim of the data mining experts are to reduce the search time of the user and to provide a better result i.e., gives the accurate information that a user wants.

The process Topic Generation includes Dataset Preparation,Topic Generation, Construction of new datasets and the final Generate optimized Topic representations.The step wise procedure is shown in the fig:1

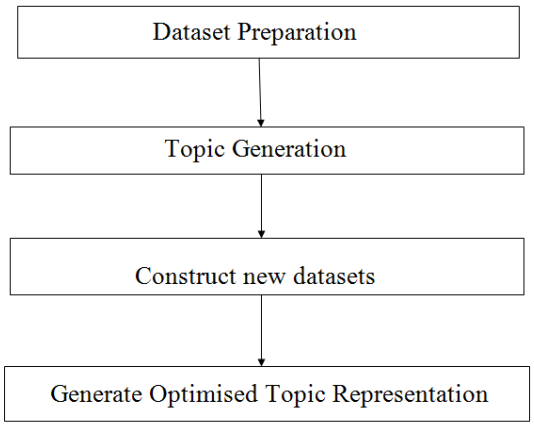

Fig. 1. Topic Generation

3.1 Term-based Representations

Term Frequency Inverse Document Frequency(TFIDF):.
TFIDF is the most effective and efficient feature extraction method that has been used widely in the field of information filtering and information retrieval.It is a numerical statistic which is used to show how important a word is to a document in a collection or corpus.It normally uses as weighting factor in text mining and information filtering.The term $tf - id_f$ is composed of two terms. The first term tf computes the normalized term frequency and second term is Inverse Document Frequency (IDF),computed as logarithm of the total number of documents in the corpus divided by the number of documents where the particular terms appears.

$$tf * idf(t,d) = tf_t * log\frac{|D|}{df_i} \quad (1)$$

$log\frac{|D|}{df_i}$ assigns high weights to more discriminative terms appearing in a few documents,whereas low weights to those common terms that are spread over many documents.

$$TF(t) = \frac{Numberoftimestermtappearsinadocument}{Totalnumberoftermsinthedocument} \quad (2)$$

$$IDF(t) = log_e\frac{Totalnumberofdocuments}{Numberofdocumentswithtermtinit} \quad (3)$$

3.1.1 Method Combined with Information-theory Function.
Another kind of weighting term approach is to adopt feature selection metrics,such as information gain, gain ratio, and mutual information. The main objective of this approach is to select the most relevant and discriminating features based on information-theoretic functions. However,this approach is not always superior over TFIDF methods, whose performances depend on different tasks.The primary aim of all term-based representations was to improve the statistics of a single term,but it neglects the semantic accuracy ,which is the main problem of term-based representations.Inorder to solve this, scholars use combinations of single words to solve the problem of semantic ambiguity.In general phrases carry more specific meaning than single words.

3.2 Phrase-based Representation

Data mining techniques were applied to text mining and classification by using word sequences as descriptive phrases (N-Gram)[6], from document collections. But the performance of N-Gram is restricted due to low frequency of phrases in documents.

Although phrases and N-Gram are stronger at interpreting semantic meaning, they perform less well with statistical properties in matching representations with documents when compared with term-based representation. In order to balance the statistical and semantic properties,researchers propose to extract pattern-based features for representing the user's interests.

3.3 Pattern-based Representations

Association rules are the patterns that are discovered from a given data set.Two important measures used for association rules are support and confidence.Two thresholds for finding user specific interesting items are minimal support and minimal confidence respectively.

Support(S) is the percentage of records that contains $X \cup Y$ to the total number of records in the transaction database. X and Y are the sets of items called itemsets and $X \cap Y = \theta$.Confidence(C) is defined as the percentage of the number of transactions that contain $X \cup Y$to the total number of records that contain X. If the percentage exceeds the threshold of confidence, $X \longrightarrow Y$ is obtained.

$$Support(XY) = \frac{CountofX \cup Y}{Totalnumberoftransactions} \quad (4)$$

$$Confidence(X|Y) = \frac{Support(XY)}{Support(X)} \quad (5)$$

Generally two phases are involved in association rule mining.Frequent itemset generation and rule generation.

Frequent itemset are used to find all the itemsets that satisfy the minimal support threshold.The most widely used algorithm for frequent pattern mining is Apriori.There exists some drawbacks for Apriori algorithms,it scans the whole database many times and it would increases the time and space complexities.Inorder to solve this drawbacks,experts uses some modified techniques such as hashing technique[7],sampling approach[8],dynamic itemset counting[9] are used.Apriori algorithm faces two bottlenecks.First one is they generate huge number of candidate itemsets.Second one is,they can scan whole database many times.Inorder to overcome these,Frequent Pattern tree(FP tree)is designed.It only passes the database twice by generating frequent patterns without candidate generation process.

Rule Generation is used to extract those itemsets with larger confidence than user-specified minimal confidence threshold,and these rules are called strong rules.

