Classification of documents based on contents using the n-gram method of MNB model

Junaina Jamil Najim Aldin AL-Bayati

Middle East University, Amman, Jordan

Abstract
Nowadays the large number of documents needs to classify by content not only by the name of the document, this research focused on the Arabic documents due to more complication. This study aims to apply this classification technique for files management to raise the level of organization and retrieval of files. In this system of study, number of ways had been utilized to increase the performance of the (MNB) or what called the multinomial naïve bays classification tool, improved the multinomial naïve Bayes model by using the n-gram. Document data was selected as consecutive pairs of keywords which called the bi-gram, or as three consecutive keywords which called the tri-gram, or as four of consecutive keywords which called the 4-gram, by used of the n-grams the classification performance was increased. In this system, we found that the bi-grams were the most efficient process in the MNB model.

Keywords.
Document classification, MNB model, n-gram, training document, testing document, recall and precision

1. Introduction

These days there are a massive number of documents files in the storage units in the computers. Usually the computer users don't know the files content and the classifying task of these documents to the proper subject is extremely a hard task to do especially when the files have a large number and the nearly the same related documents, so when the users need to find the files, the users have to spend a lot of more time to find the files that related to the wanted subject. The computer search method which is available on computers that currently the most use isn't satisfy the users demands, because the computer search depending on the name of the documents not on the documents content, so it isn't efficient to find the wanted files especially if the users don't save it with related name or when they forgot the file name.
The research idea here has the attention on the automatically classification and on enhancement of the classification of the documents to get more good performance in finding the files that concerning of special subjects depends on the content not only on the file name. This research is deal with and interested on the documents of Arabic language, due to the more issues in complication than the documents of English, so these maters must to be taken in mined during the document classification.
There were many classification problems in libraries, so to reduce these problems the document classification is the best solution which is utilized from a years ago, the document classification is a process or way to label the documents to the suitable subject, about this title there are many studies and researches but they are as yet in progress. This research target is to improve the classification of the documents, in order to raise the classification accuracy without have a bad effect on the classification time.
In real life the document classification has a lot of application like the filtering the articles such as filtering emails especially the spam emails for the consumers, also to utilize in the government sector (Goller et al, 2000). In this research the Naive Bayes will be implemented as a classification algorithm, which is improve the accuracy of the classification process, also to have more accuracy and high classification speed we will integrate the Multinomial Naive Bayes or what called the MNB model to the process. Because of the increase in document files that related to the same topic under of more subjects, the necessary of document classification and the management of the files became a must with an effective method which it let the easy to find the wanted files which is related to the wanted subject. Consequently, the automated classification of the files depends on documents of the predefined training has been witnessed a high increase in interest over the previous years. The aim of this study is to apply the technique of classification of the files monitoring and management to have a higher level of organized and high level of retrieval of the flies. This study aims to apply this classification technique for files management to raise the level of organization and retrieval of files.

2. Literature Review

According to (Goller et al, 2000) there are two levels or stages in the document classification especially the automatic one as shown in the figure below; the first stage is the learning stage while the second is the subsequent classification. In the learning stage, the users have to choose the subject as they want or according to the system need, and the choose of documents that related to the

Manuscript received October 5, 2015
Manuscript revised October 20, 2015

wanted subjects. Most of the operations of the automatic classification for the document are require a counterexamples document to all subjects, the counterexamples documents should not refer to such subject. the documents can be related to more of one subject, it could be had various subject in the hierarchy mode, but during the classification stage it's a must and responsible on the classifier to present the rake related to every document for every subject, in the classification stage it should have a high performance level, due to the number of documents which are have to be classified.

