VOL. 10, NO. 6, APRIL 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

IDENTIFICATION OF RELEVANT DOCUMENTS CONSIDERING UNLABELLED DOCUMENTS

Subin. V. B. and Sivaranjani N.
Department of Information Technology, SRM University, Chennai, India
E-Mail: Subinzubeen309@gmail.com

ABSTRACT

Active learning tackles data scarcity problem by choosing unlabelled data for labeling and training. Active learning handles large volume of data selection. Data are diverse in character or wide range. There is a problem of handling unlabelled data and certain predefined category. This can be overcome by developing a method which is flexible to handle large volume (diverse in content) are learned through single platform or group of item rather than individually. Performance analysis using data mining approaches validates accuracy and F measure, combines precision and recall and takes data relevant to query that are successfully retrieved and efficiency of active learning leading to reliable and authentic predictions.

Keywords: data mining, classification.

1. INTRODUCTION

Data mining is a process of analyzing data from different perspective and summarized in to useful information. It is a computer assisted digging, KDD (Knowledge discovery in Database) extracted meaning of data. Overall goal of data mining is to extract information from a data set and transform in to understandable form. Data set is like having labeled and unlabelled documents. Labeled data have meaning name or tag is somehow informative to and desirable to know. Unlabelled data are having useful information but they will be stored in a name that is not related to document or domain. Active learning through many domains for data selection includes classification (predefined category) and information extraction. 85% of business information like letters, surveys, and emails are in unstructured or unlabelled form. Technically data mining is the process of finding correlations or patterns among dozens of fields in relational databases.

Document classification, grouping unlabelled text documents into meaning classifier, is of substantial interest in many applications. Our assumption, taken by traditional document classifier approaches the number of classification K is known before the process of document classification. K is regarded as predefined parameter determined by users. However in reality determining the appropriate value of K is a difficult problem. First given a set of documents, users have to browse whole document collection in order to estimate K. This is not only time consuming but also unrealistic especially when dealing with large document data sets. Furthermore, an improper estimation of K might easily mislead the classification process. Classifier accuracy degrades drastically if a bigger or smaller number of classifier is used. Therefore, it is very useful if a document classifier approach could be designed relaxing the assumption of the predefined K.

2. RELATED WORK

D.D. Lewis *et al.* [2]:

Text classification from labeled and unlabelled document using expectation maximization explains learning accurate text classifiers from limited number of labeled documents by using unlabelled examples to augment available labeled documents. Use of unlabelled document reduces classification error. The algorithm trains a classifier using the available labeled documents, and probabilistically labels the unlabeled documents. It then trains a new classifier using the labels for all the documents, and iterates to convergence. This basic EM procedure works well when the data conform to the generative assumptions of the model. Drawback can be said with training labeled document is expensive with large quantity of unlabelled documents are readily available.

A. Culotta *et at.* [3]:

Gibbs sampling method for stick breaking priors explains rich class of random probable class of labeled documents and it is a simple prediction rule. A Rich and extensible class of random probability measures, which project call stick breaking priors, can be constructed using sequence of independent beta random variables. Examples of random measures that have this characterization include the dirichlet process, its two parameter extension, the two parameter Poisson-Dirichlet process, and beta two-parameter processes. The rich nature of stick breaking priors offers Bayesian, a useful class of priors for non parametric problems, while the similar construction used in each prior can be exploited to develop a general computational procedure formatting them. Limitation lies in sampling method of selecting rich class of random probable class of labeled documents.

M. Lindenbaum *et at.* [5]:

Modeling word burstiness using dirichlet model allows capturing word burstiness words in documents appears once more likely to appear again. Multinomial distributions are often used to model text documents. However they do not capture well the phenomenon that words in a document tend to appear in bursts, if a word appears once it is more likely to appear again. Dirichlet

compound multinomial model as an alternative to the multinomial. Drawbacks exist in capturing word burstiness.

