The International Society of Data Scientists
Journal of Data Science and Artificial Intelligence
Vol. 4, No. 1. Online ISSN: 2831-4794

Efficient Vietnamese Name Retrieval using Highly Discriminative N-Grams

Toan Luu*[a], Phan Quoc Hung Mai[b], Xuan Lam Pham[b]
[a]UBS Group AG, Switzerland; [b]School of Technology, National Economics University, Vietnam;

ABSTRACT

Retrieving Vietnamese names from large global databases is crucial for fostering connections within professional Vietnamese communities. However, the use of Latin characters in Vietnamese names often causes ambiguity, as they can resemble names from other countries. This paper introduces Highly Discriminative N-grams (HDNs), a novel query method designed to retrieve Vietnamese names from diverse datasets efficiently. Experimental results show that HDNs significantly outperform traditional raw unigram queries, achieving superior precision, recall, and cost-effectiveness. For instance, HDNs improve precision from 0.02 to 0.79 for Evaluation Dataset 1 (ED1) and from 0.18 to 0.93 for ED2, while reducing query results by 24 times and 3 times, respectively, minimizing the need for costly post-processing. This innovative approach not only enhances the accuracy and efficiency of Vietnamese name retrieval but also provides a scalable solution for distinguishing names in multilingual databases, benefiting name-matching systems worldwide.

Keywords: Information Retrieval, Name-Ethnicity Matching, N-grams

1. INTRODUCTION

1.1 Vietnamese names

According to a recent statistic [1], there are over 5 million overseas Vietnamese worldwide. Identifying Vietnamese names within large global databases proves valuable for various applications and community-driven initiatives. This information helps pinpoint experts across diverse fields, enabling the connection of overseas Vietnamese professionals with research and development activities in Vietnam. Such efforts support both the public and private sectors in attracting global Vietnamese talent, fostering national development, and strengthening Vietnam's integration into the globalized world.

Various public big data platforms, such as LinkedIn for professional connections, Google Scholar, ResearchGate, and Research.com for academic and research networking, can be leveraged for name discovery. Online directories can also help locate services, consultants, and advisors. This information plays a key role in identifying experts from various fields and linking overseas Vietnamese professionals with research and development initiatives in Vietnam. This approach will support both the public and private sectors in attracting Vietnamese talent globally, aiding the country's development and fostering its integration into the globalized world.

However, detecting Vietnamese names is challenging due to their use of Latin characters [2] (which stems from the history of Vietnamese [3]), and their resemblance to names from other cultures (e.g. European names, pinyin Chinese names) [4-6]. For example, while "Nguyen" (a common rendering of the surname Nguyễn or the name Nguyên) is distinctly Vietnamese, other widespread surnames like "Lê," when written without accents, may be mistaken for the French article "Le." Similarly, "Đỗ" can be confused with the English verb "Do." Additionally, names like "Vân" or "Văn," which are uniquely Vietnamese with diacritics, might be misinterpreted as the Dutch name "Van" when written in plain text.

1.2 Name Retrieval

The need for information is an essential part of daily life. To meet the demand for quick and accurate information retrieval, a variety of Information Retrieval (IR) systems have been developed [7-10]. Notable examples include search engines like Google, Bing, and Baidu, which serve as efficient tools for locating relevant web pages based on user queries, making it easier to access information online. However, the scope of IR goes beyond just web page retrieval. With the ever-expanding volume of available data, there is a growing focus in both academia and industry on improving IR systems, making it

*luuvinhtoan@gmail.com; phone +41 79 574 88 04

increasingly challenging to extract the most relevant information from databases [11, 12]. Consequently, information retrieval has garnered significant attention in recent years [13, 14].

Online databases such as Google Scholar, LinkedIn, and ResearchGate are invaluable resources for discovering names and expert profiles, offering access to a vast repository of academic publications, professional connections, and specialized expertise. These platforms play a crucial role in academic research, professional networking, and talent identification across various domains. Google Scholar, for instance, provides researchers with an extensive library of academic papers and citations. For an illustration, **Figure 1** shows the process of querying the names of Vietnamese-origin scientists with the surname Nguyen on the Google Scholar platform (a name that is both common and highly characteristic in Vietnam). ResearchGate, on the other hand, focuses on connecting researchers through shared publications and collaboration opportunities, while LinkedIn serves as a hub for professional resumes, skills, and affiliations. Together, these platforms act as powerful tools for accessing and organizing information about individuals and their contributions to their respective fields.

