ROBUST HADITH IR USING KNOWLEDGE-GRAPHS AND SEMANTIC-SIMILARITY CLASSIFICATION

Omar Shafie, Kareem Darwish, and Bernard J. Jansen

Hamad Bin Khalifa University

ABSTRACT

Hadith is the term used to describe the narration of the sayings and actions of Prophet Mohammad (p.b.u.h.). The study of Hadith can be modeled into a pipeline of tasks performed on a collection of textual data. Although many attempts have been made for developing Hadith search engines, existing solutions are repetitive, text-based, and manually annotated. This research documents 6 Hadith Retrieval methods, discusses their limitations, and introduces 2 methods for robust narrative retrieval. Namely, we address the challenge of user needs by reformulating the problem in a two-fold solution: declarative knowledge-graph querying; and semantic-similarity classification for Takhreej groups retrieving. The classifier was built by fine-tuning an AraBERT transformer model on a 200k pairs sample and scored 90% recall and precision. This work demonstrated how the Hadith Retrieval could be more efficient and insightful with a user-centered methodology, which is an under-explored area with high potential.

KEYWORDS

Hadith, Knowledge-graphs, Arabic, Semantic Similarity

1. INTRODUCTION

For more than 1,200 years, Hadith investigation methods have branched to multiple specialized study fields known as Hadith Sciences علوم الحديث. Islam highly values Hadith as it is considered the second source of legislation after the Quran. Hadith is the oral tradition of the prophet of Islam, and it was not transmitted in written form for a long time after his death. Therefore, a new discipline emerged to preserve the prophetic traditions and detect forged narrations. Each Hadith consists of two parts: the narrative text of the Hadith (*Matn*) and the chain of its narrators (*Isnad*). Therefore the chain of narrators of each Hadith is essential to be evaluated. Hadith scholars developed precise methodologies to examine and investigate the accuracy of a given historical narration.

Figure 1 shows an example of a chain of narrators tree of a Hadith narrated by Imam Muslim: "When a man dies, his acts come to an end, but three, recurring charity, or knowledge (by which people) benefit, or a pious son, who prays for him (for the deceased)."

David C. Wyld et al. (Eds): SOEN, SIPP, PDTCA, ITE, CCSIT, NLPCL, DaKM, BIGML, AISC -2023
pp. 169-182, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131215

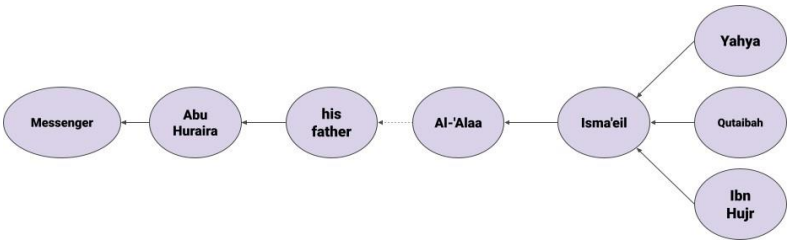

Figure 1: A tree diagram of a chain of individuals narrated by Imam Muslim. Read from right to left: Narrated Yahya, Qutaibah (Ibn Sa'eed), and Ibn Hujr, they said (to Imam Muslim): Narrated Isma'eil (Ibn Ja'afar) from Al-'Alaa, from his father, from Abu Huraira reported Allah's Messenger.

Intuitively, a hadith instance has a 2-dimensional representation; the chain of narrators as a directed graph [1], and the narrative- text as a natural language utterance(s). However, while many Hadith tasks match known Natural Language Processing (NLP) problems [2], the existing Hadith software solutions are limited. For example, existing retrieval solutions are typically more text-based, with limited support for retrieving a chain of narrators patterns. Also, most Hadith datasets were manually annotated with every new software release, rather than training on previous datasets to automate for better scalability. Therefore, investing in data science methods for Hadith applications is promising.

Typically, the Hadith investigation concludes with the acceptance or rejection of a report after a sequence of examination steps, such as determining the narrators' grades and assessing the connection between narrators in the narration graph. For example, every narrator in the chain must have met the narrator before them in the chain. Otherwise, then this Hadith will be rejected due to the disconnectedness of the chain of transmission. Even if the chain is sound and accepted, the Hadith might be graded as "weak" due to its odd text or its contradiction with other authentic Hadiths or Quran. Besides that, the individual characteristics of these narrators are taken into account for evaluating the soundness of the chain. For example, if a narrator is a known liar, then all Hadith with him in its chain of narrators will be rejected. Such evaluation tasks are labor-intensive and quite challenging due to the volume of the Hadith library with hundreds of thousands of narration instances [2,3]. These facts make Hadith a special type of data that interests the research community as Information Retrieval (IR) is perhaps the most significant problem in the entire process.

The focus problem of this paper is on one particular application of IR, where Hadith scholars need to collect the entire Takhreej group (i.e., all reported instances of a given saying, action, or story) to reinforce and maximize information gain before the investigation phase begins [4]. Moreover, as the chain of narrators is not limited to narrations of the Prophet (p.b.u.h), it is common to find Isnadic narrations in books of Arabic Literature, Tafseer, History, and other early Islamic resources. This observation shows that the chain of narrators was a general referencing practice [5].

