

Document-to-Sentence Level Technique for Novelty Detection

Sushil Kumar and Komal Kumar Bhatia

Abstract Novelty identification is accustomed to distinguishing novel data from an approaching stream of documents. In this study, we proposed a novel methodology for document-level novelty identification by utilizing document-to-sentence-level strategy. This work first splits a document into sentences, decides the novelty of every sentence, then registers the record-level novelty score in view of an altered limit. Exploratory results on an arrangement of document demonstrate that our methodology beats standard document-level novelty discovery as far as repetition exactness and excess review. This work applies on the document-level information from an arrangement of documents. It is valuable in identifying novel data in information with a high rate of new documents. It has been effectively incorporated in a true novelty identification framework in the zone of information retrieval.

Keywords Novelty identification • Sentence segmentation • Document novelty identification

1 Introduction

There is a nonstop increment in the information content that is transferred through the Internet between customers, administrations, and Internet clients [1]. Individuals who are in media, security offices get an immense measure of stories, papers, articles, and reports from an expansive number of resources. Such troublesome circumstance propelled the scientists to concoct new programmed framework which is in view of novelty identification. The most recent decade saw an expanding enthusiasm for the novelty location which expects to manufacture programmed

S. Kumar (✉) · K. K. Bhatia
YMCA, University of Science and Technology, Faridabad 121006, India
e-mail: panwar_sushil2k@yahoo.co.in

K. K. Bhatia
e-mail: komal_bhatia1@rediffmail.com

frameworks which are proficient to disregard previous stories, papers, and articles as of now read or known and tell the clients of such frameworks about any new stories, papers, reports, and articles. There is an expanding requirement for distinguishing novel and important data out of a mass of approaching content reports. Novel data for this situation alludes the message which contain new substance and novelty recognition is the procedure of singling out novel data [2–4] from a given arrangement of content documents [5]. Thus due to this procedure, clients can spare time by perusing just the new data, while the rehashed data is separated out.

2 Literature Review

Event level and sentence level are two ways for novelty identification. We provide a review for novelty identification in brief by the related research as follow.

2.1 *Novelty Identification Using Event Level*

This work is based on online new event identification [3, 6–13]. Available techniques for new event identification are related to clustering algorithms.

2.2 *Novelty Identification Using Sentence Level*

Study on novelty identification at the sentence level is identified with the TREC novelty tracks [3, 14–16]. Different exploration gatherings provide an interest in the TREC 2002–2004 novelty tracks and reported their outcomes [2].

2.3 *Comparison*

In this study, a new approach for novelty identification at document level has been proposed that is different from the available approaches in the literature in the following sense:

- (a) Available approaches assume sentences and documents as two different resources and decide novelty individually.
- (b) The proposed approach regards a document as redundant if it shares a single sentence with the history document.
- (c) The proposed work mainly focuses on sentence-level module, which, in turn, improves code reuse for novelty identification.

3 Proposed Work for Novelty Detection at Document Level

The idea of novelty detection will optimize the search engine results. Many applications have utilized novelty identification to reduce nonredundant information presented to user. In this study, a novel approach to document level has been proposed. The algorithm is accustomed to remove the redundancy of the results, which increases speed to find novel information in the documents. A novelty score is calculated by sentence segmentation instead of whole document. The document is required to be preprocessed for sentence segmentation [17].

3.1 Document-Level Novelty Detection Algorithm

Document-level novelty detection (DND) algorithm is a proposed detection algorithm which is used to find whether a document provides a novel information or not, when compared with the history documents. DND first splits the document into sentences [17] and computes the novelty score of each document based on a fixed threshold. Sentence segmentation is used a tool name Stanford parser, which splits the document into sentences. Sentences are then compared with all the history sentences to compute the similarity between those sentences.

To compute the nature of document, similarity is converted to novelty score for each sentence. A minimum value is selected out of the novelty values and finally, the decision has to be made. The architecture of the proposed system is show in Fig. 1.