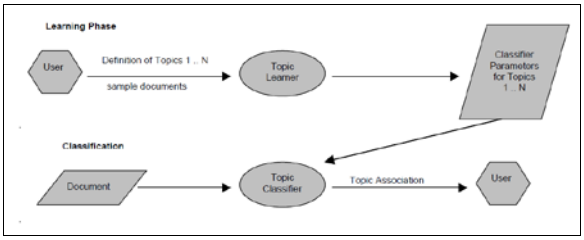

Figure 1: Document classification stages

By looking to (Li and Jain, 1998) they seek on the search engines that accomplished to lowering the potential effort and the time to the employees and to have an easier information retrieval, subsequently to raise the utilization of the World Wide Web.

Most of all the commercial search engines such as the Info seek, HotBot, and Yahoo, etc., rely on document classification in all of its procedures simply like the document retrieval, routing, categorization and the filtering order. The document classification way rely on a predefined group of labels examples that is concerning to two or more than two classes, over and above the classifying of the recent documents to the category that have the greatest similarity. There are some challenges faced the document classification such as the hardness of capturing the advanced semantic of the languages from the document keywords. Nowadays there are several types of document classification methods such as the Naive Bayes classifier, and decision trees.

Referring to (Lam et al, 1999) they have been improved a modern process in the automatic document classification also they have been utilized the automatic classification in the document retrieval. This modern process of classification is derived from a learning sample called the "instance-based learning" also depends on the new document retrieval technique called the "retrieval feedback. The performance of the modern process of classification is too high due to use of set of a two real-world text exported from the MEDLINE database. Moreover they established and achieved a high performance for the document retrieval outputs by utilizing of the manual classification of documents with equal performance level in the automatic one.

According to (Nigam, 2000) conducted a research about the document classification and organized that the number of the labeled or named documents in differentiation with the unnamed documents is too teeny. The manual labeled to unlabeled document is a potential intensive work. Automatic classification is relying on a training document to obtain the keywords and to generate the rules (classes). Because of this, they utilizing an algorithm depend on the both of Expectation-Maximization (EM) and to the Naive Bayes classification which are worked to express the classifier based on the document which is previously labeled.

By attention to (Ramdass and Seshasai ,2009) in their study about the document classification for the newspaper articles, they found that there are different scenarios in the real-lives, they are wished for the operation of the classifying of several documents in different classes, which can be accomplished by utilizing the automatic classification of documents, one of scenarios is the newspaper articles, which have to be categorized to various classes like the sports and news, etc.,.

3. Methodology.

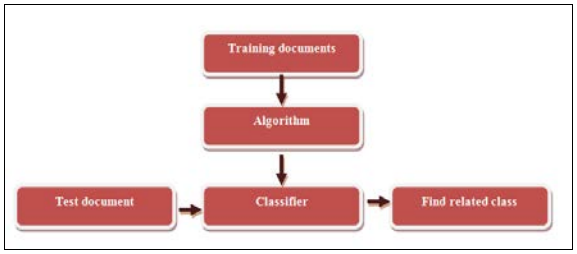

Figure 2: classification model

By looking to these days we can deserve that the Naive Bayes theorem is the most popular classification method of the document which is dealing with documents as a bag words then determine that if a particular keyword exist in a particular document or not exist (Shimodaira, 2014). Finally the methodology of this research is summarized in these points:

1. Use a certain and an open source of the data set contain a number of stages connected with similar filed, also with a huge number of documents, in each class the documents will be separated into two sections , the first is a training documents while the second one is the testing documents.