Contributions in narrative-information retrieval research can impact other digital humanities disciplines, both Islamic and non- Islamic (Some applications such as Sefaria.org support interfaith studies rely on analyzing the citation maps). Moreover, as Hadith science has received significant attention over hundreds of years, the research community could employ some of its principles in applications such as Authorship Attribution, Citation Networks, Knowledge Transmission, Information Representation, and Fact-Checking. This paper addresses the limitations of the Hadith IR problem by implementing an alternative that is more robust than a text-based search.

2. BACKGROUND

In the Hadith domain, we can start from three perspectives to define user needs. The first perspective is a traditional Islamic view of the domain that evolved over a thousand years. Which involves understanding the resources of traditional books and their authorship purposes. Secondly, it is vital to analyze modern practices of study- ing Hadiths in Academic Humanities contexts and their tasks. The third perspective is provided in the literature of Computer Science about Hadith research, where we scanned for relevant works and explored the collective efforts of the research community in this domain. This section explores all three perspectives to construct an understanding of the context of the user.

2.1. Traditional Hadith Sciences Resources

Hadith Science is defined as the set of rules that Hadith scholars have formed to evaluate the narrative-text and the chain of narrators. The purpose of this rule-set is to determine the narrations' acceptance or rejection constructively. These rules implicitly describe many tasks performed on both narrative-text and chain of narrators. For example, consider the definition of the *Sahih* صحيح (Highest level of trustworthiness to confirm the authenticity of the narration) hadiths. A narration is defined as Sahih only after it satisfies five qualifications shown in Table 1. In case a narration fails to satisfy any of the five conditions, it is not considered authentic.

Table 1. Qualifications of Authentic Hadiths

Qualification	**Description**
Chain of narrators Connection اتصال الإسناد	To establish that every narrator in the chain of narrators has received the Hadith directly from the individual next in the sequence (i.e., no hidden links).
Narrators' Integrity عدالة الرواة	To establish that every narrator in the chain of narrators is not accused of telling lies nor has the motive to fabricate the narration willingly.
Narrators' Credibility ضبط الرواة	To establish that every narrator in the chain of narrators is not accused of poor memorization nor is known to make mistakes in reporting commonly.
Narration Divergence عدم الشذوذ	To establish that the narration is not questionable for being reported from a single source when the expectation is to have a reinforcing instance from a different route.
Narration Disorder عدم العلة	To establish that the narration is not questionable for a hidden flaw, such as a conflict with other narrations about the same story.

The first step to verify a Hadith is to collect all of its instances. However, as the Hadith chain of narrators may be scattered among many multi-volume books and indexes, it requires domain expertise to identify and extract all story instances. This difficulty results from the growth of the Hadith library as one of the most extensive Islamic books genres by the 15th century. The process of identifying sources of a given narration and extracting all of its instances is known as Al-Takhreej التخريج. The objective of the Takhreej process is to verify referencing sources that documented a narration instance across the research scope. We will be referring to this problem as Hadith IR for the rest of this paper.

After researchers have collected all the available instances of the narration, they limit the investigation to that collection. A researcher's investigation is vulnerable to all kinds of mistakes by not completing the Hadith IR process [5]. Building a conclusion on a single instance is short-sighted, and we might find a counter-proof in another instance of the same Hadith. Therefore, a

researcher must complete the previous step to maximize their knowledge before they start the investigation.

2.2. Hadith in Humanities Academia

Authors of Hadith collections structured their books in different ways for efficient access to the information. For example, researchers can find instances of a Hadith reported by narrator X about topic T in chapter T in books sorted by topic and found under chapter X in books that are sorted by a narrator [3]. These authorship styles are an implicit structuring of the two retrieval methods used by Hadith scholars, i.e., retrieval using the chain of narrators and retrieval using the narrative-text.

When researchers find a new instance that reveals additional information about some narrator, it is possible to repeat the process for the new narrator. Likewise, if a narration has an additional statement in the narrative-text with relevance to a different topic, it would expand the keyword scope. This process makes Hadith IR a greedy process that expands the information space of the collection of retrieved instances.

Table 2 summarizes the main approaches Hadith researchers attempt to retrieve the instances of narration that contribute to answering their research question. These approaches provide insights into defining the querying needs and the limitations

Table 2. A summary of retrieval tasks and when researchers can use them.

Task Title	**Availability**
Task 1: Retrieval vianarrative-text topic	Only possible if narration has clear relevance to a well-served topic, such as legal jurisprudence. It can be misleading.
Task 2: Retrieval via the first statement of the Hadith	Only possible if narration starts with the same sentence in all other instances.
Task 3: Retrieval via a keyword in narrative-text	Only possible if narration has an identifiable rare vocabulary.
Task 4: Retrieval via narrator in the chain	Only possible if the narrator is a companion, book-author, or biographic books cited the narration.
Task 5: Retrieval via a unique property	Only possible if narration has an identifiable rare property. Table 3 provides a list of some known Hadith attributes and examples of the books dedicated to that property.
Task 6: Retrieval via full manual scan	Only possible on small project scopes.

