Efficient Classification of Text and Improving Learning Experience

Ms. Raja Saranya Kumari
Vel tech high tech Dr.RR Dr.SR Engineering College

Ms. R. Divya Lakshmi
Vel tech high tech Dr.RR Dr.SR Engineering College

Ms. R. Monica
Vel tech high tech Dr.RR Dr.SR Engineering College

Ms. B. Monisha
Vel tech high tech Dr.RR Dr.SR Engineering College

Abstract

Assuring relevant features in texts cannot be guaranteed, earlier techniques also suffers from problems. In the existing system those are solved using new methods which recognizes, categorizes and updates. In proposed system, Student performance will be predicted using existing text mining approaches and additionally added approaches. Allocation of questions and the evaluation knowledge can be fed into the system and can be used in an effective way. According to the performance of the student, learning materials are provided using different approach and also facility for clarifying doubts can also be done, thereby making the students to improve the performance level of them and can also be able to understand the concepts in an easier way.
Keywords: Rocchio classifier, Education, Data mining, Text Mining

I. INTRODUCTION

THE Data mining is the process of extracting patterns from data. As more data are gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform these data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery. Data mining commonly involves four classes of task: Classification - Arranges the data into predefined groups. For example an email program might attempt to classify an email as legitimate or spam. Common algorithms include nearest neighbor, Naive Bayes classifier and neural network. Clustering - Is like classification but the groups are not predefined, so the algorithm will try to group similar items together. Regression - Attempts to find a function which models the data with the least error. A common method is to use Genetic Programming. Association rule learning - Searches for relationships between variables. For example a supermarket might gather data of what each customer buys. Using association rule learning, the supermarket can work out what products are frequently bought together, which is useful for marketing purposes. This is sometimes referred to as "market basket analysis". Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' of the assignment of the higher text in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities). Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods. A typical application is to scan a set of documents written in a natural language and either model the document set or predictive classification purposes or populate a database or search index with the information extracted. National Security/Intelligence Scientific discovery, especially Life Sciences Sentiment Analysis Tools, Listening Platforms Natural Language/Semantic Toolkit or Service Publishing Automated ad placement Search/Information Access Social media monitoring Security applications Many text mining software packages are market-ed for security applications, especially monitoring and analysis of online plain text sources such as Internet news, blogs, etc. for national security purposes. It is also involved in the study of text encryption/decryption. Biomedical applications. A range of text mining applications in the biomedical literature has been described. One online text mining application in the biomedical literature is PubGene that combines biomedical text mining with network visualization as an Internet service. TPX is a concept assisted search and navigation tool for biomedical literature analyses - it runs on PubMed/PMC and can be configured, on request, to run on local literature repositories too. GoPubMed is a knowledge based search engine for biomedical texts. Software applications. Text mining methods and software is also being researched and developed by major firms, including IBM and Microsoft, to further automate the mining and analysis processes, and by different firms working in the area of search and indexing in general as a way to improve their results. Within public sector much effort has been concentrated on creating software for tracking and

monitoring terrorist activities. Online media applications. Text mining is being used by large media companies, such as the Tribune Company, to clarify information and to provide readers with greater search experiences, which in turn increases site "stickiness" and revenue. Additionally, on the back end, editors are benefiting by being able to share, associate and package news across properties, significantly increasing opportunities to monetize content. Marketing applications. Text mining is starting to be used in marketing as well, more specifically in analytical customer relationship management. Sentiment analysis may involve analysis of movie Reviews for estimating how favorable a review is for a movie. Such an analysis may need a labeled data set or labeling of the affectivity of words. Resources for affectivity of words and concepts have been made for WordNet and ConceptNet, respectively. Text has been used to detect emotions in the related area of affective computing. Text based approaches to affective computing have been used on multiple corpora such as students evaluations, children stories and news stories. Academic applications .The issue of text mining is of importance to publishers who hold large databases of information needing indexing for retrieval. This is especially true in scientific disciplines, in which highly specific information is often contained within written text. Therefore, initiatives have been taken such as Nature's proposal for an Open Text Mining Interface (OTMI) and the National Institutes of Health's common Journal Publishing Document Type Definition (DTD) that would provide semantic cues to machines to answer specific queries contained within text without removing publisher barriers to public access. Academic institutions have also become involved in the text mining initiative: The National Centre for Text Mining (NaCTeM), is the first publicly funded text mining centre in the world. NaCTeM is operated by the University of Manchester in close collaboration with the Tsujii Lab University of Tokyo. NaCTeM provides customized tools, research facilities and offers advice to the academic community. They are funded by the Joint Information Systems Committee (JISC) and two of the UK Research Councils (EPSRC & BBSRC). biological and biomedical sciences, research has since expanded into the areas of social sciences.