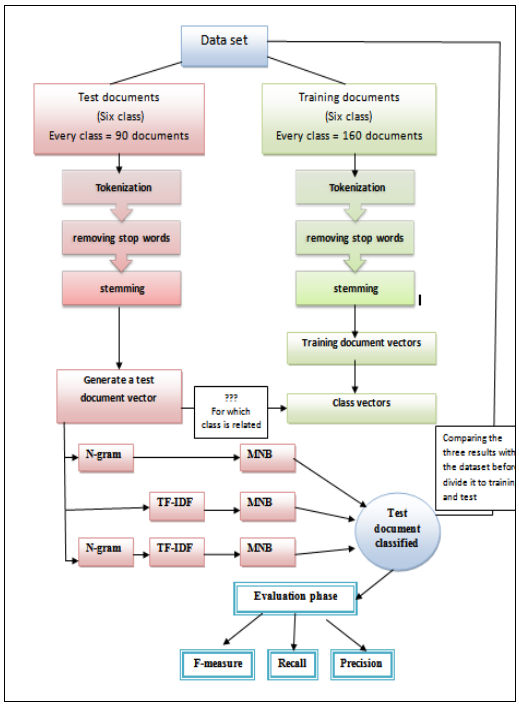

Figure 3: the classifier model building

2. In the GUI operation, the user should do the following:
   - Specify the folder path
   - Select the method of enhancement for the MNB classifier
   - Measure the Recall and precision
3. We have in the system a group of classes, the predefined classes will be defined as $C_n = \{C_1, C_2, C_3, \dots\}$.
4. Contrivance the Multinomial Naive Bayes model and the approaches which improve the classification process by utilizing the c# language programming due to the dissimilar libraries and purposes which may provision the system also to design the GUI by utilizing c# windows system.
5. Add the MNB model to raise the classification performance, enactment by utilizing n-grams: data as consecutive pairs of a keyword which is called (bigrams), or make three consecutive keywords which is called (trigrams), utilizing the n-grams can boost the classification performance.
6. The system will be classified the tested documents then to measure the recall and precision of every case of the enhancement, while the values top-up the perfectness of system increased.

$$P(c\backslash X) = \frac{P(c)P(X\backslash c)}{P(X)}$$

Bays' rule (to obtain the highest probability)

The previous class $P(c)$ may be calculated by the below equation;

$$P(c) = \frac{\textit{the number of document that belongs to class c}}{\textit{the number of the whole documents in the whole classes}}$$

The probability of tested document (X) is calculated by the equation below;

$$P(X\backslash c) = \prod_{i=1}^{m} P(X_i \backslash c) = P(X_1 \backslash c) * P(X_2 \backslash c) * \dots\dots\dots P(X_m \backslash c)$$

Where; the $P(X_i \backslash c)$ is considered as the probability for every feature vector (X_i) the equation into measure it;

$$P(X_i \backslash c) = \frac{\sum(tf(X_i, d \in c)) + \alpha}{\sum N_{d \in c} + \alpha . V}$$

Table 1: example of implement the bi-gram

Smoking, especially smoking a pipe, is a factor of canc incidence gums, tongue, mouth factors							
("التدخينلاسيماتدخينالغليون،هوعاملمنعواملالإصابةبسرطانااللثةواللسانوالفم")							
التدخين لا	سيما تدخين	الغليون هو	منع امل	الإصابةعوامل	بسرطان اللثة	اللساند و	الفم و

There are dissimilar approaches to estimate the performance of classification tool like the recall and precision:

- True positive (TP): true class.
- False positive (FP): incorrect class.
- False negative (FN): the document related to an exact class but not marked.
- True negative (TN); the document not related to an exact class and not marked.

The Recall R_i is distinct by the capability of the classifier to categorize the text document to a class; $R_i = \frac{TP_i}{TP_i + FN_i}$ *100%

The precision(ρ_i) defines the capability of the classifiers to categorize the tested document as existence below the legal class as opposite to all documents categorize in that session, both valid or invalid: $P_i = \frac{TP_i}{TP_i + FP_i}$*100%

Conjoining the precision and the recall measured, to obtain a big picture of the performance with seeing that the recall and precision have similar significance, the below equation show how to predictable it;

$$F = 2\,\frac{P_i * R_i}{P_i + R_i}$$

4. Results and Discussion

The measurement of the recall and precision of all classes have been cleared in this section for the all proposed classifier, at first, the measurements of the Recall and precision for the (2, 3, 4) gram have been assessed so as to select the furthermost efficacy n-gram.