2.3. Hadith in Computational Research

Despite many software development initiatives, a general challenge that faces computational research for Hadith is the lack of available dataset [6,2,7]. As a result, the efforts that support Hadith computation are scattered between research and the commercial market. Even among commercial projects, one can see that there are many repetitive manual efforts [8,2], and new works are not primarily novel in terms of features but mainly data volume or quality.

Here, we document seven active commercial software and present a comparison of the features they cover, ordered chronologically. We are interested in documenting: data entry, data annotation, data revision, data resources, data integration (topics, judgment, narrators), features (Hadith retrieval, visualization, text comparison, and custom dataset), languages support, and API availability.

Table 3 summarizes the software in terms of data and features. We can see that the focus on many of these systems is in the data provided. However, many data issues were found in these manually collected datasets, transmission-typologies have been mostly ignored, and they generally lack API or direct method for providing the datasets even when governments fund the project. Missing feature examples are limitations of searching methods in text search and chain pattern searches, simple visualization generation, and simple narrative-text analysis. The importance of the systems can be appreciated by gathering statistics about how many visits they get and download, and web traffic, students-surveying.

Table 3. A summary of Hadith software initiatives.

System	Data	Features
Al-Maktaba Al-Shamela	400 Hadith books	Auto-Isnad-Analysis, and basic text search.
Gawamie' Al-Kalem	Hadith judgment comments of about 200 scholars from 600 books	Basic text search.
Al-Dorar Al-Saneyyah	900 books annotated for 45000 narrators	chain segment search, basic text search.
Islamweb Hadith Library	20 Hadith book	Basic text search, and tree diagrams drawings.
Sunnah.com	7 of the Nine-Books collection	Only an exact match search
Arees Institute	Six-Books dataset	Only an exact match search, andnarrators timeline overlaps
Ifta' SunnahProject	33 major Hadith books, 18000 narrators linked to more than 86,500 biographic documents	chain segment search, basic text search,tree diagrams drawings, Grouped narrative-text feature

3. LITERATURE REVIEW

As documented in [3], MM Azmi's pioneering efforts in Hadith computational research since 1978 are well recognized due to his known reputation of being a Hadith scholar. Later, Bounhas illustrated how M.Sc. and Ph.D. contributions led the recent publication efforts in this area [2]. We managed to find 3 survey papers [2,3,9] that mapped most contributions in the domain. Noticeably, all authors have concluded the difficulty of comparing research benchmarks as no unified corpora is available.

While the research was heavily concentrated on the classification of the judgment of the Hadith [3], the research community did not serve other Hadith tasks such as Hadith Visualization, Information Retrieval, and biographical analysis.

In this paper, we focus on Hadith IR problems that are relevant to search-engine design, i.e., for a text-based query input at run time, return a ranked list of the matching Hadiths of relevance to the query. We group works into four collections by functions of our desired search-engine: Topic Clustering, Hadith Querying, Metadata Search, and Hadith Recognition and Extraction.

3.1. Topic Clustering

Hadith collections clustered by common topics can be useful to find the rest of the instances of theTakhreej group. For example, for a given Hadith, the search-engine is needed to retrieve all hadiths that share the same topic. Since most systems in Table 3 have attempted to label the topics manually, we can use this data to train a model to recommend/generate topic labels.
Authors of [10] replicated the following works that relied on term frequency methods to confirm

the accuracy of the reported results. They evaluated the results on 3150 Hadiths from Bukhari over14 categories. In [11], the authors used TF-IDF on 8 chapters of Bukhari with 15 hadiths from each chapter and 5 for testing. They reported 83% while in [10] it was reported 70% accuracy using the replication of the method. Authors of [12 –14] attempted to classify 453 hadiths from 14 subjects. Their best score used Artificial Neural Networks (ANN) with dictionary-lookup stemming [12]. They reported a score of 88% F1-measure. These results were confirmed using a bigger scale dataset of [10] with 90% accuracy, while the replication of the methods of [14] performed 94% accuracy. Authors of [15] used TF-IDF and Rachio algorithm to classify 1350 hadiths over eight subjects and reported 67% precision. Replication of [10] reported 83% accuracy.

On the other hand, authors of [16] used chapters of Bukhari as the class names for a topic classification problem. They constructed a similarity coefficient table for each word in the collection of 13 chapters with 1321 hadiths and reported an overall better recall and precision against word-based classification.

The work of [17] of detecting similarity is unique in the use of a Shiite dataset from Wasa'il al-Shi'ah collection. They reported a 97% F-score score of their classifier by calculating the cosine similarity between pairs using TF-IDF of n-gram of 3 characters representation. However, their experiment was conducted only on 400 instances from 100 groups, and it was not clear how the instances were selected and labeled as similar. As described in [2],
the sample size is not representative for generalizing such results.