S. Tong *et at.* [6]:

Classification of documents with an exponential family approximation of the dirichlet compound multinomial distribution derives a new family of distributions that are approximations to DCM distributions and constitute an exponential family, unlike DCM distributions. System use these so called EDCM distributions to obtain insights in to the properties of DCM distributions and then derive an algorithm for EDCM maximum likelihood training that is many times faster than corresponding method for DCM distributions. Drawbacks exist in faster training of word burstiness.

B. Settles *et at.* [7]:

Classification by deterministic annealing and wishart based distance measures for fully polarimetric SAR (synthetic aperture radar) data can be characterized by distribution of dictionary entries that match its contents. Deterministic annealing explains recombining of classified data. The goal of classifier is to find and represent this dissimilar groups of similar elements or in other words to subdivide the data space in to number of partitions .Biggest drawback exist in grouping data which are in isolation

H. S. Seung *et at.* [9]:

Dirichlet process prior in Bayesian Non parametric inference with partial exchange ability considers Bayesian non parametric inference for continuous valued partially exchangeable data, when the partition of the observations in to the group is unknown. When the observed data are all distinct the effect on Bayes factors is to favor more groups. In a hierarchical model with a dirichlet process as the second stage priors, prior can also have a large effect on inference but in the opposite direction towards more unbalanced partitions. Drawbacks exist in calculating probability distribution of documents and recombining classified data.

3. PROPOSED SYSTEM

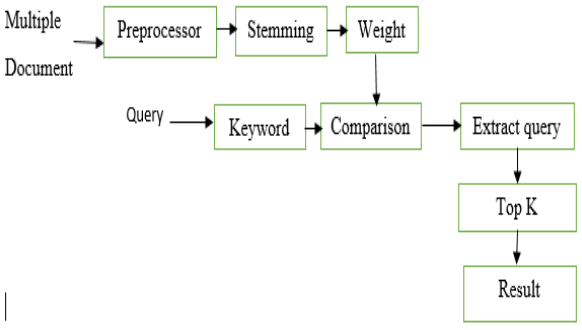

Figure-1. System architecture.

Proposed system develops an automated system for labeled and unlabelled documents. Control system (software) technique of making a process. Proposed system implement search based on keyword rather than documents name. Here we apply stemming algorithm for Root word extraction, Based on scoring algorithm documents are principally categorized in corresponding classification.

In server, data will be uploaded in the early stage of preprocessing stopword removal occurs and next stage stemming starts with extraction of root words and then calculate weight of document by TF-idf (Frequency-inverse document frequency)

TF (t) = (Number of times term t appears in a document) / Total number of terms in the document). TF (Term frequency)

TF-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to document in a collection or corpus. Term frequency which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in along document than shorter ones. Inverse Document Frequency, which measures how important a term is. While computing TF (Term Frequency), all terms are considered equally important. However it is known that certain terms, such as “as”, ”of”, ”that”, may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones by computing

IDF (t) =log_e (Total number of documents /Number of documents with term t in it).

Consider an example with a document containing 100 words where word cat appears 3 times term frequency for cat is then (3/100) = 0.03. Now assume we have 10 million documents and the word cat appears in one thousand of these. Then, the inverse document frequency (ie., idf) is calculated as log(10,000,000/1,000) = 4. Thus, the Tf-idf weight is the product of these quantities: 0.03*4 = 0.12.

Compare weight and keyword of query and extract query keyword. Then apply top K query algorithm to find best possible results. Top K query works with taking keyword and compare with other documents ie doc1, doc1 …docn. And then rank the document. From that top 10 results are taken out and final output is obtained. Top K query algorithm works with taking out the query and selecting keyword and comparing with other documents present in the server and providing weight for doc1=weight, doc1=weight like that up to docn. Rank the documents according to weight calculated using tfid. Suppose we want 10 best results, Top 10 results are obtained.