Table 2: the Recall and precision for the MNB classifier with bi-gram classifier

Class name	Precision	Recall	F-measure
Business	85.56%	91.67%	88.50%
Entertainment	82.22%	57.36%	67.57%
Middle_east	84.44%	76.77%	80.42%
Scitech	80.00%	91.14%	85.20%
Sport	87.78%	96.34%	91.86%
World	65.56%	88.06%	75.15%
Average	80.93%	83.56%	81.46%

Table 3: the Recall and precision for the MNB classifier with tri-gram classifier

Class name	Precision	Recall	F-measure
Business	84.44%	73.79%	78.75%
Entertainment	33.33%	56.60%	41.95%
middle_east	81.11%	73.00%	76.84%
Scitech	76.67%	56.10%	64.78%
Sport	91.11%	91.11%	91.11%
World	65.56%	83.10%	73.29%
Average	72.04%	72.28%	71.12%

Table 4: the Recall and precision for the MNB classifier with 4-gram classifier

Class name	Precision	Recall	F-measure
Business	84.44%	73.79%	78.75%
Entertainment	33.33%	55.56%	41.66%
Middle_east	81.11%	73.00%	76.84%
Scitech	76.67%	56.56%	65.09%
Sport	91.11%	91.11%	91.11%
World	65.56%	83.10%	73.29%
Average	72.04%	72.18%	71.12%

The figures below represent the recall and precision also the F-measures for unlike n-gram (bigram, trigram, and 4-gram).

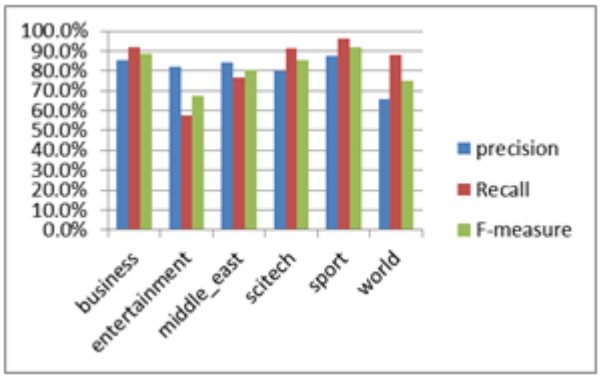

Figure 4: Recall and precision and F-measure for bigram

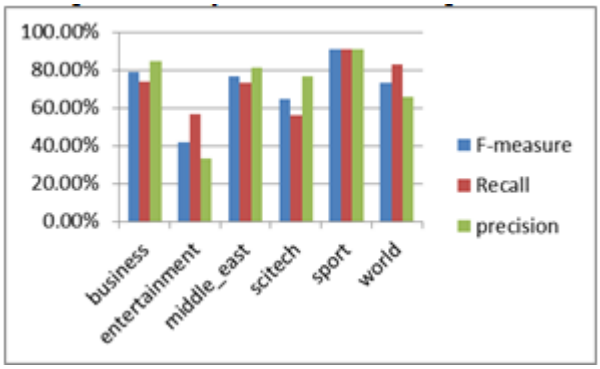

Figure 5: Recall and precision and F-measure for trigram

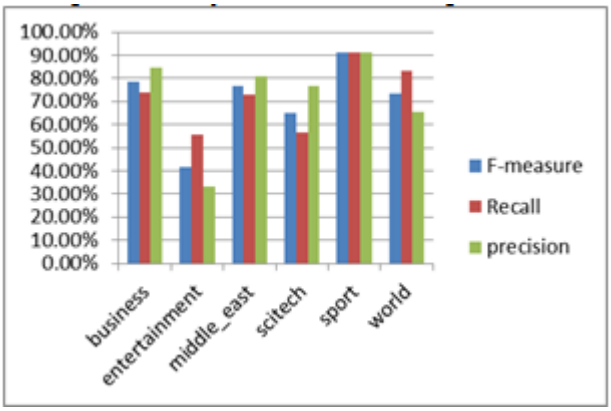

Figure 6: Recall and precision and F-measure for 4-gram

By looking to the previous figures and depends on it, it's too clear to show the difference between recall and precision measure amid dissimilar n-gram sorts. The figure below represents the f-measures of the whole classes between the n-gram, bi-gram accomplished the maximum values, so demonstrate that the greatest efficiency sort of n-grams is the bi-gram, else the tri-gram and 4-gram done very convergent values.

