Information Retrieval on Indian Knowledge Systems - Data Curation, Semantic Search & Retrieval-Augmented-Generation

KID: 20250312 | Mr Avinash H N

Introduction

India's intangible heritage most notably comprises its traditions of knowledge - oral and textual, formal and informal. The intangible heritage includes codified and informal knowledge, together referred to as Indian Knowledge Systems (IKS). It includes a large gamut of formal texts, traditional best practices, and oral wisdom of various communities. It includes fields as diverse as philosophy, architecture, grammar, mathematics, astronomy, metrics, sociology, economy, politics, ethics, geography, logic, military science, weaponry, agriculture, mining, trade, commerce, metallurgy, shipbuilding, medicine, poetics, biology, and veterinary science. Indic knowledge systems also include pervasive aspects of life and society, such as dress, language, and cuisine.

Textual Data of IKS is scattered across physical books, manuscripts, digital PDFs, text documents in private archives, and popular websites. Popular Search, Retrieval, and Generation engines rely largely on content from sources like blogs from the World Wide Web. IKS lacks specialised search engines like Google Scholar to separate this clutter and allow retrieval based exclusively on scholarly sources. Our work under Project Dhārā aims to rectify this shortcoming by providing tools for information retrieval grounded in authentic scholarly sources. Dhārā consists of a family of interlinked databases along with specialised retrieval algorithms for fuzzy search, semantic searches and retrieval-augmented-generation (RAG).

Understanding IKS Data Sources

Building Information Retrieval (IR) Systems for IKS presents several peculiar challenges beyond the commonly known issues of multilinguality and sparse data. The structure of scholarly literature in IKS is different from the peer-reviewed journal format. IKS scholarly literature has a multilayered annotation format of primary sources (Mantras, Ślokas, Sūtras) and their elaborations and commentaries (Bhāṣya, Ṭīkā, Vārttika). Primary literature has constraints of spontaneity and prosody (e.g. Vedic Mantras) or of extreme brevity (sūtras), which often decide the choice of words. Hence, the elaborations and commentaries become essential tools in understanding and interpreting. Newer books are published from time to time to consolidate the broad understanding of a field so far, to communicate the state of the art or as an expression of creative genius. These books follow a structure similar to the contemporary practice of literature survey, followed by results and discussion. The books which get accepted by peers are commented upon extensively. All of these, including primary sources and their commentaries, are broadly accepted as primary literature for our purpose. They are mostly in verse form or dense prose. A common characteristic of this primary literature is that it needs further morphological analyses before modern NLP can reliably interpret it.

A second class of literature are books authored in modern times, in the form of translations, analyses or scholarly articles published in journals. These are mostly in free-running prose that is easily understood by current NLP algorithms. For our purpose, they will be designated as secondary literature. A third category of academic literature is the Indexed data or glossaries comprising Compendia (e.g. Dhātupāṭha [1]), Encyclopaedias (e.g. Śabdakalpadruma [2], Vācaspatyam [3]) and Dictionaries (Nighaṇṭus, e.g. Amarakośa [4]). While glossaries are compiled even in modern NLP, in IKS discourse, they have a special formal role. It is a common practice in IKS to start the analysis of a concept or a problem statement by first clarifying the various terms of relevance. Discourse is gradually built by layering the meanings of the akṣaras, padas, and vākyas together in a coherent whole. They are also used as primers and for disambiguation. Dictionaries give several senses and connotations of a word with illustrative references for each of the connotations. Within the Dhārā system, the three types of data are stored in separate databases with distinct retrieval modes and algorithms. The three databases and their contents will henceforth be referred to respectively as Verse (V), Book (B) and Dictionary (D).

Information Structure of IKS Databases

Design of Information storage and retrieval systems for IKS must be cognisant of the difference between the properties of V, B and D databases and their contents. V needs significant morphological processing before it can be tokenised and processed further. B is better processed with NLP techniques than V. D is usually well structured, often with syntax that can be easily parsed with regular expressions. In spite of different properties, the contents of D, V and B are mappable to each other. If B is a translation of the same textual variant as V, then the mapping is one-to-one. If B is a summarised translation or an expanded translation, the mapping of textual chunks can be one-to-many or many-to-one. Within V itself, textual variants can be mapped to each other. Elements of V are to be stored and read along with their annotations in the form of commentaries. Verses in V are understood by pooling together annotations, mapped elements within V (textual variants) and translations (in B) and word meanings (in D). A reading of B frequently requires support from the primary reference from V.

The richness and flexibility of Indic languages generate several pseudonyms and aliases for persons and objects, leading to problems in text understanding of V. However, B and D usually do not have these issues. D can be used to disambiguate V.

Citations and references are today commonly denoted in the (Author, Year) format. However, within IKS, references are made by quoting a partial verse (typically the opening phrase).