In order to compress large amount of data into useful and manageable form,topic modelling is used.Latent Semantic Analysis (LSA)[10] uses a singular value decomposition of a collection,forming a reduced linear subspace.Another step to this concept is Probabilistic Semantic model,which is a generative data model.Almost all models that can be used are statistical mixture models,in which each word in a document form a mixture model,where the mixture components are multi-national random variables that can be viewed as a representation of topics. Topic modelling techniques can be generally divided into two categories,supervised and unsupervised;where bag of words and sequence words approaches are used respectively.In the field of Information retrieval ,Document clustering and Summarization[11],uses an unsupervised bag of word technique [12], due to its simplicity.Whereas in the case of supervised models are used in supervised manner,using a preassigned labels for training set.

Latent Dirichlet Allocation(LDA)[13],[14] algorithm is a well known algorithm used in the field of topic modelling.LDA is a technique that automatically finds topics in a collection of documents.In LDA,each document can be viewed as a mixture of topics which splits each word with certain probabilities.Consider a set of of documents and each document has a fixed number of topics to be discovered and let it be K.

Yang Gao ,Yue Xu and Yuefung Li,2015,[15], proposes a two-stage model for modelling the documents in a collections.One of the main discriminative feature of this model is, it combines the data mining techniques to statistical topic modelling to generate discriminative and pattern based representations for modelling topics in documents.In the first stage ,it generates word distributions over topics for documents in the collection whereas in the stage second stage ,it uses the topic representations that are generated in the first stage for representing the topics by using term weighting method and pattern mining methods.The pattern based and discriminative term based representations generated in the second stage are more accurate and efficient than the representations generated by typical statistical topic modelling method LDA.Another important feature of this representation is ,patterns carry more structural and inner relationship within each topic.

Hong Cheng and Xifeng Yan, 2007,[16] designed discriminative frequent pattern analysis for effective classification.Here the authors give more importance to ”frequent patterns”.Frequent patterns is a set of items,subsequence,sub-graphs etc.Frequent patterns have the capacity of reflecting strong associations between each items and carries the underlying semantics of the data.They are also potentially useful features for classification.But it also have some disadvantages.Due its limited predictive power,the inclusion of infrequent patterns does not increases the accuracy.By building a connection between pattern frequency and discriminative measures like Fischer score,information gain,here a strategy is developed to set a minimum support in frequent pattern mining for generating more useful patterns.A feature selection algorithm was also proposed for building high quality classifiers.The two drawbacks of this approach is the scalability issue and over-fitting issue(Features are not representative).The new feature selection algorithm solves the scalability issue and facilitate the pattern generation.

Roberto J and Bayardo Jr, 1998,[17] proposed a pattern mining algorithm which scales linearly in the number of maximal patterns embedded in a database irrespective of the length of the longest pattern.Apriori like algorithms are used for finding frequent patterns in a database.In apriori algorithms generally uses a bottom up search mechanism and it involves a phase for finding frequent patterns.A frequent itemset is a set of items appearing together in a database records meeting a user specified threshold.There are several disadvantages are there for this apriori like algorithms.Most important among them are it restricts apriori like algorithms to discovering only short patterns. Here they uses a new algorithm called Max-Miner algorithm.Look ahead approach is used instead of bottom up search. and it can also prune all its subset from the consideration.By using the Max-Miner algorithm it can efficiently and effectively extracting only the maximal frequent itemsets. The primary task behind every data mining operation is to find the patterns in a database.The current techniques in the field of data mining leaves data outside the mold becomes unexplorable.

Ning Zhong, Yuefeng Li and Sheng-Tang Wu,2012,[18]proposed a technique for effective pattern discovery for data mining.Almost all data mining techniques have been used for finding or extracting useful patterns in a text document,and most of these techniques are adopted from the concept of term based approach,they all suffer from the problems of polysemy and synonymy.Pattern based approaches can well peform than term-based approaches.The most prominent reasons for not using phrases are,they have inferior statistical properties compared with respect to terms,low frequency of occurrence and there exist large amount of redundant and noisy phrases among them.In this effective pattern discovery technique,first calculates the specificity of the discovered patterns and then calculates term weights according to the distribution of terms in the discovered patterns and thus solving the misinterpretation problem.This approach refine the discovered patterns by means of two methods called pattern deploying and pattern evolving.And this method shows an improvement in accuracy of evaluating term weights because discovered patterns are much more specific than the whole documents.

Yves Bastide,Rafik Taouil and Nicolas Pasquier,2000,[19] presents an algorithm called Pascals algorithm which is an optimization of the apriori algorithm.The optimization was performed on a strategy known as pattern counting inference using the key patterns.Key patterns are a subset of equivalence class and it reduces the number of patterns counted in each database pass.A key pattern is defined

as "minimal pattern of an equivalence class gathering all patterns that have the same objects.The supports of some frequent and infrequent patterns can be determined with the pattern counting inference.The support of all other frequent patterns are derived from the frequent key pattern.Key patterns have a property which is compatible with the apriori,hence the optimization is valid.