II. RELATED WORK

Over the years, a variety of algorithms for finding frequent sequential patterns in very large sequential databases have been developed. The key feature in most of these algorithms is that they use a constant support constraint to control the inherently exponential complexity of the problem. In general, patterns that is the best while working with it and getting through them it contain only a few items will tend to be interesting if they have a high support, whereas long patterns can still be interesting even if their support is relatively small. Ideally, we desire to have an algorithm that finds all the frequent patterns whose support decreases as a function of their length. In this paper we present an algorithm called SLPMiner, that finds all sequential patterns that satisfy a length-decreasing support constraint. SLPMiner combines an efficient database-projection-based approach for sequential pattern discovery with three effective database pruning methods that dramatically reduce the search space. Our experimental evaluation shows that SLPMiner, by effectively exploiting the length-decreasing support constraint, is up to two orders of magnitude faster, and its runtime increases gradually as the average length of the sequences (and the discovered frequent patterns) increases. Text mining is the technique that helps users find useful information from a large amount of digital text documents on the Web or databases. Instead of the keyword-based approach which is typically used in this field, the patternbased model containing frequent sequential patterns is employed to perform the same concept of tasks. However, how to effectively use these discovered patterns is still a big challenge. In this study, we propose two approaches based on the use of pattern deploying strategies. The performance of the pattern deploying algorithms for text mining is investigated on the Reuters dataset RCV1 and the results show that the effectiveness is improved by using our proposed pattern refinement approaches. In the literature of feature selection, different criteria have been developed to evaluate the goodness of features. In our investigation, we notice that a number of existing selection criteria implicitly select features that preserve sample similarity, and can be unified under a common framework. We further point out that any feature selection criteria covered by this framework cannot handle redundant features, a common drawback of these criteria. Motivated by these observations, we propose a new "Similarity Preserving Feature Selection" framework in an explicit and rigorous way. We show, through theoretical analysis, that the proposed framework not only encompasses many widely used feature selection criteria, but also naturally overcomes their common weakness in handling feature redundancy. In developing this new framework, we begin with a conventional combinatorial optimization formulation for similarity preserving feature selection, then extend it with a sparse multiple-output regression formulation to improve its efficiency and effectiveness. A set of three algorithms are devised to efficiently solve the proposed formulations, each of which has its own advantages in terms of computational complexity and selection performance. Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis. This paper presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves encouraging performance. In this survey we review work in machine learning on methods for handling data sets containing large amounts of irrelevant information We focus on two key issues the problem of selecting relevant features and the problem of selecting relevant examples We describe the advances that have been

made on these topics in both empirical and theoretical work in machine learning and we present a general framework that we use to compare different methods We close with some challenges for future work in this area.

III. DEFINITIONS

To learn term features within only relevant document and unlabeled documents, paper used two term-based models. In the first stage, it utilized a Rocchio classifier to extract a set of reliable irrelevant documents from the unlabeled set. In the second stage, it built a SVM classifier to classify text documents. A two-stage model was also proposed in which proved that the integration of thorough analysis (a term-based model) and pattern mining is the best way to design a two-stage model for information filtering systems.

A. Natural Language Processing

Natural language analysis based on semantic grammar is bit similar to syntactically driven parsing except that in semantic grammar the categories used are defined semantically and syntactically.

B. Word Net technique

Word net tool is like dictionary. The process is giving the meaning of nouns, verbs in the content of files. The efficiency of search is improved by word net processing on user queries.

