International Journal of Pure and Applied Mathematics
Volume 116 No. 9 2017, 423-427
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

EXTRACTION-BASED TEXT MINING TECHNIQUES FOR FLAW VERDICT OF RAILWAY SYSTEMS

[1]K.Shanmugapriya, [2]I.Mary Linda
[1,2] Assistant Professor, Department of CSE,
BIST,BIHER, Bharath University, Chennai
[1]shanmugapriya.cse@bharathuniv.ac.in, [2]marylinda.cse@bharathuniv.ac.in

Abstract: An immense amount of content data is recorded inside the types of repair verbatim in railroad upkeep areas. Prudent content mining of such upkeep data assumes an essential part in sleuthing inconsistencies and rising issue distinguishing proof strength. In any case, unstructured verbatim, high-dimensional data, and lopsided blame class circulation make challenges for include options and blame recognisable proof. We tend to propose a bi-level element extraction-based content mining that coordinates alternatives extricated at every linguistic structure and semantic levels with the expect to support the blame order execution. We tend to introductory perform Associate in Nursing enhanced $\chi 2$ measurements based component decision at the sentence structure level to beat the instructive trouble caused by Associate in Nursing unequal data set. At that point, we play out an earlier idle Dirichlet assignment based component decision at the semantic level to downsize the information set into a low-dimensional theme space. At long last, we tend to meld blame choices got from every punctuation and phonetics levels by means of serial combination. The arranged system utilizes blame alternatives at totally extraordinary levels and improves the exactitude of blame distinguishing proof for all blame classes, essentially minority ones. Its execution has been approved by utilizing a railroad support informational index gathered from 2008 to 2014 by a railroad partnership. It beats old methodologies.

Keywords: SDU(speed distance unit), Dirichlet allocation,virtual and click features based learning to rank.

1. Introduction

From repair verbatim information, content mining methods can be utilized to build up the relationship between blame terms and blame classes with the end goal that these affiliations can be utilized to enhance the exactness of blame determination. In support records[1-3], there are several thousand or even a huge number of unmistakable terms or tokens. After end of stop words and stemming, the arrangement of components is still too expensive for some learning calculations. In upkeep reports, the quantity of cases in one blame class (i.e., dominant part class) is altogether more noteworthy than that of the others (i.e., majority classes). Such imbalanced class circulations have represented a genuine trouble to most classifier learning calculations which accept a relatively balanced distribution. We have improved x^2 statics for syntax level. This work proposes a bi-level feature extraction-based text mining for fault diagnosis to meet the aforementioned challenges by automatically analyzing the repair verbatims. Our main idea is to extract fault features at syntax and semantic levels respectively and then fuse them to achieve the desired results. Considering the fact that the extracted features at each level gives a different emphasis to a particular aspect of feature spaces and has its deficiencies, the proposed feature fusion of two levels may enhance the precision of fault diagnosis for all fault classes, especially majority ones[4-6].

2. Existing System

A vast amount of text data is recorded in the forms of repair verbatim in railway maintenance sectors. Efficient text mining of such maintenance data plays an important role in detecting anomalies and improving fault diagnosis efficiency. However, unstructured verbatim, high-dimensional data, and imbalanced fault class distribution pose challenges for feature selections and fault diagnosis. In the event of malfunctioning, the diagnostic trouble symptoms are generated and transmitted to the monitoring center database. After every diagnosis episode a repair verbatim is recorded, which consists of a textual description of the mixture of fault symptom[7].

2.1 Existing System Disadvantages

a)High-dimension data.
b)Imbalanced fault class distribution.
c)Unsupervised text mining models.

3. Proposed System

We propose a bilevel feature extraction-based text mining that integrates features extracted at both syntax and semantic levels with the aim to improve the fault classification performance. We first perform an improved $\chi 2$ statistics-based feature selection at the syntax level to overcome the[8] learning difficulty caused by an imbalanced data set. Then, we perform a prior latent Dirichlet allocation-based feature selection at the semantic level to reduce the data set into a low-dimensional topic space.

3.1 Proposed System Advantages

a)To reduce the data set into a low-dimensional.
b)To overcome the learning difficulty caused by an imbalanced data set.