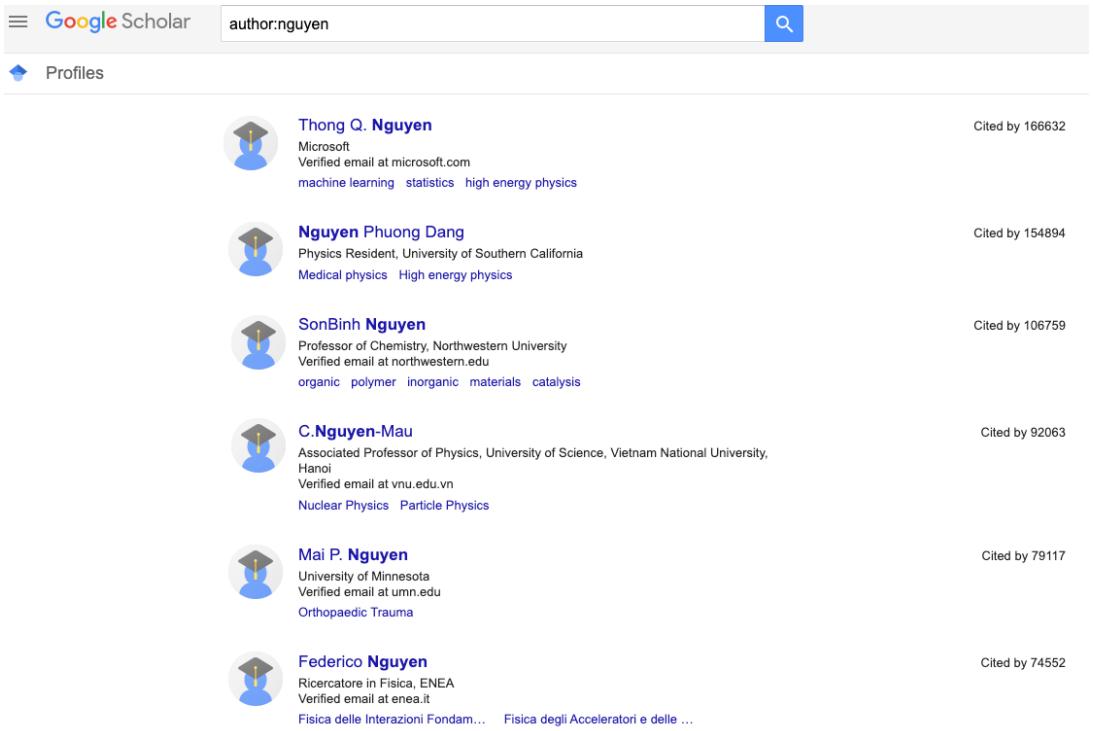


Figure 1. The process of querying profiles on the Google Scholar platform via names

Despite their utility, these databases have significant limitations that hinder their effectiveness in large-scale data collection. One major constraint is the built-in restrictions on the number of queries that can be made and the number of results that can be retrieved [15, 16]**.** These limits are often implemented to manage server loads, prevent abuse, and protect user privacy. However, they pose significant challenges for researchers and data scientists aiming to extract comprehensive information. For example, attempting to compile a complete list of Vietnamese names from these databases becomes infeasible due to query caps and truncated result sets. Even when relevant profiles exist, they may not appear in the returned results, especially for complex or exhaustive queries [17, 18].

This issue becomes even more pronounced when using APIs provided by these platforms. APIs are essential tools for automating searches and integrating database information into workflows, but they typically come with stringent rate limits, often enforcing maximum daily or hourly query thresholds. For example, Google Search Console API limits 1,200 queries per minute [19]. Google Scholar's unofficial API or similar third-party services are known for their restrictive policies, while LinkedIn's API access requires approval and enforces strict usage limits at a maximum of 100,000 calls per day [20].