C. Project and Batch Allocation

In this module coordinator upload project base paper on behalf of each and every student, and also allocate batches for all projects. Batches were created by coordinator by selecting number of student in batch and student ID's. In upload process the base paper will be checked for duplication with previous batches in title level as well as in content level. The content level checking is done by stripping down the pdf contents to text contents. If the base paper is not duplicated server accepts the upload and updates the student record.

D. Text Mining for Assessment

Teacher prepares questions and answers for student assessment. Text mining process is done by natural language processing and word net tools. Pos tagger is implemented to extract the important keywords in the answer given by staff before assessment is done. The extracted Keywords are categorized mandatory keywords, subordinate keywords, and technical keywords. Wordnet tool is used to give the related synonyms to literal word in the subordinate terms. Now Teachers can feed the servers with the eligible terms in the categories to be present for student evaluation.

E. Project Review and Student Assessment

Student login with his credentials and then uploads the review materials in server. Reviewer gives the review marks for each student based on performance. Here we allotted three reviews, and give marks for student based on review performance. Student can write the assessment test and can submit the answers to the server. Student answers are evaluated later in server by extracting keywords using NLP technique and wordnet tool. The Machine will evaluate the answers by comparing it with the categorized terms given by the teachers. Depends upon the student answer they will give marks and prepare performance report. Review performance and assessment score are aggregated to find the overall performance.

F. Material Recommendation and Interactive Student Learning:

Teachers prepare the material for each subject and also give tags (good, best) for student material recommendation. Here we upload the materials like video, text, pdf. Video transcoding is applied while video materials are uploaded for below average students. After finishing the assessment test, in student portal they get the materials based on overall performance calculated by server. If they have doubt while watching video content, students can interactively raise questions by simply clicking on the video frame. The video frames are previous indexed so that appropriate Meta information's can be extracted for each frame. The student's questions and Meta information from the current frame are sending to server and can be reviewed by the staff. Once the staff login they will be notified with the questions and then staffs can reply to the question. The Student can now be able to view the answers given by the staffs.

IV. APPROACHES

We proposed a material recommendation system based on student performance which will be predicted from the project work documents and the online examination results. The online tests are conducted for each individual subjects and the questions will be allocated for the same by the respective staff in their portal. The staffs will feed the evaluation knowledge to the system for student examination score generation. The aggregate of project work and online test decides the student's overall performance in three levels which is used to categorize them. The study materials will be rendered to the particular student depending upon the

performance category he belongs. Videos can also be played in the student portal for effective learning which we innovate to interactive learning by simple click events on the doubtful content of the video.

A. String Matching Algorithm

1) Input: one string
2) Read the contents from database
3) Check the character of two string
4) Compare character by character in first file to character by character in second String.
5) If matches found, title will be same.
6) If no matches found, title is unique.

In upload process the base paper will be checked for duplication with previous batches in title level as well as in content level

B. Content Based Algorithm

1) Input :PDF file(base paper uploaded)
2) Read the files using pdf parser technique.
3) Pdf parser technique the process of analyzing a string of symbols, either in natural language or in computer languages, conforming to the rules of a formal grammar.
4) Let x=Read the abstract lines.
5) Check the already uploaded pdf files
6) Let y=Then that file is to be parsed using pdf technique
7) Compare x and y.
8) If x is equal to y.
9) Store it in Vector.
10) Check and compare line by line.
11) If every line is equal then file is same

The content level checking is done by stripping down the pdf contents to text contents. If the base paper is not duplicated server accepts the upload and updates the student record.

C. Transcoding Algorithm

1) Input the videos
2) Convert the videos into multiple frame
3) Finding the similar frame(x).
4) x->remove
5) Extract the audio from video using ffmpeg tool
6) Audio is extracted->a
7) X->converted into videos->output(Y)
8) Y->embedded into audio
9) It gives the minimum leghth of video compare to original videos

Video transcoding is applied while video materials are uploaded for below average students. After finishing the assessment test, in student portal they get the materials based on overall performance calculated by server.