Such references to verses in V are extensively found within B and D. Within B, the references often follow multiple styles - either the partial verse style or Text-Chapter-VerseNumber (e.g. Ṛg Veda 3.1.1). Due to the presence of several textual variants, uniquely mapping references to verses is a challenge.

Due to these properties, the typical operations performed on these databases vary. Fuzzy search for partially corrupted phrases is frequently performed on V for dereferencing. D is referenced directly as a lookup and indirectly to provide additional context. In most cases, it is retrieved using the primary key. B is more amenable to Question-answer systems. But are often considered insufficient within the IKS domain unless the mapped entities from V and D are also furnished as references.

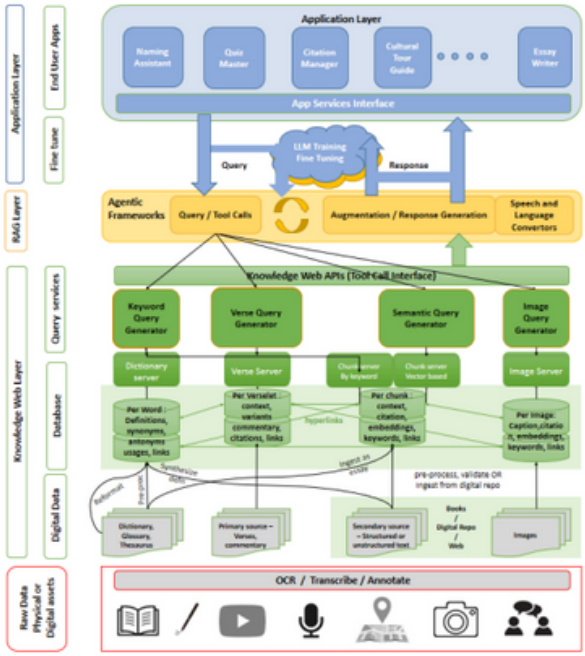

Fig 1. Architecture diagram of Project Dhārā

Structure of Database and retrieval APIs

To exploit the inherent structure of the IKS databases and address the needs of the IKS domains, we create a set of databases D, V and B with mappings between elements. The set of

access methods on each of them is different. The primary access method for the V database is a fuzzy search that outputs the full original verse that contains an input partial verse (possibly corrupted). The semantic searches are the primary operation on B. Primary operation on Dictionary D is a fuzzy lookup with the key. V and B databases also support traversal to next and previous elements like a list, or a hierarchical traversal across Category->Book->Chapter->Page/Verse. Automated glossaries are also created on the contents of B to improve factual question-answer accuracy [5].

Currently, our database contain about 10+ lakh word definitions curated from dictionaries and compendia and 7+ lakh place names in D, 8+ lakh verses curated from primary texts in V and 95+ lakh paragraphs of text curated from books in B. Project Dhārā is an ongoing effort and the numbers are growing rapidly each passing day. The mappings between the databases are being developed in a phased manner.

Using the mapped databases B, D and V, we provide a variety of services as APIs, which include semantic search and RAGs. Open LLMs running locally on our servers perform the generation based on the retrieved contents. A comprehensive set of benchmarks is being developed for various tasks within the purview of IKS.

Conclusion

Project Dhārā is a massive effort to create an information storage and retrieval system for the Indic Knowledge System that takes cognisance of its unique properties and mappings. Using this system, the project also provides information retrieval as a service in multiple modes, viz, Semantic Search and Retrieval Augmented Generation. We also develop example applications and benchmarks to demonstrate the utility and power of the system. In the future, we envisage these systems becoming gold standards for a variety of applications and tasks in IKS. This system can also be distilled to build foundational models.

References

[1] Pāṇinī Maharṣi. Dhātupāṭha. Chaukhamba Amarabharati Prakashan, 2008. [2] Amarasiṃha. Amarakośa. New Bharatiya Book Corporation, 2023.

[3] Rādhākāntadeva, Vasu, and Haricaraṇa Vasu, editors. Śabdakalpadrumaḥ. Vol. 1, Vārāṇasī, Caukhambā Saṃskṛta Sīrīja Āphisa, 1886.

[4] Bhaṭṭācārya, Tārānātha Vaidyanātha, editor. Vācaspatiyam: Bṛhat Saṃskṛtābhidhānam. Vol. 5, Vārāṇasī, Caukhambā Saṃskṛta Sīrīja Āphisa, 1873.

[5] Lita, Lucian Vlad, et al. "Qualitative dimensions in question answering: Extending the definitional QA task." Proceedings of the national conference on artificial intelligence. Vol. 20. No. 4. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005

> *Project Dhārā is a massive effort to create an information storage and retrieval system for the Indic Knowledge System that takes cognisance of its unique properties and mappings. Using this system, the project also provides information retrieval as a service in multiple modes, viz, Semantic Search and Retrieval Augmented Generation*

Mr Avinash H N
Research scholar
Dept of Heritage Science and Technology