ISSN:0975-9646

Novel Approach to Identify Relevant and Irrelevant Questions Using Text Categorization

C.M.Nalayini, J.keerthika
Department of Information Technology, Velammal Engineering College, Chennai.

Abstract-Micro-blogging lets users to publish online short text messages in real time via the web, SMS, instant messaging clients, etc. Micro-blogging is an effective tool in the classroom and has gained more interest from the education community. This paper proposes a text categorization strategy for differentiating between the relevant and irrelevant questions. Apart from the question correlation technique to rate the questions, personalization and correlation between the questions have been proved effective in future works. In this work in addition to the above techniques the use of keywords to extend this question categorization for more than one course is proposed. This can help the instructor greatly having one platform for all courses and in the same way to select the course easily for answering.

1.INTRODUCTION

Micro-blogging is a web technology that allows the users to post small messages with up to characters. The wide acceptance of this technology in social networking media is due to the fact that the user can communicate with each others through different channels such as the web, mobile devices, e-mail etc. Micro-blogging allows several users to post their messages in one's blog, to which he/she can respond. In this way a professional networking can be achieved using Micro-blogging technology. Due to its popularity among students and its ease to use, this micro-blogging can be effectively and productively employed for education purposes.

Due to the high number of students and limited time available for the instructors, it gets difficult to clarify the doubts of all students during the lecture hours. To overcome this problem micro-blogging can be used as a medium of communication between students and their instructors. In this way the students can post their queries on the instructor's blog, which he/she can answer in his free time.

Even though it sounds interesting, it can become difficult for the instructor to answer all the questions, once the number of questions posted increases. There may be some irrelevant questions also. So he has to spend more time in recognizing the relevant questions first which he should answer first.

The previous works have proposed a tool to categorize the posted questions into relevant and irrelevant questions to assist the instructor. This method is developed for one particular course only. If an instructor handles more than one course, the proposed method can not be employed. This work proposes a strategy for handling the multiple courses offered by an instructor in an effective way.

Proposed Method

Different techniques have been proposed in [2] for identifying the relevant and irrelevant questions. In addition to this the concept of using keywords has been used in this work to differentiate between different the questions posted on different courses. These different techniques are explained shortly in this section.

Keyword

The students posting their questions are advised to use a course specific keyword (e.g. JAVA) as the starting word in their questions. In this way the questions posted in an instructor's blog can be categorized easily using the keywords. With this the instructor can have different set of questions corresponding to each course he/she has been handling. As a next step the relevant and irrelevant questions in each of the course has to identify to help him/her in saving time.

Removing stop words

As a first step to identify the relevant and irrelevant questions, the stop words are removed from the posted questions. Stop words are a selected set of words which are filtered from the given text before processing them further. They are normally small function words in English like the, is, that, which, etc,. [7] Shows that removing the stop words makes the process of identifying the relevant and irrelevant questions more effective.

Correlation between questions and course material

The stop word removed question sets are correlated with the course material which are stored in an accessible database. In this way the questions can be ranked depending on how relevant they are to the course. For doing this techniques like Support Vector Machines (SVM) as proposed in [5] can be employed.

Personalization

The students are ranked based on their previous history of questioning. In this way questions coming from a student who has been asking good questions get favored.

Correlation between questions

This also makes the categorization more effective, especially when the course material database is not available. In this method the similar questions from the posted set of questions are identified. This can further help in eliminating the irrelevant questions. A final assessment of the ranks provided by each of the above method leads to an effective categorizing of the questions.

2.METHODOLOGIES

Support Vector Machine

Support Vector Machines (SVMs) have been successfully applied to a number of real world problems such as handwritten character and digit recognition face detection text categorization and object detection in machine vision. They manifest an impressive resistance to over fitting a feature which can be explained using VC theory and their training is performed by maximizing a convex functional, which means that there is a unique solution that can always be found in polynomial time. For simple binary classification tasks they work by mapping the training points into a high-dimensional feature space where a separating hyper plane can be found which has a maximal distance from the two classes of labeled points. This minimizes the effective VC dimension of the system enforcing good generalization. The task of finding the maximal margin hyper plane is reduced to a quadratic programming (QP) problem which can be solved using optimization routines.

Micro-blogging questions/messages are in textual format; therefore detecting their types can be treated as a text categorization (TC) problem. Support Vector Machines (SVM) has been shown to be one of the most accurate as well as widely used text categorization techniques. In this work the simplest linear version of SVM (i.e. with a linear kernel) was used as the TC classifier that can be formulated as a solution to an optimization problem. Note that, the positive class (i.e. y= +1) represents the relevant questions, and the negative class (i.e., y= -1) represents the irrelevant questions.

Support Vector machines implement complex decision rules by using a non-linear function to map training points to a high-dimensional feature space where the labeled points are separable. A separating hyper plane is found which maximizes the distance between itself and the nearest training points. The hyper plane is, in fact, represented as a linear combination of the training points. Theoretical results exist from VC theory, which guarantee that the solution found will have high predictive power, in the sense that it minimizes an upper bound on the test error.