V. EVALUATION

This section discusses the testing environment, and reports the experimental results and the discussions. It also provides recommendations for offender selection and the use of specific terms and generl terms for describing user information needs. The proposed model is a supervised approach that needs a training set including both relevant documents and irrelevant documents.

A. Baseline Models and Setting

One-to-many data linkage is an essential task in many domains, yet only a handful of prior publications have addressed this issue. Furthermore, while traditionally data linkage is performed among entities of the same type, it is extremely necessary to develop linkage techniques that link between matching entities of different types as well. In this paper we propose a new one-to-many data linkage method that links between entities of different natures. The proposed method is based on a one-class clustering tree (OCCT) which characterizes the entities that should be linked together. The tree is built such that it is easy to understand and transform into association rules, i.e., the inner nodes consist only of features describing the first set of entities, while the leaves of the tree represent features of their matching entities from the second dataset. We propose four splitting criteria and two different pruning methods which can be used for inducing the OCCT. The method was evaluated using datasets from three different domains. The results affirm the effectiveness of the proposed method and show that the OCCT yields better

performance in terms of precision and recall (in most cases it is statistically significant) when compared to a C4.5 decision tree-based linkage method. In this survey we review work in machine learning on methods for handling data sets containing large amounts of irrelevant information We focus on two key issues the problem of selecting relevant features and the problem of selecting relevant examples We describe the advances that have been made on these topics in both empirical and theoretical work in machine learning and we present a general framework that we use to compare di erent methods We close with some challenges for future work in this area.

B. Hypotheses

To learn term features within only relevant document and unlabelled documents, paper used two term-based models. In the first stage, it utilized a Rocchio classifier to extract a set of reliable irrelevant documents from the unlabeled set. In the second stage, it built a SVM classifier to classify text documents. A two-stage model was also proposed in which proved that the integration of thorough analysis (a term-based model) and pattern mining is the best way to design a two-stage model for information filtering systems. Ethernet on the AS/400 supports TCP/IP, Advanced Peer-to-Peer Networking (APPN) and advanced program-to-program communications (APPC). You can connect your AS/400 to an Integrated Services Digital Network (ISDN) for faster, more accurate data transmission. An ISDN is a public or private digital communications network that can support data, fax, image, and other services over the same physical interface. Also, you can use other protocols on ISDN, such as IDLC and X.25.

C. Results

The software may be safety-critical. If so, there are issues associated with its integrity level.The software may not be safety-critical although it forms part of a safety-critical system. For example, software may simply log transactions.If a system must be of a high integrity level and if the software is shown to be of that integrity level, then the hardware must be at least of the same integrity level.There is little point in producing 'perfect' code in some language if hardware and system software (in widest sense) are not reliable. If a computer system is to run software of a high integrity level then that system should not at the same time accommodate software of a lower integrity level.Systems with different requirements for safety levels must be separated.Otherwise, the highest level of integrity required must be applied to all systems in the same environment.