These limitations significantly hinder the ability to perform large-scale data extraction or conduct research that demands high-volume queries.

Another drawback of some search engines, including Google and Bing, is their accuracy, which is one of the reasons why these platforms are considered to be limited, despite their great applicability [21, 22]. It so happens that when posting content, many platforms do not focus on its accuracy but on its context's relevance to the audience, for example, how many people search for it, its keywords, or even how well the material is presented from an SEO point of view. Additionally, search engines have a hard time with queries that are not specific, or for that matter, with vague or complex queries, which sometimes lead to outcomes that are not quite what was expected, or intended, which diminishes their efficiency in certain circumstances. Specifically, the problem is compounded by the fact that search engines within these platforms often prioritize relevance over exhaustiveness [23, 24]. Algorithms are designed to rank and return results based on perceived relevance, popularity, or keyword matching, which may inadvertently exclude less prominent but equally relevant profiles. For culturally specific or nuanced searches, such as identifying Vietnamese names across global databases, these algorithms may struggle to deliver accurate or complete results. Names with diverse spelling variations, transliterations, or regional influences often fall through the cracks, leaving gaps in the extracted data.

Many previous studies have explored the use of N-grams for efficient querying [25-27]. However, these methods often rely on complicated approaches such as building Word2Vec [28], bag-of-words [29] models, and machine learning, deep learning approaches like neural networks, k-nearest neighbors, latent Dirichlet allocation (LDA) [29-31]. These approaches can be computationally expensive and challenging to scale, making them less practical for large databases and real-time applications. While effective for capturing context in long and topic-based queries, these approaches are less suitable for short queries, such as person names, where a more efficient and lightweight method is needed. This paper introduces an approach that utilizes Highly Discriminative N-Grams (HDNs) to improve the retrieval of Vietnamese names from large global databases.

This method is inspired by the research by Podnar et al. [32], which utilizes Highly Discriminative Keys (HDK) to efficiently query documents within a P2P search engine system. Unlike existing query methods, HDNs are designed to capture statistical and probabilistic patterns that distinguish Vietnamese names from non-Vietnamese names via N-grams. A token or token pair qualifies as an HDN if its frequency in positive name examples surpasses a predefined threshold (e.g., 10) and its confidence score—derived from the estimated probability of it being a Vietnamese name—exceeds a high threshold (e.g., 0.995). To enhance discriminability, the confidence score penalizes tokens frequently found in non-Vietnamese names, reducing the likelihood of irrelevant matches. Finally, N-grams that are determined as Highly Discriminative are used for queries. This approach significantly improves precision while maintaining high recall, reducing the need for costly post-processing. Additionally, we leverage Elasticsearch as the retrieval framework in our experiment, ensuring scalable and efficient search capabilities.

In general, this work aims to make the following contributions:

- Introduce and propose an efficient method for constructing HD N-Grams, applicable to various tasks such as querying, classification of Vietnamese name.
- Present an overview statistical analysis of linguistic tokens in Vietnamese names.
- Contribute to open data community a data set including Vietnamese name features.

2. LITERATURE REVIEW

2.1 Information Retrieval Optimization

As shown in **Figure 2**, an information retrieval system consists of several key components. First is crawling, where a crawler retrieves documents by following links, such as on a specific website for targeted searches [33]. Next is indexing, where documents are organized into a searchable structure. Users then input queries in the query representation phase, enabling

the system to search the index and return ranked results based on relevance. Finally, users can provide feedback to refine the search engine's performance [34].