3.2. Hadith Querying

Despite the uniqueness of the hadith chain-narrative-text structure, however, not many options have been proposed that support advanced querying about the chain of narrators. In Hadith IR, the chain is what makes a narrative interesting, therefore, allowing the users to express querying aboutthe chain is a must-have feature.

Some authors [18] used regular expressions to construct a more flexible query processing search on Android devices. However, regular expressions were only part of the backend implementation rather than provided directly to the user. The search options provided include searching using the root, controlling the order of the words, and searching for documents not having words. Such methods were not evaluated nor tested by users from the domain.

There are more advanced designs of IR proposed in the literature such as [19] of using TF-IDF approach or the use of regular expressions to build root-based search and flexible querying such as in [18]. However, most such tools have not been implemented for Hadith search engines in software support [18]. Currently, existing search engines have various text-matchingconfigurations that are primarily deterministic. Hadith similarity and topic classification may improve the results as well.

While two retrieval methods are implicit in Hadith books' structures (i.e., by a narrator and by topic), surprisingly, searching by chain-pattern is widely ignored in the existing solutions. Only Ifta' and Gawamie' Al-Kalem systems provided a way to construct a chain sequence and search for Hadith instances that match it. However, it is limited to exact pattern matching using a user interface.

3.3. Metadata Search

It is common that users are interested in particular properties of the narrator or the text rather than the narrator/text itself. For example, Hadiths that are entirely narrated by Iraq-based narrators along the chain. Or Hadiths that are accepted by one Imam while rejected by another. For this problem, users are interested in data about the narration, i.e., Metadata. However, narrative-text and chain properties require even more sophisticated Hadith parsing and annotation. Current manually annotated software such as Ifta' has included some cases that can be used for training and testing models. Some authors [20] tested their ontology construction using a Description Logic Query which is easy to use to find within the range of the defined Ontology. No further works can be found in the literature on this problem.

3.4. Hadith Recognition and Extraction

Due to the large size of the isnadic library and the spread of a chain of narrators in many Islamic books, recognition and extraction of narrations from an available book to prepare a dataset for processing will expand our resources. This is essential for scalability of Hadith processing in general, where search-engine designs need to be flexible for the IR scope of the user. Further, all previous systems in Table 3 have been manually annotated on a selection of books, with hundreds of books receiving little or no support. For example, Ifta' system was annotated by 200 users over six years. This problem is even more exciting for Hadith recognition off the web or social media platforms. One challenge to this problem would be distinguishing between the author's narrations and those the author has copied from other sources with/without explicit citation. This problem is essential for scalability of Hadith processing in general.

Some authors [21] proposed a crawling system to extract hadith texts from web pages and lookup their degrees from a pre-defined dataset. They tested their system on 63 hadiths from 5 websites and reported 50.7% and 38.6% for recall and precision, respectively. Similarly, authors of [22] shared how they identified chains of narrators in their large OpentITI corpus. These efforts can be used to gather more Hadith data in new platforms.

4. METHODOLOGY

To best support the user experience, we had gathered user requirements using the surveyed works in computational research in the literature review. Then, we conducted semi-structured interviews with 3 domain experts to construct scenarios and use-cases of their experiences using existing software solutions. Users are considered expert faculty in 3 universities of Islamic Studies that are lecturing Hadith with more than 20 years of experience in the domain. In addition, we observed 10 users (different levels of Hadith expertise) performing their tasks using available software. Further, we built a comparison matrix of the limitations of existing software that supports the hadith domain summarized in Table 3. Lastly, we used a task-oriented approach to evaluate our proposed solutions performance in solving the scenarios of the user requirements.

4.1. Defining User Need

A widely used definition of information retrieval is "obtaining material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)."[23 , p. 1]. While Hadith in its textual form fits this definition for unstructured data, Hadith embeds a semi-structured dimension within its most exciting part, i.e., the chain of narrators [2]. Further, from our user-centric methodology, we have seen that 'information need' defined by the user sometimes interchangeably combines/switches between

these two dimensions. Therefore, representing Hadith querying for the Hadith IR tasks needs to consider both parts, the chain of narrators as a directed graph and the Matn as textual data, which produces an unusual problem of combining two data types in one IR application. However, existing methods in Table 2 are limited in being not able to serve for the relational user needs.

Unlike traditional text retrieval problems, a suitable Hadith search engine must be able to: 1) allow the user to express multi-dimensional needs as a query and 2) Retrieve Takhreej groups from unannotated datasets.

The presented framework in this research satisfies this by solving the corresponding sub-tasks: 1) robust relational expressions querying and 2) retrieval of Takhreej groups, which is an alternative to all six methods of retrieval that is more powerful.

4.2. Querying Hadith Using Knowledge Graph

Table 4. Example of common query statements inspired by traditional Hadith resources.