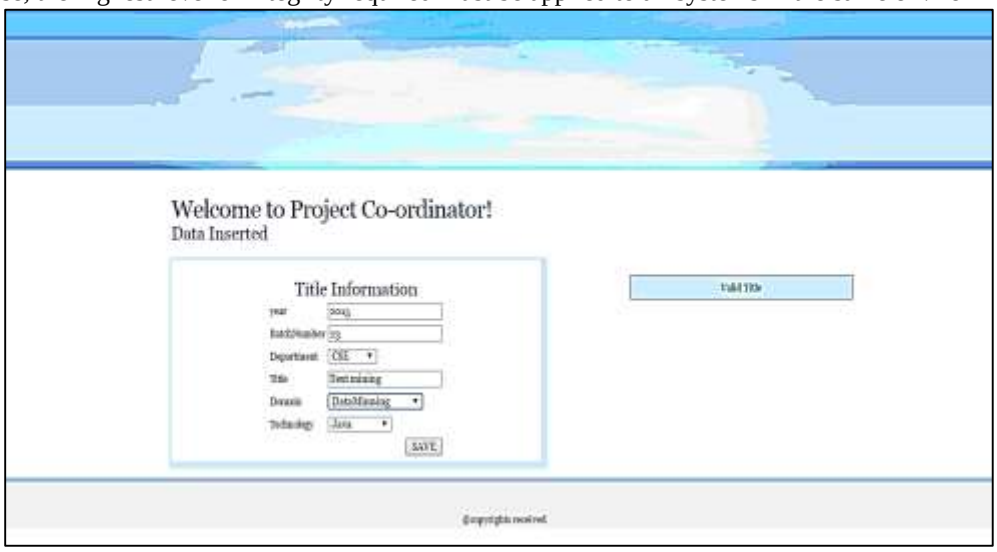

Fig. 1: Message for Title Updation

Batches were created by coordinator by selecting number of student in batch and student ID's. In upload process the base paper will be checked for duplication with previous batches in title level as well as in content level. The content level checking is done by stripping down the pdf contents to text contents. If the base paper is not duplicated server accepts the upload and updates the student record.

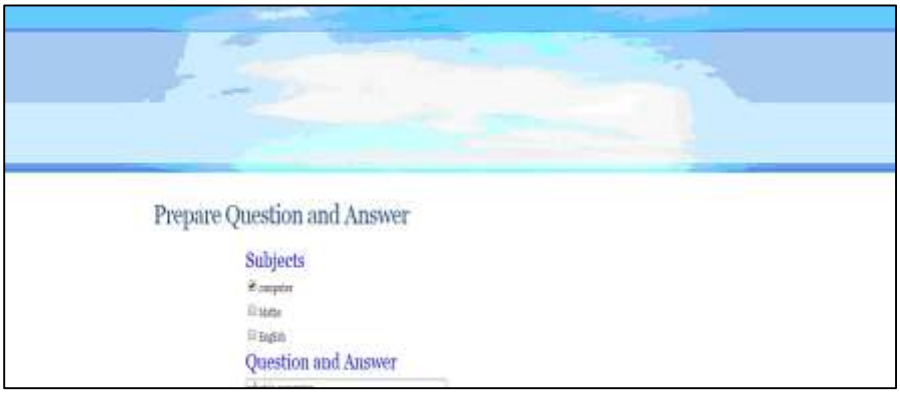

Fig. 2: Updation of Questions by Staff

Wordnet tool is used to give the related synonyms to literal word in the subordinate terms. Now Teachers can feed the servers with the eligible terms in the categories to be present for student evaluation.

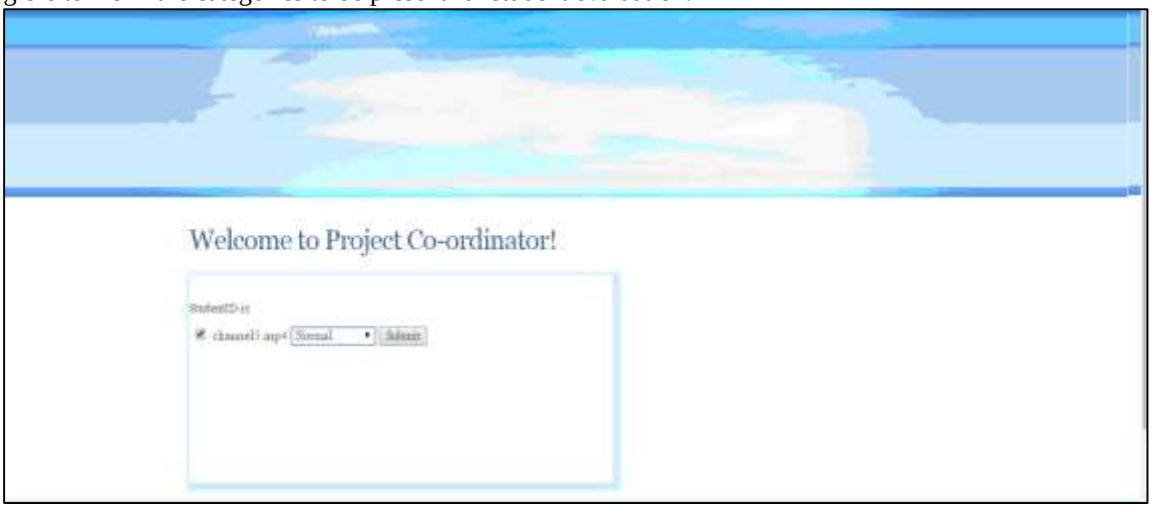

Fig. 3: Choosing Normal/Transcoding

Video transcoding is applied while video materials are uploaded for below average students. After finishing the assessment test, in student portal they get the materials based on overall performance calculated by server.

