Towards Chronological Stratification of *Ṛg* Veda by Data Mining

A V S D S Mahesh
Supervisor: Dr. Arnab Bhattacharya

Department of Computer Science and Engineering, Indian Institute of Technology Kanpur

July 15, 2021

Abstract

This poster addresses the possible realization of linguistically established chronological stratification of *Ṛg* Veda, the collection of sacred hymns in oldest form of Sanskrit, through computational approaches at the level of hymns. As a first step, k-means clustering is performed over the corpus and it is observed that the hymns cluster according to their topics. It is then discussed how this poses a challenge to the desired task of dating the texts.

Keywords: Computational Historical Linguistics, Vedic Sanskrit, Clustering

Introduction

Ṛg Veda is an ancient Indian text attested in the oldest and archaic form of Sanskrit. It is a collection of about thousand hymns distributed among 10 books or *maṇḍalas*. From linguistic and internal evidences in the text, it has been well established that the hymns correspond to different times. Many hymns have been suggested to belong to an earlier time and similarly many else to a later time. Yet precise periods have not been assigned to every hymn owing to the difficulty of the task [Wit12]. Thus it makes sense to attempt at trying computational approaches to address this problem.

Previous Work

Bayesian Mixture Model has been employed in [Hel20] to stratify several vedic texts and the results do correspond to existing consenses of the linguists. It has also been attempted in the same work to date the books of *Ṛg* veda. The results assign a later date to book 10 which is again correct from linguistics perspective. From here, there is still need for finer methods to match with or even complement the current linguistic understanding in order to match at the hymn-level.

Corpus

Ṛg veda consists of 1,027 hymns with about 10,600 verses. All of which are available in the tree bank [HSAW20] with annotations for lemmas and other grammatical aspects like case, gender, tense, mood etc. Each hymn is associated with divinity praised in it like *Indra, Agni, Vāyu etc.*

Clustering

Lemmas are extracted from each hymn with the stop words, which are arrived manually, removed and further each hymn is represented by an indicator vector i.e. 1 at i^{th} position if the corresponding lemma is present in the hymn, else zero. This vector is further normalized using $l2$ norm. The resulting hymn-vectors are clustered by k-means clustering with number of clusters 6. The number of clusters has been decided based on the number of major topics. It has been observed that the hymns are getting clustered according to divinities whoever so is invoked in them or in other words, according to their topics. This can be observed in the distribution of clusters for different divinities in Fig 2. The clusters can also be visualized in the 2-PCA plot in Fig 1.

Results of Clustering

The topic labels are assigned to each cluster depending to the topic to which the clusters attach with maximum proportion. For example cluster 6 happens to be in the topic *Agni* with maximum proportion, thus cluster 4 is assigned label *Agni*. If this is done so, the resulting precision, recall and f1 scores are as follows.

Topic	Precision	Recall	F1-score
Agni	**0.72**	**0.87**	**0.79**
Indra	**0.79**	**0.88**	**0.83**
Soma Pavamāna	**0.88**	**0.93**	**0.91**
Aśvins	**0.89**	0.57	**0.70**
Dawn	**0.80**	0.08	0.14
Others	0.51	**0.84**	0.64

It can seen that precision wise the five topics get significant score where as considering recall the three main topics, which together cover more than half of the text, get good score. Overall it can be inferred that the hymns are clustering according to the topics.