Stemming is morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of IR applications. For this reason, a number of so-called stemming Algorithms, or stemmers, have been developed, which attempt to reduce a word to its stem or root form. Thus, the key terms of a query or document are represented by stems rather than by the original words.

This not only means that different variants of a term can be conflated to a single representative form - it also reduces the dictionary size, that is, the number of distinct terms needed for representing a set of documents. A smaller dictionary size results in a saving of storage space and processing time. Algorithms for stemming have been studied in computer science since the 1960s. Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation.

Ranking algorithms With the growing number of Web pages and users on the Web, the number of queries submitted to the search engines are also growing rapidly day by day. Therefore, the search engines needs to be more efficient in its processing way and output. Web mining techniques are employed by the search engines to extract relevant documents from the web database documents and provide the necessary and required information to the users. The search engines become very successful and popular if they use efficient ranking mechanisms. It is a numeric value that represents the importance of a page present on the web.

4. PERFORMANCE

Performance testing to determine system performs in terms of responsiveness and stability under a particular workload. System performs well in web environment. Performance analysis approaches validates accuracy and efficiency of active learning leading to reliable and authentic predictions. Performance measures by f-measure, a measure that combines precision and recall ie fraction of retrieved documents and relevant documents. Precision takes all retrieved documents and recall takes data relevant to query that are successfully retrieved.

5. EXPERIMENTAL OUTCOME

Table-1. Characteristics of data sets used in the experiments.

Data sets	Size	#Features	#Classes	#Iterations	Window size
isolet	7797	617	26	200	25
letter	20000	16	26	500	28
madelon	4400	500	2	364	5
magic	19020	10	2	1000	13
mfeat-factors	2000	216	10	280	5
mfeat-fourier	2000	76	10	280	5
mfeat-karhunen	2000	64	10	280	5
mfeat-pixel	2000	240	10	280	5
mfeat-zernike	2000	47	10	280	5
mushroom	8124	22	2	1000	5
optdigits	5620	64	10	786	5
page	5473	10	5	766	5
pendigits	10992	16	10	1000	7
segment	2310	19	7	323	5

6. CONCLUSION AND FUTURE ENHANCEMENT

Project implements document classification for both labeled and unlabelled documents so that proposed system retrieve more results for given query. Also implementation of stemming algorithm and top k query algorithm for securing documents and display best matched results respectively.

Project helps us to sweep through databases and identifies hidden unlabelled documents in one step i.e. automated discovery of unlabelled patterns. Active learning automates process of finding predictive information from large database. Predictive analysis is proposed systems strength.

REFERENCES

[1] Z. Lu, X. Wu, and J. Bongard, “Active learning with adaptive heterogeneous ensembles,” In: Proceedings of the 9th IEEE International Conference on Data Mining (ICDM). 2009, pp. 327-336.

[2] D. D. Lewis and W. A. Gale, “A sequential algorithm for training text classifiers,” In: Proceedings of Research and Development in Information Retrieval. 1994, pp. 3-12.

[3] A. Culotta and A. McCallum, “Reducing labeling effort for structured prediction tasks,” In: Proceedings

of the National Conference on Artificial Intelligence (AAAI). 2005, pp. 746-751.

[4] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," In: Proceedings of the International Conference on Machine Learning (ICML). 1994, pp. 148-156.

[5] M. Lindenbaum, S. Markovitch, and D. Rusakov, "Selective sampling for nearest neighbor classifiers," Machine Learning. Vol. 54(2), pp. 125-152, 2004.

[6] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," Journal of Machine Learning Research. vol. 2, pp. 45-66, 2001.

[7] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," In: Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP). 2008, pp. 1069-1078.

[8] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," In: Proceedings of the International Conference on Machine Learning (ICML). 2004, pp. 79-86.

[9] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," In: Proceedings of the Fifth Workshop on Computational Learning Theory. 1992, pp. 287-294.

[10] L. Breiman, "Bagging predictors," Machine Learning. Vol. 24(2), pp. 123-140, 1996.