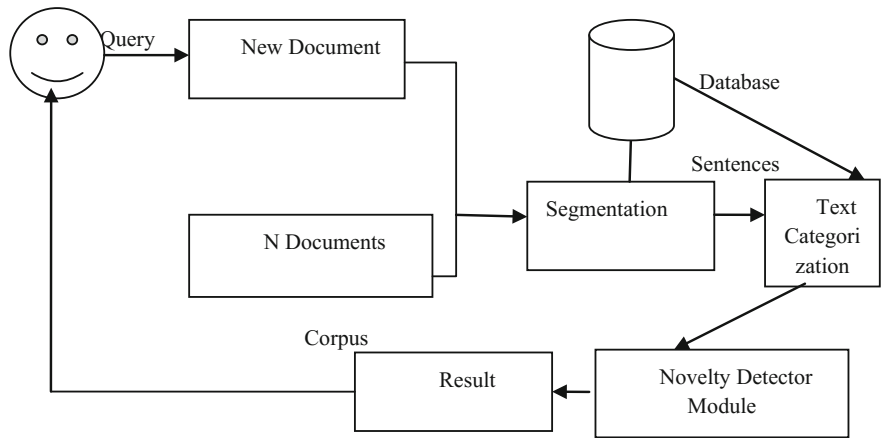


Fig. 1 Architecture of proposed system

3.2 Novelty Detector Module

This module helps in discovering the document novelty. The procedure of this module is as demonstrated in Fig. 2. The document is divided into sentences; register the novelty score of every sentence by utilizing the sentence-level novelty identification. At that point, the normal of novelty score is compared with threshold, and if the estimation of novelty score is more prominent than the fixed threshold, then the document is considered as novel generally not.

For similarity measure, cosine similarity is used for good performance to identify the novel information between sentences. This has been cleared from the existing study. The cosine similarity is defined between two sentences as:

$$\cos(s_t, s_i) = \frac{\sum \text{wk}(s_t) \text{wk}(s_i)}{\|s_t\| \cdot \|s_i\|}$$

The novelty similarity score is calculated as (1-cosine similarity score). Each of the history sentences is separately compared with the current sentence. Then minimum novelty score from them provides the novelty score of the current sentence [18].

4 Experimental Result Analysis

The proposed architecture has been simulated by using an example. User chooses a new document and that document is compared with three documents. The result is computed by finding the novelty score for each document based on a fixed threshold.