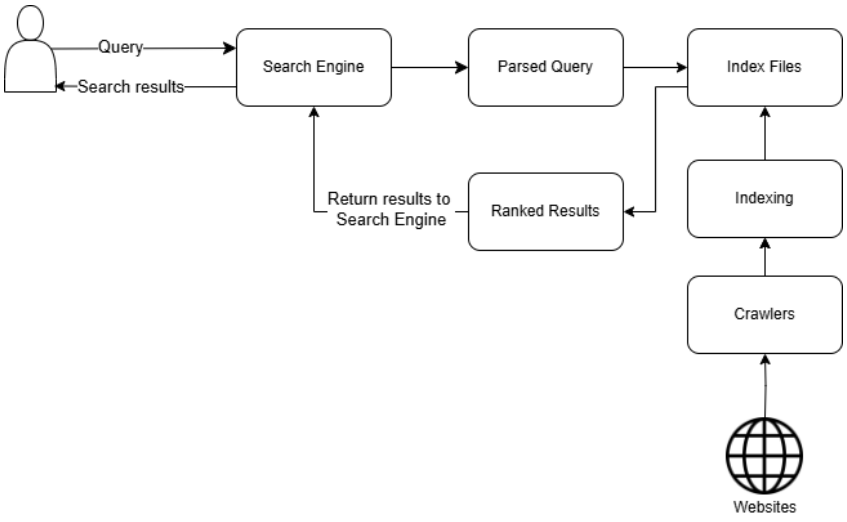


Figure 2. Information Retrieval System

Modern information retrieval systems aim to deliver the most relevant information in response to user queries through two primary stages: retrieval and ranking [35]. In the retrieval stage, the system identifies an initial set of documents potentially relevant to the query, focusing on efficiency. Traditional methods like the Boolean model [36], vector space model, Latent Semantic Indexing [37], and Latent Dirichlet Allocation [38] are commonly used, alongside modern pre-trained models like BERT [39].

The ranking stage refines this initial set by reassessing document relevance based on similarity scores and adjusting their order using advanced algorithms. While retrieval emphasizes efficiency, ranking prioritizes effectiveness. Models like BM25 are often employed for initial retrieval [40], while more sophisticated deep-learning ranking models [41, 42] leverage techniques like reinforcement learning, contextual embeddings, and attention mechanisms to improve result accuracy and relevance.

Furthermore, the integration of Large Language Models (LLMs) has significantly advanced modern IR systems recently, enriching various components such as query rewriting, retrieval, reranking, and reader modules. Practical applications, like New Bing, utilize LLMs to extract information from diverse web pages and condense it into concise, user-specific summaries, enhancing the overall search experience [43]. In research, LLMs have improved query understanding by addressing ambiguous or complex queries, enabling more accurate intent identification. They have also enhanced retrieval accuracy through context-aware matching between queries and documents and refined reranking by considering linguistic nuances to reorder results effectively [11, 44, 45]. Moreover, the incorporation of reader modules allows IR systems to generate comprehensive, informative responses rather than simple document lists. These advancements have transformed IR systems into more precise, personalized, and user-centric tools, reshaping how users engage with information and paving the way for future innovations in search technology [46-48].

However, deep-learning and LLM techniques face several challenges in their application. Deep learning methods require vast amounts of data and computational resources, driving the development of more advanced algorithms to meet these demands and improve NLP performance [49]. LLMs, moreover, encounter issues like concept drift, where their extensive knowledge base can introduce irrelevant or redundant content, potentially diluting the focus of queries. Maintaining a balance between enhancing and preserving query intent is essential for effective IR [50]. Additionally, studies reveal a negative correlation between retrieval performance and query expansion benefits, where expansions help weaker models but can harm stronger ones. This underscores the need for strategic application of expansions, considering model strength and dataset characteristics [51].

Many research, for decades, have shown significant attention to the retrieval stage, with most efforts focused on increasing retrieval accuracy and reducing retrieval costs, which leads to fewer documents being retrieved [52-55]. This is because the primary challenge in information retrieval lies in the gap between how a user expresses the information they are looking

for and how an IR system presents the information. This can be seen as a mismatch between the user's and the system's vocabularies. Furthermore, users often face difficulties in precisely defining the information they need, due to limitations in their ability to articulate their requirements. Language ambiguities and uncertainties also contribute to the challenges users encounter during the retrieval process [56].

2.2 N-gram Models

In natural language processing, n-grams are sequences of *n* consecutive words or characters from a text, used to model language patterns and dependencies. They help analyze text structure, predict the next word, and optimize queries by capturing the statistical relationships between tokens in a lightweight and efficient manner [57].