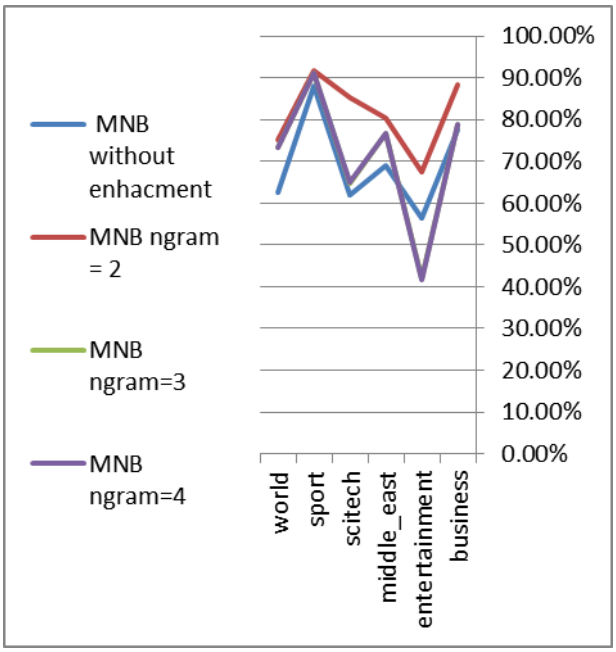

Figure 7: variation of f-measures between different n-gram

Conclusion

In this research the Naive Bayes is implemented as a classification algorithm, that is improve the accuracy of the classification process, also to have more accuracy and high classification speed we integrated the Multinomial Naive Bayes model MNB to the process. This study has been proposed a process to raise the performance of the multinomial naïve Bayes classifier tool. The improved Multinomial Naïve Bayes has been assessed in order to discover the greatest improvement method, the multinomial naïve Bayes has been enhanced by applying the n-gram. Dissimilar approaches have been utilized to measure the performance of the classifier like the recall, precision, and F-measure which associations the precision and the recall measured , in order receiving a large image of the performance with taking in mind that the recall and precision have alike significance in determining the performance. Bigram which accomplished the maximum values is the greatest efficiency classifier among the three planned classifiers.

References.

[1] Goller, C., Löning, J., Will, T., & Wolff, W. (2000). Automatic Document Classification-A thorough Evaluation of various Methods. ISI, 2000, 145-162.

[2] Lam, W., Ruiz, M., & Srinivasan, P. (1999). Automatic text categorization and its application to text retrieval. Knowledge and Data Engineering, IEEE Transactions on, 11(6), 865-879..

[3] Li, Y. H., & Jain, A. K. (1998). Classification of text documents. The Computer Journal, 41(8), 537-546.

[4] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval (Vol. 1, p. 496). Cambridge: Cambridge university press.

[5] McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, pp. 41-48).

[6] Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. Machine learning, 39(2-3), 103-134.

[7] Shimodaira, H. (2014). Text Classification using Naive Bayes. Learning and Data Note 7. Informatics 2B.

[8] Ramdass, D., & Seshasai, S. (2009). Document classification for newspaper articles.

[9] Pop, I. (2006). An approach of the Naive Bayes classifier for the document classification. General Mathematics, 14(4), 135-138.

[10] Mustafa, S. H. (2012). Word Stemming for Arabic Information Retrieval: The Case for Simple Light Stemming. Abhath Al-Yarmouk: Science & Engineering Series, 21(1), 2012.