4. Implementation

We propose a bi-level feature extraction-based text mining that integrates features extracted at both syntax and semantic levels with the aim to improve the fault classification performance. We first perform an improved $\chi 2$ statistics-based feature selection at the syntax level to overcome the learning difficulty caused by an imbalanced data set. Then, we perform a prior latent Dirichlet allocation-based feature selection at the semantic level to reduce the data set into a low-dimensional topic space[9-11]

4.1 User Interface Design

To Login into the website. User must register their Details into the server. Then only the user can able to login into the website. So all the user details can be able to store in the database. The database will maintain all the user details.

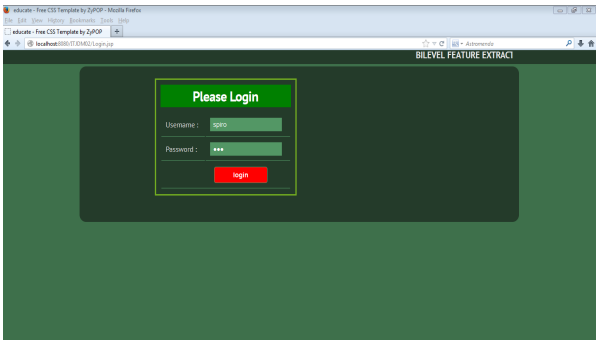

Figure 1. Customer Login

4.2 To Start Train

Once the user has registered their details and logged into the website the user have to click on start train tab to generate the fault verbatim record and it will show a popup message.

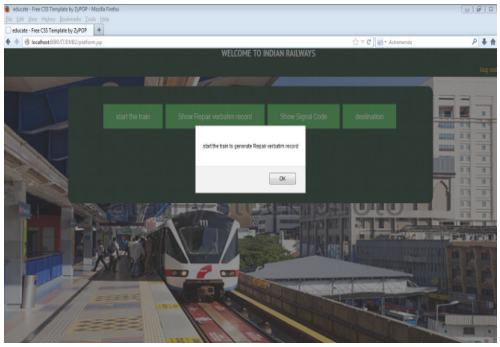

Figure 2. Generation of verbatim record

4.3 Railway Records Maintenance

It helps to extract the railway maintenance records from 2008 to 2015. So we may notice how the railway has performed those years.

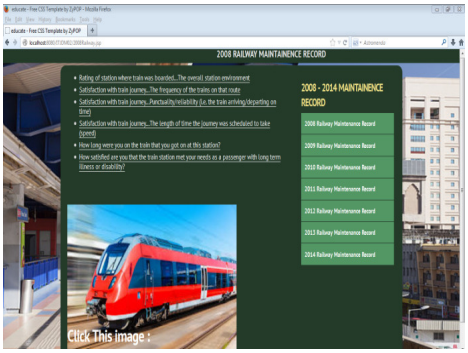

Figure 3. Maintenance record

4.4 Generate Fault Verbatim Records

This module helps to generate the Fault Verbatim Records. When the Train starts from the specified station. The reason to generate fault verbatim record is, if the train exceed the speed limits and if the train does not throws the signal at particular time, etc.

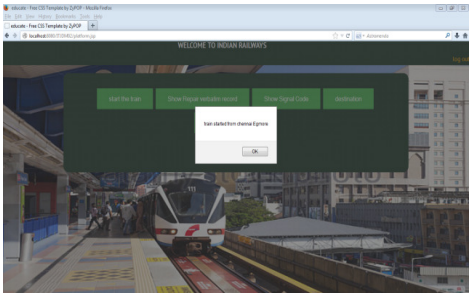

Figure 4. Faultverbatim record

4.5 Generate Signal Code

If the train starts from the particular location. Then the signal code will generate for the particular train. So the person can identify that the particular train has been started. If it delays to generate the signal code then the message will send to the fault verbatim [11-13]

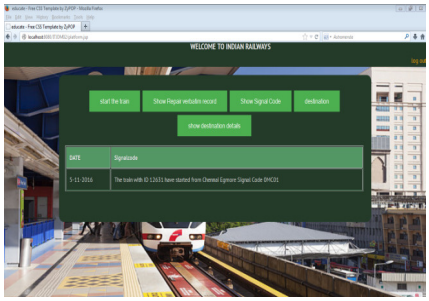

Figure 5. Signal code

4.6 Generate Destination Details

This module helps to generate the destination details, if the train reach the destination. So we come to know the destination details of date, time, km, station Code, Arrival time. All the details will store into the database