In Name-Ethnicity Classification works, Treeratpituk et al. [58] explored various features, including basic n-grams and external techniques like Double Metaphone and Soundex, demonstrating their contribution to robust name classification. Similarly, Lee et al. [59] showed that n-gram embeddings effectively capture character-level features for name classification tasks. Other studies [60, 61], also report high performance using n-grams in similar tasks. These findings highlight the versatility and effectiveness of n-grams in capturing complex patterns within textual data, particularly for name classification.

For optimizing queries in IR systems, N-Grams models also show their efficiency in recent studies. For instance, Aliwy et al. [62] incorporate several features for the Named Entity Recognition (NER) system, including n-grams at the letter level (from 1 to 3) to capture word prefixes and suffixes, as well as word-level n-grams. These features, along with others like word type (letters or alphanumeric), part-of-speech (POS) tags for the current and neighboring words, and the presence of gazetteers, contribute to improving the accuracy of the named entity recognition process. The inclusion of these n-gram features enhances the system's ability to identify and classify entities effectively, especially in the context of word structure and surrounding information. Fonseca et al. [63] explore the application of machine learning techniques to identify researchers' main areas of expertise using numerical representations of their scientific production titles. By utilizing a TF-IDF character n-gram representation for the titles, the authors achieve an accuracy of 95.91%, surpassing current state-of-the-art results in the task of classifying expertise areas, which are crucial for verifying the quality of data in platforms like Lattes. N-gram models have also been researched and integrated into query optimization and related fields in IR systems [64-67].

Our research focuses on leveraging N-Grams to optimize queries before sending requests to search engines. By relying on the statistical distribution of unigrams and bigrams in names, this approach avoids the high computational cost associated with deep learning techniques. This lightweight method ensures efficiency while maintaining the effectiveness of query optimization, making it a practical solution for systems with limited resources.

3. APPROACH

3.1 Highly Discriminative N-Grams

Based on our analysis of a corpus containing both Vietnamese and international names, we identified certain tokens that are unique to Vietnamese names. For instance, names like "Nguyen," "Huyen," and "Phuong" are rarely found in other countries and can be considered highly discriminative unigrams. Conversely, some unigrams are common in both Vietnamese and foreign names and are non-discriminative by themselves. However, when combined, these unigrams can form highly discriminative bigrams.

For example, the unigrams "Tran" and "Van" are not discriminative individually, as they appear in names from other cultures. Yet, when paired together, they form a distinctive feature in Vietnamese names, such as the commonly used "Tran Van." This bigram typically represents the family name "Trần" paired with the middle or first name "Văn" in Vietnam.

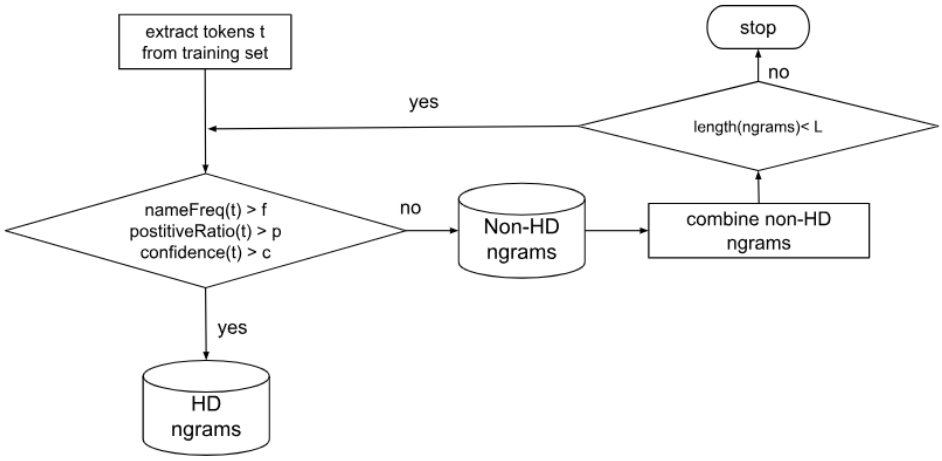


Figure 3. The process to generate Highly Discriminative N-Grams to retrieve Vietnamese names

