**Comparison Among Cluster Labeling Approaches**

**Abstract**

Document clustering techniques often produce a set of clusters which require human intervention for interpreting the meaning of clusters. Automatic cluster labeling refers to the method of assigning a meaning phrase to the cluster as a label. This paper proposed an unsupervised method of cluster labeling. Furthermore, the proposed method is compared with the three other n-gram based scoring methods which include Z-Order, M-Order, and T-Order. The evaluations were performed on the dataset of Urdu News Headlines. It was found that the labels produced by the proposed approach are more descriptive in contrast to other methods of cluster labeling. Future work will focus on the development of a framework to evaluate the quality of labels. The evaluation framework will consider human judgment while assessing cluster labels.

(Mei, Shen, & Zhai, 2007) proposed probabilistic methods for automatic labeling of topic models. The proposed method consists of two steps. First, the candidate labels were generated using two basic approaches which include a) Chunking and b) N-gram testing. Then these labels were ranked using a) Zero-order relevance score and b) First-Order relevance score. In contrast to the Zero-order relevance scoring method, the First-Order relevance score also considered the context while interpreting the semantics of the topic model. For context, the scoring method required the domain-specific reference corpus. The authors also described that the topic labeling methods can also be applied for labeling document clusters.

(Gourru, Velcin, Roche, Gravier, & Poncelet, 2018) presented three unsupervised N-gram based topic labeling methods which include:

1. M-Order Labeler
2. T-Order Labeler
3. Document-based Labeler

M-Order improves the Zero-Order relevance scoring method (Mei et al., 2007) by penalizing the candidate labels which are likely to be generated by other clusters/topics. T-Order captures the notion of “term hood” such that the candidate label is considered as short term if it is nested in another candidate label. For instance “ڈینگی” is a short term as it is nested in a longer-term “ڈینگی وائرس”

In this document, the proposed method of cluster labeling is compared with Zero-Order Labeler (Mei et al., 2007), M-Order Labeler, and T-Order Labeler (Gourru et al., 2018). The results are presented in Table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster** | **Zero-Order** | **M-Order** | **T-Order** | **Proposed Method** |
| 0 | ترکی شام | ترک فوجی کارروائی | ترکی شام | ترکی کی شام میں فوجی کارروائی |
| 1 | ڈی ایچ کیو ہسپتال وہاڑی کادورہ ہسپتال | ڈی ایچ کیو ہسپتال وہاڑی کادورہ ہسپتال | ڈی ایچ کیو ہسپتال وہاڑی کادورہ ہسپتال | ڈینگی کے مریضوں |
| 2 | ڈینگی مچھر | سی ای او ہیلتھ خانیوال | سی ای او ہیلتھ خانیوال | انعقاد احتیاطی تدابیر اپنا کر ڈینگی پر قابو |
| 3 | سی پیک اتھارٹی بارے | حکام | حکام | حکام سی پیک اتھارٹی |
| 4 | وزیراعظم عمران خان | وزیراعظم عمران خان | وزیر اعظم عمران خان | پاکستان اور چین سی پیک |
| 5 | ضلعی انتظامیہ لاہور | ضلعی انتظامیہ لاہور | ضلعی انتظامیہ لاہور | ضلعی انتظامیہ نے ڈینگی کی رپورٹ |
| 6 | پریشان عمران خان | پی ٹی آئی حکومت عمران خان | پی ٹی آئی حکومت عمران خان | No title constructed |
| 7 | کشمیری عوام کیساتھ اظہار یکجہتی | طاقت کشمیری حریت پسندوں | کشمیری عوام کیساتھ اظہار یکجہتی | کشمیریوں سے اظہار یکجہتی کیلئے ریلی |
| 8 | پاک سری لنکن ٹی ٹونٹی کرکٹ سیریز سکیورٹی | سری لنکن | سری لنکا | No title constructed |
| 9 | محکمہ صحت ڈینگی وائرس | محکمہ صحت ڈینگی وائرس | محکمہ صحت ڈینگی وائرس | محکمہ صحت ڈینگی وائرس |
| 10 | خیبر پختونخوا | مزید کیسز تعداد | مزید کیسز تعداد | ڈینگی کے مریضوں کی تعداد |
| 11 | ڈپٹی کمشنر محمد سیف انور جپہ | ڈپٹی کمشنر جہلم محمد سیف انور جپہ | ڈپٹی کمشنر محمد سیف انور جپہ | No title constructed |
| 12 | زیر صدارت انسداد ڈینگی | زیر صدارت انسداد ڈینگی | زیر صدارت انسداد ڈینگی | زیر صدارت انسداد ڈینگی کے حوالے |

For Zero-Order, M-Order, and T-Order, first, the candidate labels were generated using Noun Phrase chunking. Then these labels were scored using the relevance scoring mechanism. The highest scoring candidate label is chosen as the cluster label.

For clusters with ids [0, 2,3,4,5, 12], the label generated by the proposed scheme seems more relevant and descriptive. For clusters with ids [6, 8, 11] the proposed method was unable to produce any label.

**Conclusion**

This paper presents n-gram based unsupervised approach of cluster labeling. An empirical comparison among the proposed method and other cluster labeling methods (Zero-Order, M-Order, and T-Order) is also performed on the dataset of Urdu News headlines. The quantitative assessment of the cluster labels requires the development of a human evaluation framework. Future work will focus on the improvement of the proposed cluster labeling method and the development of human evaluation framework.

**References**

Gourru, A., Velcin, J., Roche, M., Gravier, C., & Poncelet, P. (2018). United we stand: Using multiple strategies for topic labeling. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-91947-8\_37

Mei, Q., Shen, X., & Zhai, C. (2007). Automatic labeling of multinomial topic models. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/1281192.1281246