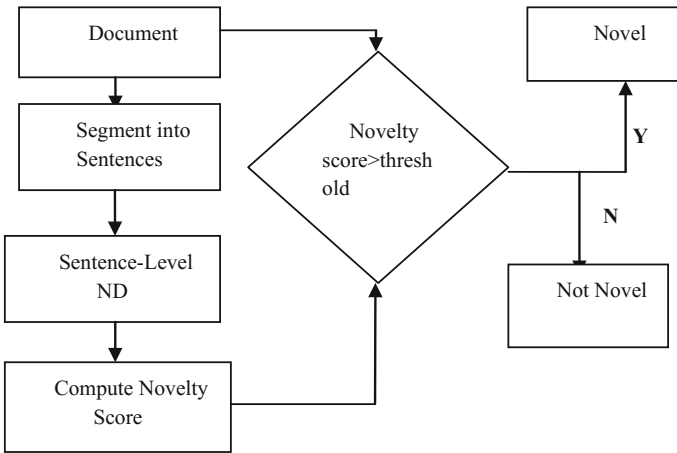


Fig. 2 Process of novelty detection

Example

- Step 1 Three documents (N1, N2, and N3) have been taken for the basic analysis. N2 and N3 documents did not show here, but the calculation is performed on these documents for results comparison (Fig. 3)
- Step 2 Now user selects a new document (newDoc) (Fig. 4)
- Step 3 All the documents are segmented into sentences, and each sentence of the new document are compared with all the sentences of history documents
- Step 4 Sentences of N1 document are taken one by one
- Step 5 $\text{Len}(\text{newDoc1}) == \text{Len}(\text{senN1})$
- Step 6 Find cosine similarity of each sentence (applied the same for N2 and N3) (Table 1)
- Step 7 Now the maximum values of cosine similarity from each table is selected (Table 2)
- Step 8 Find novelty score for each document (Table 3)
- Step 9 Now compute minimum novelty score for each document (Table 4)
- Step 10 Find the average novelty score
 $\text{avgNovel} = (0.15 + 0.10 + 0.02)/3 = 0.09$
- Step 11 Now we compare the average novelty score with the fixed threshold value Threshold = 0.45
 $\text{avgNovel} = 0.09$ which is less than the threshold value

So, new document ND is not novel.

Our society is suffering from various social evils at the moment. The dowry system is common almost in all parts of India. Dowry has been stated as "the value paid by the parents for getting their daughters the place of a daughter-in-law". Parents pay huge sums of money so that their daughters may secure a satisfactory and permanent post. The groom's parents try to mine the maximum from a matrimonial association. They insist on getting huge amount of price, luxury items like television sets, VCR's, refrigerators, cars, and even houses.

Fig. 3 Document N1

Dowry has been defined as "the price paid by the parents for getting their daughters the post of a daughter-in-law". In due course of time demand for the dowry became most essential condition of the marriage settlement. The groom's parents try to extract the maximum from a matrimonial alliance. The amount of the dowry depends on the jobs the grooms may be holding at the time of marriage. The devil of dowry has put an end to the happiness of several couples even after marriage. When demands for dowry are not met, the bride is subject to torture, and often even killed.

Fig. 4 New document

Table 1 Cosine similarity values of N1 document

ND1	0.842	0.389
ND2	0.428	0.11
ND3	0.85	–
ND4	0.127	–
ND5	0.427	0.117
ND6	0.252	0.06

Table 2 Maximum values from each table

	N1	N2	N3
ND1	0.84	0.33	0.13
ND2	0.43	0.90	0.16
ND3	0.85	0.23	0
ND4	0.13	0.59	0.98
ND5	0.43	0.86	0.28
ND6	0.25	0.26	0.88

Table 3 Novelty scores

New document	N1	N2	N3
ND1	0.16	0.67	0.87
ND2	0.57	0.10	0.84
ND3	0.15	0.77	1
ND4	0.87	0.41	0.02
ND5	0.57	0.14	0.72
ND6	0.75	0.74	0.12

Table 4 Minimum novelty scores

ND	N1	N2	N3
	0.15	0.10	0.02

From the result analysis, it has been proved that this proposed method provides proper result in lesser amount of time and with better efficiency.

5 Conclusion

In this paper, a system has been suggested that aptly applies document-level novelty identification. This structure makes record-level novelty identification more powerful by receiving the procedures for the sentence level. Results demonstrate that proposed strategy significantly enhances the document-level novelty identification execution as far as repetition accuracy and excess review [19–22]. The perceptions

are exceptionally useful for effectively coordinating DND to a true novelty identification framework in information processing [23–32].