Figure 3 illustrates the approach for generating Highly Discriminative N-Grams to retrieve Vietnamese names. Motivated by the work of Podnar et al. [32], which study showed that, due to the adherence of word frequency to Zipf's law [68, 69], the growth of word combinations for generating HDKs is constrained, meaning it cannot increase too rapidly as the document collection expands. This suggests that storing a manageable number of HDKs is feasible. In Vietnamese names, token frequency distribution follows Zipf's law, where a small set of tokens frequently appears across most names. When tokens are arranged in descending order of frequency, the nth token's value is inversely proportional to n. As a result, we can apply the same method to generate combinations of name tokens to create new, discriminative n-grams without worrying about exponential growth in combinations. While the HDK approach uses document frequency with a fixed threshold to limit the size of the inverted index during transmission over low-latency networks, our approach uses a confidence score to decide whether to combine tokens to create more discriminative n-grams.

A token or token pair is considered a **Highly Discriminative N-Gram (HDN)** if it meets the following criteria:

1. The frequency of the token(s) in positive name examples exceeds a predefined threshold, which helps filter out spelling errors or unusual name forms in the training set. Initially, this threshold is set to 10. A higher threshold increases the confidence in the estimated probability.
2. The confidence score for the token(s) exceeds a high threshold (e.g., 0.995).

The confidence score is calculated based on the estimated probability of a token being a Vietnamese name, derived from the dataset. However, we enhance the negative cases of tokens using a function to ensure that tokens frequently appearing in negative label names receive a very low confidence score. This negative amplification ensures that if a token is highly prevalent in Vietnamese names but is not rare in negative names, it is not considered highly discriminative.

$$confidence(t) = \left(\frac{f(t|1)}{f(t)}\right)^{1+\frac{f(t|0)}{2}}$$

In the above formula, f (t|1), f (t|0), and f (t) are respectively frequencies of positive, negative, and total names containing token t in the dataset. The frequency of negative names is put on power to make the confidence score reduce drastically if the number of negative labels increases. The confidence score always has a value between 0 and 1 (when f(t|0) = 0). For example, if the name "le" appears in 180 negative names, which is only 0.3% of the total 58152 names containing "le", but the confidence score is 0.75, then it is not a Highly Discriminative Unigram.

3.2 Training Dataset

To measure the discriminability of N-Grams on Vietnamese names, we assembled a dataset consisting of positive and negative examples of Vietnamese names gathered from various sources: (1) names extracted from the Vietnamese

phonebook through web crawling, (2) a compilation of student names from several universities in Vietnam, (3) names extracted from a European country's phonebook via web crawling, (4) names obtained from Wikipedia, and (5) names sourced from LinkedIn profiles. To ensure label accuracy, we utilized multiple validation methods, including cross-checking by volunteers (Each name is independently labeled by two people; if there is a discrepancy, a third person verifies it), verifying Vietnamese names through Wikipedia or LinkedIn profiles, and examining personal homepages or biographies available on the Internet and social media. The European country's list of personal names was chosen for crawling because we know that a high proportion of overseas residents (~25%) live there, contributing to a wide variety of name variations.

Table 1. Overview of the name dataset used in the research

Total Names	**4,126,815**
Positive names	635,678
Negative names	3,491,137
Total tokens	9,423,808
Average tokens per name	2.38
Unique unigrams	276,890
Positive unigrams	4,099

All names undergo normalization, which involves removing accents. For example, "Trần Văn" is transformed into "tran van." Additionally, single-character tokens are excluded to reduce the risk of errors with abbreviations. This normalization process is implemented because Vietnamese individuals and professionals often present their names in the international format on global websites and publications. Dataset statistics are shown in **Table 1**.

It appears that the negative set is much larger than the positive set. However, considering that Vietnam's population in 2024 is approximately 100 million compared to the global population of over 8 billion, this sample size is quite reasonable.