Hanna M Wallach, 2006,[20] presents a new concept known as Beyond Bag-of-Words.Normally uses bag of words for topic modelling.But it have some disadvantages. The most important among them are it ignores the word order. Generative topic models are generally of two categories,Bigram language models and N-gram topic models.N-gram models does not considers the word order,while bigram models consider pairs of words with the leading word defining a context.Bigram language models uses Hierarchical Dirchlet Language Models whereas N-Gram topic models uses Latent Dirchlet Allocation.It creates a model which considers both topics and word order.It uses a simple extension of LDA algorithm.The bigram topic model shows improved performance compared to both the bi-gram language model and LDA.It is more feasible to consider word level models when the word order is not ignored.

4. PARAMETERS USED FOR EXPERIMENTAL EVALUATION

In order to evaluate the efficiency of the topic modelling techniques, some standard parameters are there. They are Precision,Recall and Accuracy.Precision is defined as the ratio between number of relevant documents retrieved to the total number of retrieved documents.Whereas Recall can be defined as the ratio between retrieved relevant documents and Accuracy can be calculated as relevant document retrieved in top T returns divided by T. The formulas for calculating these evaluation parameters are:

$$Precision = \frac{Numberofretrievedrelevantdocuments}{Totalnumberofretrieveddocuments} \quad (6)$$

$$Recall = \frac{Numberofretievedrelevantdocuments}{Totalnumberofrelevantdocuments} \quad (7)$$

$$Accuracy = \frac{Relevantdocumentsretrievedintop T}{T} \quad (8)$$

Perplexity is another evaluation criteria used for evaluating Topic models and it can be calculated by estimating the probability of words in held out /test data based on training data.

5. CONCLUSION

Pattern mining is the process of mining patterns from a collection of document on the basis of some criteria which are discussed above.In this paper,concepts associated with topic modelling are reviewed that categorize different approaches in this area.The literature review explore the recent trends in topic modelling that comes from the novice procedures ,where data mining is used to classify.Almost all the techniques found for pattern mining have discussed here.The study of pattern mining techniques provides a productive region for further research Around 10 papers have been discussed here. There exists some other techniques similar with those described in this paper, the discussion of which has been not included here as it will be a large corpus.But it is expected that this paper is useful for researchers who are working with pattern mining techniques.Anyone can also get direction for better perception of the diversified sorts of abstraction, which will help to construct new procedures for next generation.

6. REFERENCES

[1] Ma, Bing Liu Wynne Hsu Yiming. "Integrating classification and association rule mining." Proceedings of the fourth international conference on knowledge discovery and data mining. 1998.

[2] Borgelt, Christian. "Frequent item set mining." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2.6 (2012): 437-456.

[3] Masseglia, Florent, Maguelonne Teisseire, and Pascal Poncelet. "Sequential Pattern Mining." (2009): 1800-1805.

[4] Cerf, Loc, et al. "Data Peeler: Contraint-Based Closed Pattern Mining in n-ary Relations." SDM. Vol. 8. 2008

[5] Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4):77 84.

[6] Cavnar et al,Furnkranz," N-gram-based text categorization" Ann Arbor MI, 48113(2):161175.1998

[7] Park, J. S., Chen, M.-S., and Yu, P. S. (1995). An eective hash-based algorithm for mining association rules, volume 24. ACM.

[8] Savasere, A., Omiecinski, E. R., and Navathe, S. B. (1995). An ecient algorithm for mining association rules in large databases.

[9] Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In ACM SIGMOD Record, volume 26, pages 255264. ACM

[10] Yang Gao ,Yue Xu and Yuefung Li, "Pattern based Topics for Document Modelling in Information Filtering" IEEE Transactions on Knowledge Engineering and Data Mining,Vol 27,No:6,June 2015.

[11] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, Discriminative frequent pattern analysis for effective classication, in Proc. IEEE 23rd Int. Conf. Data Eng., 2007, pp. 716725.

[12] R. J. Bayardo Jr, Efciently mining long patterns from databases, in Proc. ACM Sigmod Record, 1998, vol. 27, no. 2, pp. 8593.

[13] N. Zhong, Y. Li, and S.-T. Wu, Effective pattern discovery for text mining, IEEE Trans. Knowl. Data Eng., vol. 24, no. 1, pp. 3044, Jan. 2012.

[14] Dumais, Susan T. "Latent semantic analysis." Annual review of information science and technology 38.1 (2004): 188-230.

[15] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal, Mining frequent patterns with counting inference, ACM SIGKDD Explorations Newslett., vol. 2, no. 2, pp. 6675, 2000.

[16] Ball, Geoffrey H., and David J. Hall. "A clustering technique for summarizing multivariate data." Behavioral science 12.2 (1967): 153-155.

[17] H. M. Wallach, Topic modeling: Beyond bag-of-words, in Proc. 23rd Int. Conf. Mach. Learn., 2006, pp. 977984.

[18] Purver, Matthew, et al. "Unsupervised topic modelling for multi-party spoken discourse." Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006.

[19] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.

[20] Hoffman, Matthew, Francis R. Bach, and David M. Blei. "Online learning for latent dirichlet allocation." advances in neural information processing systems. 2010.