User Need	Statements	Cypher Query

To retrieve all **hadiths** that follow a given chain of narrators pattern	Ibn-Rajab: 'There are about 6or 7 narrations acceptable in the path of Sufyan Ibn Uyaynah from Al-Zuhri fromAnas.'
	ابن رجب : سفيان بن عيينة عن الزهري عن أنس عن النبي ٧ أو ٦
		`MATCH path =`
		`(n0:Narrator {n_id:2478}--> (n1:Narrator {n_id:5917}-->(n2:Narrator {n_id:822},`
		`(:Takhreej)<--(:Hadith)-->(i:Isnad)-[r0]->(n0),`
		`(i)-[r1]->(n1),`
		`(i)-[r2]->(n2)`
		`WHERE r0.rank = r1.rank+1 AND r0.rank = r2.rank+2`
		`RETURN path;`

To retrieve all **narrators** that are single sourceof a given Takhreej group	Al-Tabarani: 'Hadith [...] is reported from a single sourceby the narrator [...] in all six books.'
	الطبراني: وتفرد به فلان في الكتبالستة
		`MATCH (t:Takhreej {id:5}<--()-->(i:Isnad),`
		`(n:Narrator)`
		`WITH t, n, collect(i) AS asaneed`
		`WHERE ALL(s in asaneed WHERE (s)-->(n))`
		`RETURN n;`

To retrieve all **chains** that match a givenproperty	Al-Nawawi: 'Isnad of Hadith [...] is entirely narrated by narrators from Al-Kufah' كله
	النووي: إسناده كوفي
		`MATCH (i:Isnad)-->(n:Narrator)`
		`WITH i, collect(n) AS narrators_list`
		`WHERE ALL (n IN narrators_list`
		`WHERE n.places CONTAINS "Kufah")`
		`RETURN i;`

To retrieve all **Takhreej Groups** that match a given property	Ibn-Hajjar: 'Hadith [...] is one of 4 Hadiths where Muslim has a higher rank (i.e., shorter chains) than Bukhari to the same source.' أحد الأحاديث الأربع
	ابن حجر: وهذا التي علا فيها مسلم على البخاري عننفس الشيخ
		`MATCH (b {n_id:5495}<-[b_b]-()<--(h_b:Hadith {book: "Bukhari"}-->(t:Takhreej)`
		`WITH t, min(b_b.rank) AS minb`
		`MATCH (t)<--(h_m:Hadith{book: "Muslim"}-->(i_m:Isnad)-[m_m]->(m {n_id:6116}`
		`WHERE minb > m_m.rank`
		`RETURN DISTINCT t;`

Consider the statements collected from our users' observations in Table 4; user need for these Hadith IR tasks is to verify the correctness of such statements by collecting all Hadiths instances that match the same scenario from traditional resources of Hadiths. Such statements inherently describe a graph relationship among the hadith knowledge domains rather which are hard to express in a text retrieval query. However, we see that most modern Hadith search engines that support retrieval are mostly narrative-text-based, with limited features supporting chain querying.

We propose a robust approach to support queries that express the entire Hadith knowledge-base for all user needs. Knowledge graphs have been one of the most effective forms of information representation [24]. In our exhaustive study of the domain with the user-centric method, we were able to construct a comprehensive knowledge map. This map of available data types enabled us to build a graph database model using a graph DBMS.

Considering the existing software support, we compiled seven data types for Hadith's concepts that can be integrated into a knowledge base for answering research questions. The data types can be found in Table 5. These data types can be modeled into an Entity Relational (ER) database framework shown in Figure 2. Interestingly, there are many possible ways to integrate the data, producing many Hadith visualizations and analysis projects.

Table 5. Hadith Knowledge Data Types and their definition

Domain Knowledge Data Type	**Definition**
Ahadith (plural of Hadith)	Historical narrations documented via an chain of narrators and narrative-text
Biography	Descriptive texts about a narrator that documents their narrative credibility and integrity and documents their educationaljourney.
Categorical Topics	Knowledge base of Islamic studies concepts. Every Hadith is relevant to several topics (for example, prayer)
Degree judgment	Summary conclusion of the study of the Hadith about the degreeof acceptance/authenticity of a collection of Hadiths.
Entities/narrators	Individuals that narrate any Hadith, which is a structured formof the data in the biographic texts.
Fawa'ed (insights)	It is common for researchers in Hadith to keep a collection of 'useful' quotations from their lateral readings of Hadith literature. These collections resemble 'interesting' findings that can be compiled to construct material for further research. Such quotations, also known as (*fawa'ed*) or (*taqyeedat*), are fundamentally unstructured and can be anything within the author's scope of 'interest'.
Groupings/Takhreej	Collection of instances of a report that is most likely the same Hadith from the same top-narrator but different chain ofnarrators or different wordings.