References

1. Greengrass, E.: Information Retrieval: A Survey, DOD Technical Report TR-R52-008-001 (2000)
2. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, ISBN 0070544840 (1983)
3. ciir.cs.umass.edu
4. www.sersc.org
5. Soboroff, I., Harman, D.: Novelty detection: the TREC experience. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada, pp. 105–112 (2005)
6. Allan, J., Wade, C., Bolivar, A.: Retrieval and novelty detection at the sentence level. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, pp. 314–321 (2003)
7. Ng, K.W., Tsai, F.S., Chen, L., Goh, K.C.: Novelty detection for text documents using named entity recognition. In: 2007 6th International Conference On Information, Communications And Signal Processing, ICICS (2007)
8. Allan, J., Paka, R., Lavrenko, V.: On-line new event detection and tracking. In: Proceedings of SIGIR-98, pp. 37–45 (1998)
9. Yang, Y., Zhang, J., Carbonell, J., Jin, C.: Topic-conditioned novelty detection. In: SIGKDD, pp. 688–693 (2002)
10. Stokes, N., Carthy, J.: First story detection using a composite document representation. In: Proceedings of HLT01 (2001)
11. Franz, M., Ittycheriah, A., McCarley, J.S., Ward, T.: First story detection, combining similarity and novelty based approach. Topic Detection and Tracking Workshop (2001)
12. Allan, J., Lavrenko, V., Jin, H.: First story detection in TDT is hard. In: Proceedings of CIKM (2000)
13. Yang, Y., Pierce, T., Carbonell, J.: A study on retrospective and on-line event detection. In: Proceedings of SIGIR-98 (1998)
14. Brants, T., Chen, F., Farahat, A.: A system for new event detection. In: Proceedings of ACM SIGIR2003 (2003)
15. Kumaran, G., Allan, J.: Text classification and named entities for new event detection. In: Proceedings of ACM SIGIR, pp. 297–304 (2004)
16. Harman, D.: Overview of the TREC 2002 Novelty Track. In: TREC (2002)
17. Tsai, F.S.: D2S: document-to-sentence framework for novelty detection. Knowl. Inf. Syst. (2010)
18. Verhaegen, P.-A., Vandevenne, D., Dufloy, J.R.: Originality and novelty: a different universe. In: Proceedings of DESIGN 2012, the 12th International Design Conference, Dubrovnik, Croatia, pp. 1961–1966 (2012)
19. Brants, T., Chen, F., Farahat, A.: A system for new event detection. In: Proceedings of SIGIR-03, pp. 330–337 (2003)
20. Soboroff, I., Harman, D.: Overview of the TREC 2003 novelty track. In: TREC (2003)
21. Soboroff, I.: Overview of the TREC 2004 novelty track. In: TREC (2004)
22. Allan, J., Bolivar, A., Wade, C.: Retrieval and novelty detection at the sentence level. In: Proceedings of SIGIR-03 (2003)
23. Kazawa, H., Hirao, T., Isozaki, H., Maeda, E.: A machine learning approach for QA and novelty tracks: NTT system description. In: TREC-10 (2003)

24. Qi, H., Otterbacher, J., Winkel, A., Radev, D.T.: The University of Michigan at TREC2002: question answering and novelty tracks. In: TREC (2002)
25. Eichmann, D., Srinivasan, P.: Novel results and some answers, The University of Iowa TREC-11 results. In: TREC (2002)
26. Zhang, M., Song, R., Lin, C., Ma, S., Jiang, Z., Jin, Y., Zhao, L.: Expansion-based technologies in finding relevant and new information: THU TREC2002 novelty track experiments. In: TREC (2002)
27. Kwok, K.L., Deng, P., Dinstl, N., Chan, M.: TREC2002, novelty and filtering track experiments using PRICS. In: TREC (2002)
28. Zhang, Y., Callan, J., Minka, T.: Novelty and redundancy detection in adaptive filtering. In: Proceedings of SIGIR (2002)
29. Tsai, M., Hsu, M., Chen, H.: Approach of information retrieval with reference corpus to novelty detection. In: TREC (2003)
30. Jin, Q., Zhao, J., Xu, B.: NLPR at TREC 2003: novelty and robust. In: TREC (2003)
31. Sun, J., Yang, J., Pan, W., Zhang, H., Wang, B., Cheng, X.: TREC-2003 novelty and web track at ICT. In: TREC (2003)
32. Litkowski, K.C.: Use of metadata for question answering and novelty tasks. In: TREC (2003)

Speech and Language Processing for Human-Machine
Communications

Proceedings of CSI 2015

Agrawal, S.S.; Devi, A.; Wason, R.; Bansal, P. (Eds.)

2018, XIV, 215 p. 104 illus., Softcover

ISBN: 978-981-10-6625-2