Fig. 4: Assigning Material

Teachers prepare the material for each subject and also give tags (good, best) for student material recommendation. Here we upload the materials like video, text, pdf. The video frames are previous indexed so that appropriate meta information's can be extracted for each frame.

Fig. 5: Screen after Posting Questions

The student's questions and Meta information from the current frame are send to server and can be reviewed by the staff. Once the staff login they will be notified with the questions and then staffs can reply to the question. The Student can now be able to view the answers given by the staffs.A hierarchical structuring of relations may result in more classes and a more complicated structure to implement. Therefore it is advisable to transform the hierarchical relation structure to a simpler structure such as a classical flat one. It is rather straightforward to transform the developed hierarchical model into a bipartite, flat model, consisting of classes on the one hand and flat relations on the other. Flat relations are preferred at the design level for reasons of simplicity and implementation ease. There is no identity or functionality associated with a flat relation.

VI. CONCLUSION

Hence we developed a method to find and classify terms based on their teacher answer. Student test answer also classifies terms and evaluates answers based on teacher answer, and then teacher will provide material based on student performance. And hence we proposed a material recommendation system based on student performance which is predicted from the project work documents and the online examination results. The online tests were conducted for each individual subjects and the questions ware allocated for the same by the respective staff in their portal. The staffs fed the evaluation knowledge to the system for student examination score generation. Thus aggregate of project work and online test decided the student's overall performance in three levels which is used to categorize them. The study materials are rendered to the particular student depending upon the performance category he belongs. Videos are played in the student portal for effective learning which we innovated to interactive learning by simple click events on the doubtful content of the video. Our Project will be effective in calculating the results of the student and make staff to provide materials and also interactive session. But Our Project ask Students to ask questions from specific topic and does not provide them to ask a specific doubt in that particular topic. This provides advantage for even deeper understanding

REFERENCES

[1] Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana in Relevance Feature Discovery for Text Mining, VOL. 27, NO. 6, JUNE 2015
[2] M. Aghdam, N. Ghasem-Aghaee, and M. Basiri, "Text feature selection using ant colony optimization," in Expert Syst. Appl., vol. 36, pp. 6843–6853, 2009.
[3] A. Algarni and Y. Li, "Mining specific features for acquiring user information needs," in Proc. Pacific Asia Knowl. Discovery Data Mining, 2013, pp. 532–543.
[4] A. Algarni, Y. Li, and Y. Xu, "Selected new training documents to update user profile," in Proc. Int. Conf. Inf. Knowl. Manage., 2010, pp. 799–808.
[5] N. Azam and J. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization," Expert Syst. Appl., vol. 39, no. 5, pp. 4760–4768, 2012.
[6] R. Bekkerman and M. Gavish, "High-precision phrase-based document classification on a modern scale," in Proc. 11th ACM SIGKDD Knowl. Discovery Data Mining, 2011, pp. 231–239.
[7] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," Artif. Intell., vol. 97, nos. 1/2, pp. 245–271, 1997.
[8] C. Buckley, G. Salton, and J. Allan, "The effect of adding relevance information in a relevance feedback environment," in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1994, pp. 292–300.
[9] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson, "Selecting good expansion terms for pseudo-relevance feedback," in Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 243–250.
[10] G. Chandrashekar and F. Sahin, "Asurvey on feature selection methods," in Comput. Electr. Eng., vol. 40, pp. 16–28, 2014.
[11] B. Croft, D. Metzler, and T. Strohman, Search Engines: Information Retrieval in Practice. Reading, MA, USA: Addison-Wesley, 2009.