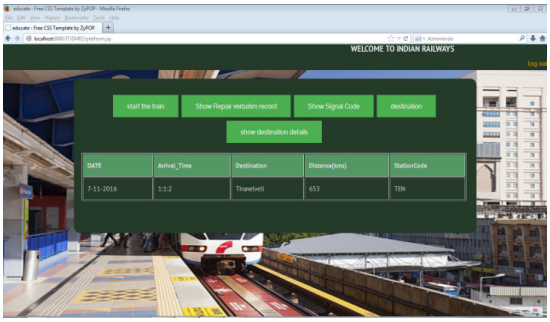

Figure 6. Destination details

4.7 Train Destination Detail

This allows us to check the destination schedule and status of a particular train with arrival time and Halt time at each station .From this wecome to know whether the train has reached the destination[20-26]

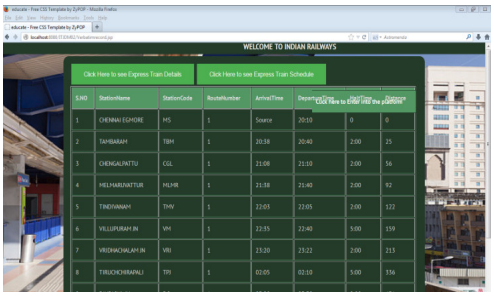

Figure 7. Train schedule

5. Conclusion

Text mining of repair verbatim for fault diagnosis of railway systems poses a big challenge due to unstructured verbatim ,high-dimension data, and imbalanced fault classes. In this paper, to improve the fault diagnosis performance, especially on minority fault classes, we have proposed a bi-level feature extraction-based text mining method .We first adjust the exclusive feature weights of various fault classes based on $\chi 2$statistics and their distributions. Then we reselect the common features according to both relevance and Hellinger distance. This can be categorized as feature selection at the syntax level. Next, we extract semantic features by using a prior LDA model to make up for the limitation of fault terms derived from the syntax level. Finally, we fuse fault term sets derived from the syntax level with those from the semantic level by serial fusion.