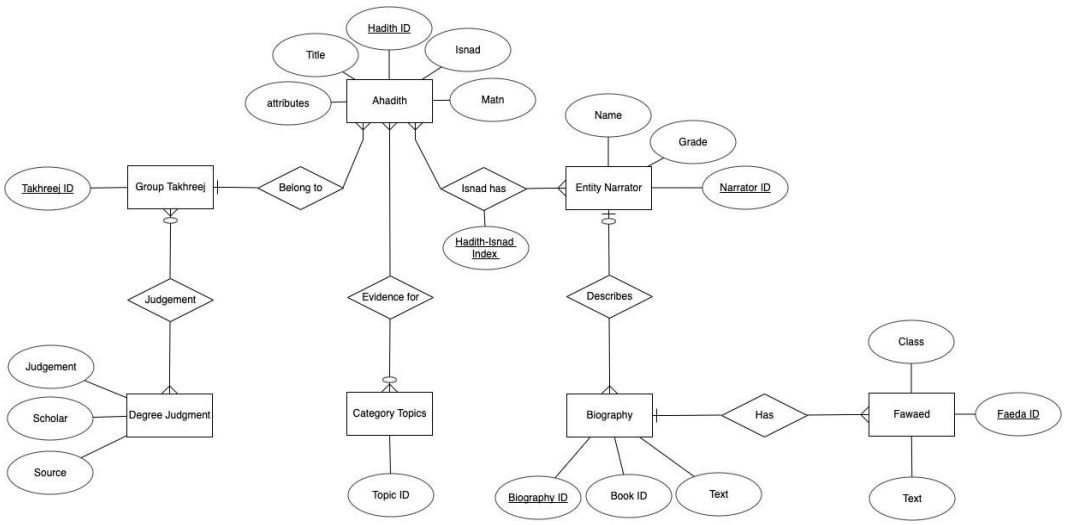

Figure 2: Entity Relational Data Schema of the Hadith datasets

4.3. Querying Takhreej Groups Using Text Pair Classification

Previous works on hadith similarity can be used as a baseline to determine the distance between Hadiths for topic classification and clustering. However, all of the previous works did not attempt a Takhreej group classification variation of the problem, which is the main contribution of this paper regarding Hadith IR. In such variation, the chain of narrators plays a significant factor, as Takhreej groups are defined by the story's narrator (last narrator). Evaluating an experimental text search engine is challenging, requiring manually-collected ground truth results that satisfy the user's need.

Using a graph database is sufficient for most user needs of finding a specific instance of a Hadith, which is the beginning of a typical Hadith IR process. The next natural step is to retrieve all instances of the Hadith that are in the same Takhreej group. Identifying a Takhreej Group is challenging due to the large number of classes, i.e., 10,000 Takhreej groups. Therefore, a multi-class model would perform poorly. Instead, we can train a model to classify a pair of Hadith texts if they are of the same group and then use that model to compute the Takhreej Groups. Notice that the results of this step are static for a given Hadith dataset. Hence, we can perform this task offline for the entire collection. At run-time, users can retrieve all of the instances of any Takhreej Group in a single mouse-click the instances in hand.

The problem of Takhreej Group classification is equivalent to the known binary semantic-similarity classification problem. Therefore, we collected an experimental dataset from Ifta' collection of the Nine-Books. Then, we randomly sampled text pairs, and trained a model to classify whether the given pair are from the same Takhreej Group. In our dataset, we can find more than 500,000 positive pairs of the same Takhreej Group that we can use for training. We can see in Figure 3 that using a traditional method such as cosine similarity of TF-IDF vector representation of the text pairs presents a significant overlap between the positive (same Takhreej) and negative (different Takhreej) pairs. We use this as a baseline, which is quite a poor f1 measure of less than 0.7.

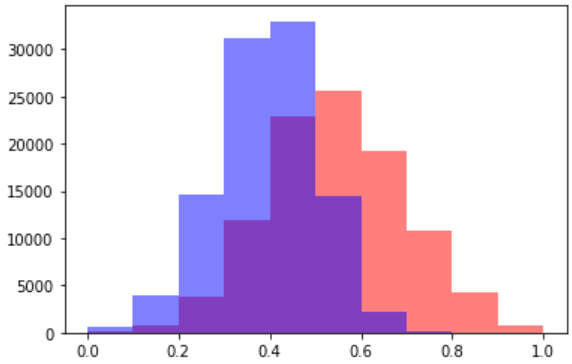

Figure 3: Cosine Similarity of TF-IDF vector representations of Hadith pairs.

In our classification experiment, we used the state-of-the-art NLP, the transfer learning approach. Due to limited computational resources, we sampled around 200,000 pairs of the dataset, evenly distributed between positive and negative classes. Using the trans- former library, we fine-tuned the **bert-base-arabertv02** model on 80% of the dataset and evaluated the performance on the remaining 20%.

5. RESULTS

Our knowledge map implementation supported the knowledge retrieval of The Nine-Books collection of Iftaa', which has annotated chains of narrators from 68,561 hadiths, more than 10,000 Takhreej Groups, and more than 18,800+ narrators with their insights. For our framework, we used Neo4j, a graph DBMS that uses Cypher, a declarative query language. A graph model has been proposed previously for Hadith in [25] but was only for narrators' chain network representation, while we propose that the representation would be used for the entire knowledge domains of Hadith querying. With a query language as expressive as Cypher, we can solve the scenarios in Table 4 with ease.