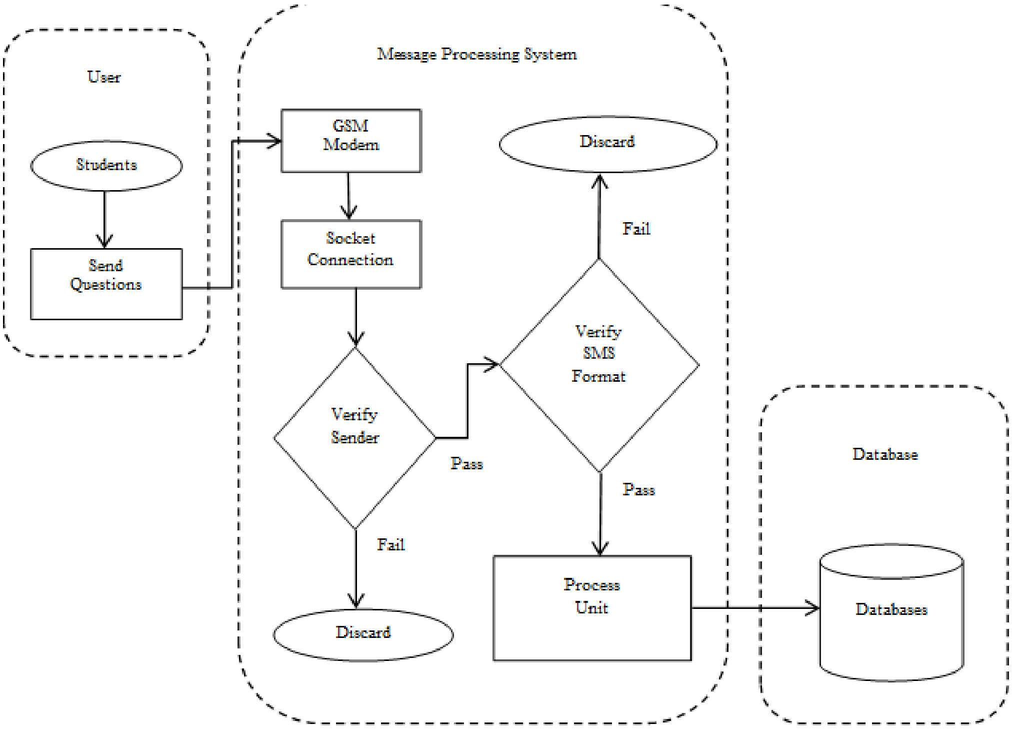

Fig 1: Flow Chart

3. CONCLUSION AND FUTURE WORK

In this paper a novel strategy for automatic sorting of the course specific questions, and indentifying the relevant questions posted on an instructors blog by students attending more than one course has been presented. It has been shown in [7] to be beneficial to utilize the correlation among questions and available lecture materials as well as the correlations between questions asked in a lecture. Furthermore, it is found to be significantly more effective to remove stop words when calculating the correlations among questions themselves. Finally, utilizing students' votes on questions is found not to be effective, although it has been shown to be useful in community question answering environments for question quality assessment. As of now micro-blogging system don't send any alert information to particular subject staff. In future we can send alert message to particular subject staff for quick response.

REFERENCES

[1] Java, A., Song, X., Finin, T. and Tseng, B. (2007). Why we Twitter: Understanding Micro-Blogging Usage and Communities. *Proceedings of the 9th WEBKDD Conference, pp. 56-65.*

[2] Jeon, J., Croft, B. and Lee, J. (2005). Finding Semantically Similar Questions Based on Their Answers. *Proceedings of the 28th ACM SIGIR Conference, pp. 617-618.*

3] Jeon, J., Croft, B., Lee, J. and Park, S. (2006). A Framework to Predict the Quality of Answers with Non- Textual Features. *Proceedings of the 29th ACM SIGIR Conference, pp. 228-235.*

[4] Jikoun, V. and de Rijke, M. (2005). Retrieving Answers from Frequently Asked Questions Pages on the Web. *Proceedings of the 14th ACM CIKM Conference, pp. 76-83.*

[5] Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of 10th ECML Conference, pp. 137-142.*

[6] Li, B., Liu, Y., Ram, A., Garcia, E. and Agichtein, E. (2008). Exploring Question Subjectivity prediction in Community QA. *Proceedings of the 31st ACM SIGIR Conference, pp. 735-736*

[7] Suleyman cetintas, luo Si, Hans Aagard and Kyle Bowen(2011).Micro-blogging in classroom: Classifying Students' Relevant and irrelevant Questions in a Micro-blogging Supported Classroom.

[8]] Porter, M. F. (1980). An algorithm for suffix stripping. *Program: Electronic Library and Information Systems, 14(3), pp. 130–137.*

[9] Song, Y., Cao, C. and Rim, H. (2008). Question Utility: A Novel Static Ranking of Question Search. *Proceedings of the 23rd AAAI Conference, pp. 1231-1236.*

[10] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34(1), pp. 1-47.*

[11] Tang, T. Y. and Mccalla, G. (2005). Smart Recommendation for an Evolving E-Learning System. *International Journal on ELearning, 4(1), pp. 105–130.*

[12] Ullrich, C., Borau, K., Luo, H., Tan, X., Shen, L. and Shen, R. (2008). Why Web 2.0 is good for learning and for research: Principles and prototypes. *Proceeding of 17th ACM WWW Conference, pp. 705-714.*