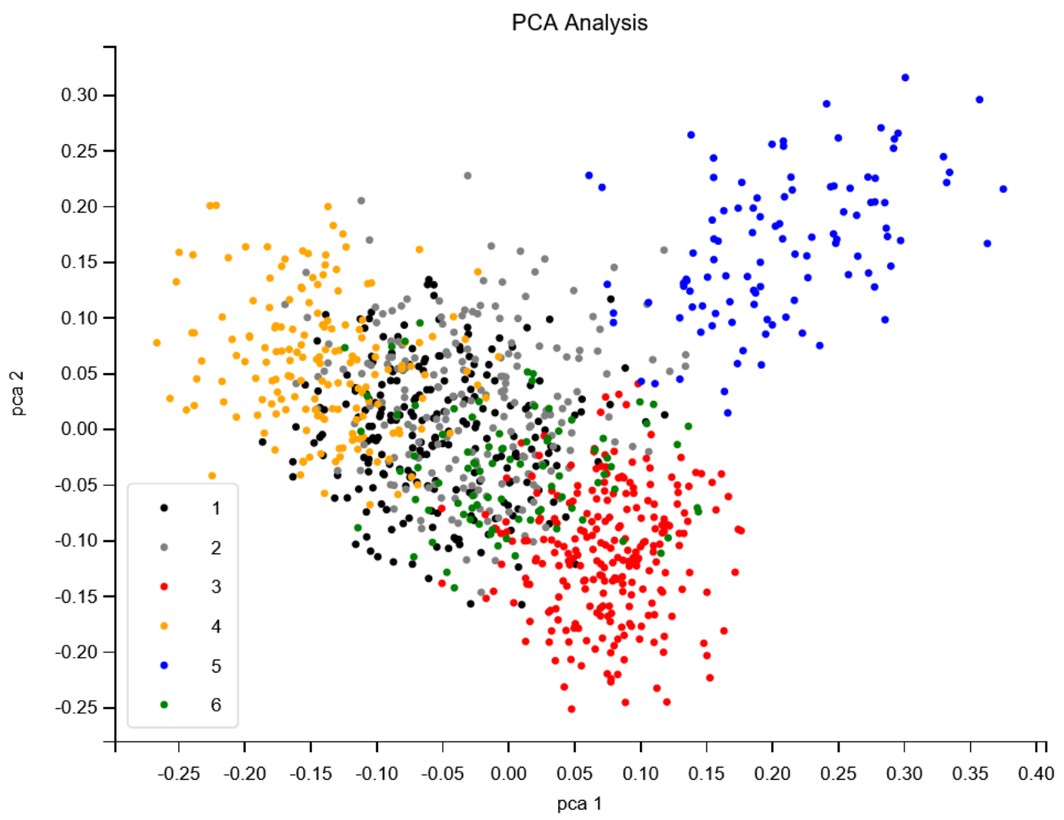

Figure: 1 2-PCA of the hymn-vectors and their clusters

Conclusion

Thus it is observed that the hymns cluster according to their respective topics instead of chronology. In other words the underlying chronological differences in the vocabulary are masked or dominated by the words that associate with a specific topic. Thus this poses a challenge in order to achieve the desired task. Currently, we are working to overcome this challenge. Other possibilities of word representations like word-embeddings are yet to be tried which may be of some help.

References

Oliver Hellwig.
Dating and stratifying a historical corpus with a Bayesian mixture model.
In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 1–9, Marseille, France, May 2020. European Language Resources Association (ELRA).

Oliver Hellwig, Salvatore Scarlata, Elia Ackermann, and Paul Widmer.
The treebank of vedic Sanskrit.
In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5137–5146, Marseille, France, May 2020. European Language Resources Association.

Michael Witzel.
4. early indian history: Linguistic and textual parametres.
In *The Indo-Aryans of Ancient South Asia*, pages 85–125. de Gruyter, 2012.

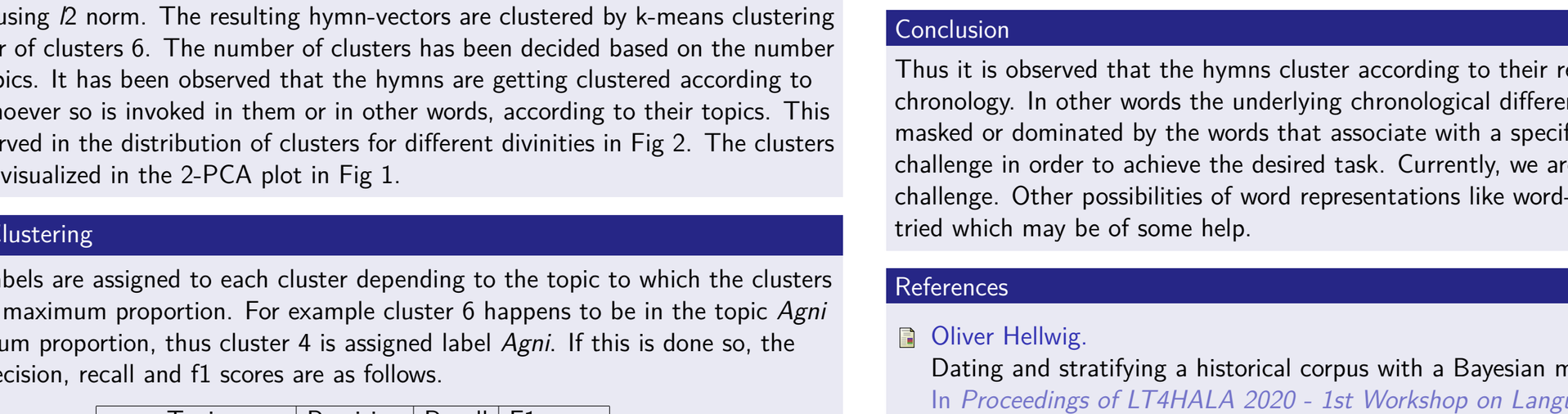

Figure: 2 Proportions of clusters among various topics