We presented our work to Hadith students of our user-study group and received good impressions on the capabilities of the new method. However, the main challenge faced by most users is to translate the user's need to Cypher query as they did not have previous experience with similar literacy. At the time of the experiment, Neo4j did not implement ChatGPT interface feature, which allows the user to translate natural language statements into the necessary Cypher.

Then, for identifying Takhreej Groups, our classifier scored 90% f1-measure, recall, and precision after two epochs. Which is establishing a new baseline, compared to only topic classification, a relaxed variant of the problem from the literature, and far better than Cosine Similarity TF-IDF approach.

5.1. Implications and Discussion

In the surveyed tasks of the Hadith IR, six problems are identified for Hadith IR tasks. However, Hadith IR has several unique challenges, such as having 2 data parts represented differently. Further, as seen from the example queries, users' needs in Hadith IR can be more knowledge-based rather than text-based. In our reformulation of Hadith IR problem by building a knowledge map to be represented as a graph DB for expressive querying, this research overcame the limitations of text-based querying by far.

The produced solution utilized the latest trends using knowledge graphs representation. Further, it was complemented by training a state-of-the-art classifier model to retrieve similar Takhreej groups that scored 90% in both recall and precision.

These solutions attempt to solve Hadith IR problems using new ways that overcome the limitations of previous methods. Also, it has established a new problem baseline for Hadith IR. This work is unique as it explores a highly under-explored area of Classical Arabic with very high potential. The datasets/code used for our experiment are the largest in size and are entirely made available online.

6. CONCLUSIONS

Although most efforts of data annotation for the Hadith domain have been repetitive, such efforts offered a tremendous amount of data that can be used to conduct many exciting experiments in the future. The concern here is that these datasets were never annotated to conduct computational research, which indicates that the stakeholders of the domain are yet to be informed about the substantive value of digitization. However, this burden lies on the research community to convince the users with tangible results.

Despite being mapped to known data science problems, many of the tasks of the Hadith domain are not solved with state-of-the-art solutions, some were little explored, and many are not well approached with user needs in mind. In this research, we addressed the Hadith IR problem and proposed robust solutions with a user-centric approach.

With the many data options for Hadith processing, it is unfortunate that no data provider has considered an easy access protocol for researchers. Nevertheless, it would greatly support the research community to evaluate the available data options for data accuracy and integration.

The presented framework for Hadith IR in this research is encouraging to look for more challenging query tasks that combine the 2 data representations of the Hadith. However, perhaps a more appropriate direction would be to simplify the graph and regular expression querying design to adapt to the level of digital literacy of the users.

REFERENCES

[1] Syed Irfan Hyder and S Ghazanfer. Towards a database oriented hadith research using relational, algorithmic and data-warehousing techniques. The Islamic Culture, Quarterly Journal of Shaikh Zayed Islamic Center for Islamic and Arabic Studies, 19:14, 2008.

[2] Ibrahim Bounhas. On the usage of a classical arabic corpus as a language resource: related research and key challenges. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(3):1–45, 2019.

[3] Aqil M Azmi, Abdulaziz O Al-Qabbany, and Amir Hussain. Computational and natural language processing based studies of hadith literature: a survey. Artificial Intelligence Review, 52(2):1369–1414, 2019.

[4] Ibrahim Bounhas, Bilel Elayeb, Fabrice Evrard, and Yahya Slimani. Information reliability evaluation: from arabic storytelling to computer sciences. Journal on Computing and Cultural Heritage (JOCCH), 8(3):1–33, 2015.

[5] Alsharif Hatim Ibn 'Arif Al-'Auni. Muqarrar at-takhreeg we-manhag al-hukm alal-hadith. Markaz Nama' li-l-Buhut wa-d-Dirasat, 2018.

[6] Ahsan Mahmood, Hikmat Ullah Khan, Fawaz K Alarfaj, Muhammad Ramzan, and Mahwish Ilyas. A multilingual datasets repository of the hadith content. International Journal of Advanced Computer Science and Applications, 9(2):165–172, 2018.