References

[1] L. Huang and Y. L. Murphey, "Text mining with application to engineering diagnostics," in *Proc. 19th Int. Conf. IEA/AIE*, Annecy, France, 2006, pp. 1309–1317.
[2] D. G. Rajpathak, "An ontology based text mining system for knowledge discovery from the diagnosis data in the automotive domain," *Comput. Ind.*, vol. 64, no. 5, pp. 565–580, Jun. 2013.
[3] J. Silmon and C. Roberts, "Improving switch reliability with innovative condition monitoring techniques," *Proc. IMechE, F C J. Rail Rapid Transit*, vol. 224, no. 4, pp. 293–302, 2010.
[4] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[5] J. Chang, J. Boyd-Graber, C.Wang, S. Gerrish, and D. Blei, "Reading tea leaves: How humans interpret topic models," *Neural Inf. Process. Syst.*, vol. 22, pp. 288–296, 2009.
[6] D. A. Cieslak and N. V. Chawla, "Learning decision trees for unbalanced data," in *Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases-Part I*. Berlin, Germany: Springer-Verlag, 2008, pp. 241–256.
[7] T. Kailath, "The divergence and Bhattacharyya distance measures in signal
selection," *IEEE Trans. Commun. Technol.*, vol. 15, no. 1, pp. 52–60, Feb. 1967.
[8] W. Wang, H. Xu, and X. Huang, "Implicit feature detection via a constrained topic model and SVM," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Seattle, WA, USA, 2013, pp. 903–907. [9] J. Yang, J. Yang, D. Zhang, and J. Lu, "Feature fusion: Parallel strategy vs. serial strategy," *Pattern Recognit.*, vol. 36, no. 6, pp. 1369–1381, Jun. 2003.
[10] C. Drummond and R. C. Holte, "C4. 5, class imbalanced, and cost sensitivity: Why under-sampling beats over-sampling," in *Proc. Workshop Learn. Imbalanced Datasets II, ICML*, Washington, DC, USA, 2003, pp. 1–8.
[9] Udayakumar R., Kaliyamurthie K.P., Khanaa, Thooyamani K.P., Data mining a boon: Predictive system for university topper women in academia, World Applied Sciences Journal, v-29, i-14, pp-86-90, 2014.
[10] Kaliyamurthie K.P., Parameswari D., Udayakumar R., QOS aware privacy preserving location monitoring in wireless sensor network, Indian Journal of Science and Technology, v-6, i-SUPPL5, pp-4648-4652, 2013.
[11] Brintha Rajakumari S., Nalini C., An efficient cost model for data storage with horizontal layout in the cloud, Indian Journal of Science and Technology, v-7, i-, pp-45-46, 2014.
[12] Brintha Rajakumari S., Nalini C., An efficient data mining dataset preparation using aggregation in relational database, Indian Journal of Science and Technology, v-7, i-, pp-44-46, 2014.
[13] Khanna V., Mohanta K., Saravanan T., Recovery of link quality degradation in wireless mesh networks, Indian Journal of Science and Technology, v-6, i-SUPPL.6, pp-4837-4843, 2013.
[14] Khanaa V., Thooyamani K.P., Udayakumar R., A secure and efficient authentication system for distributed wireless sensor network, World Applied Sciences Journal, v-29, i-14, pp-304-308, 2014.
[15] Udayakumar R., Khanaa V., Saravanan T., Saritha G., Retinal image analysis using curvelet transform and multistructure elements morphology by reconstruction, Middle - East Journal of Scientific Research, v-16, i-12, pp-1781-1785, 2013.
[16] Khanaa V., Mohanta K., Saravanan. T., Performance analysis of FTTH using GEPON in direct and external modulation, Indian Journal of Science and Technology, v-6, i-SUPPL.6, pp-4848-4852, 2013.
[17] Kaliyamurthie K.P., Udayakumar R., Parameswari D., Mugunthan S.N., Highly secured online voting system over network, Indian Journal of Science and Technology, v-6, i-SUPPL.6, pp-4831-4836, 2013.
[18] Thooyamani K.P., Khanaa V., Udayakumar R., Efficiently measuring denial of service attacks using appropriate metrics, Middle - East Journal of Scientific Research, v-20, i-12, pp-2464-2470, 2014.
[19] R.Kalaiprasath, R.Elankavi, Dr.R.Udayakumar, Cloud Information Accountability (Cia) Framework Ensuring Accountability Of Data In Cloud And Security In End To End Process In Cloud Terminology, International Journal Of Civil Engineering And Technology (Ijciet)Volume 8, Issue 4, Pp. 376–385, April 2017.
[20] R.Elankavi, R.Kalaiprasath, Dr.R.Udayakumar, A fast clustering algorithm for high-dimensional data, International Journal Of Civil Engineering And Technology (Ijciet), Volume 8, Issue 5, Pp. 1220–1227, May 2017.
[21] R. Kalaiprasath, R. Elankavi and Dr. R. Udayakumar. Cloud. Security and Compliance - A Semantic Approach in End to End Security, International Journal Of Mechanical Engineering And Technology (Ijmet), Volume 8, Issue 5, pp-987-994, May 2017.
[22] Thooyamani K.P., Khanaa V., Udayakumar R., Virtual instrumentation based process of agriculture by automation, Middle - East Journal of Scientific Research, v-20, i-12, pp-2604-2612, 2014.
[23] Udayakumar R., Thooyamani K.P., Khanaa, Random projection based data perturbation using geometric transformation, World Applied Sciences Journal, v-29, i-14, pp-19-24, 2014.
[24] Udayakumar R., Thooyamani K.P., Khanaa, Deploying site-to-site VPN connectivity: MPLS Vs IPSec, World Applied Sciences Journal, v-29, i-14, pp-6-10, 2014.
[24] T. Padmapriya and V. Saminadan, "Improving Throughput for Downlink Multi user MIMO-LTE Advanced Networks using SINR approximation and Hierarchical CSI feedback", International Journal of Mobile Design Network and Innovation- Inderscience Publisher, ISSN : 1744-2850 vol. 6, no.1, pp. 14-23, May 2015