[7] Sumaira Saeed, Sania Yousuf, Faiza Khan, and Quratulain Rajput. Social network analysis of hadith narrators. Journal of King Saud University-Computer and Information Sciences, 2021.
[8] Aqil Azmi and Nawaf Al Badia. Mining and visualizing the narration tree of hadiths (prophetic traditions). In Applied Natural Language Processing: Identification, Investigation and Resolution, pages 495–510. IGI Global, 2012.
[9] Saqib Hakak, Amirrudin Kamsin, Wazir Zada Khan, Abubakar Zakari, Muhammad Imran, Khadher bin Ahmad, and Gulshan Amin Gilkar. Digital hadith authentication: Recent advances, open challenges, and future directions. Transactions on Emerging Telecommunications Technologies, page e3977, 2020.
[10] Mohammad Arshi Saloot, Norisma Idris, Rohana Mahmud, Salinah Ja'afar, Dirk Thorleuchter, and Abdullah Gani. Hadith data mining and classification: a comparative analysis. Artificial Intelligence Review, 46(1):113–128, 2016.
[11] Mohammed Naji Al-Kabi, Ghassan Kanaan, Riyad Al-Shalabi, Saja I Al-Sinjilawi, and Ronza S Al-Mustafa. Al-hadith text classifier. Journal of Applied Sciences, 5(3):584–587, 2005.
[12] Fouzi Harrag and Eyas El-Qawasmah. Neural network for arabic text classification. In 2009 Second International Conference on the Applications of Digital Information and Web Technologies, pages 778–783. IEEE, 2009.
[13] Fouzi Harrag, Aboubekeu Hamdi-Cherif, Abdul Malik S Al-Salman, Eyas El-Qawasmeh, et al. Experiments in improvement of arabic information retrieval. In 3rd International Conference on Arabic Language Processing (CITALA), Rabat, Morocco, pages 71–81, 2009.
[14] Fouzi Harrag, Eyas El-Qawasmah, and Abdul Malik S Al-Salman. Stemming as a feature reduction technique for arabic text categorization. In 2011 10th International Symposium on Programming and Systems, pages 128–133. IEEE, 2011.
[15] Manar Alkhatib. Classification of al-hadith al-shareef using data mining algorithm. In European, mediterranean and middle eastern conference on information systems, EMCIS2010, Abu Dhabi, UAE, pages 1–23, 2010.
[16] Khitam Jbara. Knowledge discovery in al-hadith using text classification algorithm. Journal of American Science, 6(11):409–419, 2010.
[17] Hossein Juzi, Ahmed Rabiei Zadeh, Ehsan Barati, and Behrouz Minaei-Bidgoli. A new framework for detecting similar texts in islamic hadith corpora. In Workshop Organizers, page 38, 2012.
[18] Aqil M Azmi, Fahad Alkhalifah, Abdulaziz Alsaeed, and Yasser Barnawi. Using non-conventional search schemes to retrieve hadiths. In The 5th international conference on Arabic language processing (CITALA'14), Oujda, Morocco. http://www.citala.org/citala2014/papers/paper_39.pdf. Accessed, volume 11, 2017.
[19] Fouzi Harrag and Aboubekeur Hamdi-Cherif. Uml modeling of text mining in arabic language and application to the prophetic traditions "hadiths". The 1st international sysmposium on computers and Arabic language and exhibition, KACST & SCS, pages 11–20, 2007.
[20] Rebhi S Baraka and Yehya M Dalloul. Building ijs ontology to support the process of judging hadith isnad. 2014.
[21] Mohammed Q Shatnawi, Qusai Q Abuein, and Omar Darwish. Verification hadith correctness in islamic web pages using information retrieval techniques. In Proceedings of International Conference on Information & Communication Systems, pages 164–167, 2011.
[22] Ryan Muther. Tracking traditions: Identifying isnads in the openiti corpus, Feb 2020. URL https://kitab-project.org/Tracking-Traditions-Identifying-Isnads-in-the-OpenITI-Corpus/.
[23] Christopher D Manning and Prabhakar Raghavan. and schutze, h.[2008] introduction to information retrieval, 2008.
[24] Meenakshi Malhotra and TG Nair. Evolution of knowledge representation and retrieval techniques. International Journal of Intelligent Systems and Applications, 7(7):18, 2015.
[25] Emha Taufiq Luthfi, Nanna Suryana, and Abdulsamad Hasan Basari. A novel graph-based representation for hadith sanad. 8(1.5):355–363, November 2019. doi: 10.30534/ijatcse/2019/5881.52019. URL https://doi.org/10.30534/ijatcse/2019/5881.52019.

AUTHORS

Omar Shafie is a graduate student from the Hamad Bin Khalifa University, where he holds Masters in Data Science and Engineering. Omar is a graduate of Carnegie Mellon University in Qatar wherehe received his Computer Science Bachelorof Science.

Dr. Kareem Darwish is a principal scientistat aiXplain Inc working on efficient human-in-the-loop ML and speech processing. Previously, he was the acting research director of the Arabic Language Technologies group (ALT) at the Qatar Computing Research Institute (QCRI) where he worked on information retrieval, computational social science, and natural language processing. Kareem Darwish worked as a researcher at the Cairo Microsoft Innovation Lab and the IBM Human Language Technologies group in Cairo. He also taught at the German University in Cairo and Cairo University.

Dr. Jim Jansen has 380 or so authored research publications. He is the co-authorof the book, Data-Driven Personas, co-author of Web Search: Public Searching of the Web, co-editor of Handbook of Research on Weblog Analysis, author of Understanding User - Web Interactions ViaWeb Analytics, and author of Understanding Sponsored Search: A Coverage of the Core Elements of Keyword Advertising. Jim is a principal scientist at the Qatar Computing Research Institute (QCRI) and an adjunct professor at the College of Information Sciences and Technology at the Pennsylvania State University. At QCRI, he is actively conducting research in various areas of web analytics. At Penn State, he is activelyinvolved in teaching undergraduate (IT project management, keyword advertising)and graduate courses (searching, retrieval, analytics).