The top 30 unigrams and bigrams from Vietnamese names are shown in **Table 2**, highlighting that a few tokens dominate the majority of names. Around 4,000 unigrams make up over 635,000 Vietnamese names, with 124 classified as highly discriminative, primarily found in Vietnamese names and rarely in others. These 124 discriminative unigrams account for over 84% of the names in our dataset, with "nguyen" being the most common, representing nearly 9%. The most frequent bigram, "nguyen thi," is excluded because bigrams are only generated when the unigram is non-discriminative. Since "nguyen" is a discriminative unigram, no bigrams containing it are created. This strategy helps keep the number of highly discriminative n-grams manageable for efficient classification. In total, only 37,914 bigrams were generated, with 8,476 being discriminative.

Table 2. Top 30 unigrams from ~635K Vietnamese names

Unigram	**Percentage (%)**	**Highly Discriminative**	**Unigram**	**Percentage (%)**	**Highly Discriminative**
nguyen	8.90	Yes	duc	1.10	
thi	7.90	Yes	thu	1.05	
van	3.74		dang	0.99	
tran	2.97		bui	0.95	
le	2.83		do	0.95	
hoang	2.27	Yes	ha	0.94	
thanh	2.11	Yes	quang	0.89	Yes
ngoc	2.06	Yes	hong	0.89	
pham	2.01	Yes	duong	0.83	
anh	1.82	Yes	phan	0.82	

minh	1.62	Yes	truong	0.80
vu	1.41		xuan	0.80
thuy	1.13	Yes	tuan	0.79
phuong	1.12	Yes	linh	0.76
dinh	1.11		mai	0.72

The top 10 bigrams and their respective percentages are shown in **Table 3**. It is important to note that the most common bigram, "nguyen thi" is not included in this list because bigrams are only generated when the unigram is not a discriminative token. Since "nguyen" is considered a discriminative unigram for Vietnamese names, any bigrams containing it are excluded from generation. This strategy helps keep the number of highly discriminative n-grams small, which in turn ensures the classifier remains efficient in practice. Only 37,914 bigrams were generated, of which 8,476 are discriminative, making them easier to manage.

Table 3. Top 10 bigrams from ~635K Vietnamese names

Bigram	**Percentage**
tran van	0.99
le van	0.88
van vu	0.45
bui van	0.34
do van	0.34
duc tran	0.30
hong le	0.30
thu tran	0.30
dinh van	0.29
duong van	0.29

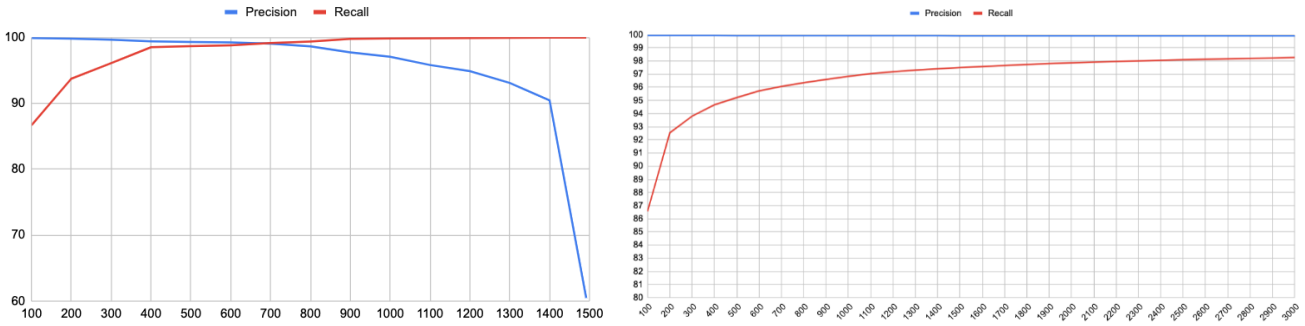
After the HD-Ngrams set is generated with n=1,2, it is used as queries sent to name databases to collect Vietnamese names. Using n=3 can improve the accuracy of Vietnamese name collection; however, in practice, combining more than 10,000 non-discriminative bigrams with a few hundred non-discriminative unigrams results in a very large number of possible trigrams. This leads to inefficiency, as it would require sending too many queries to retrieve Vietnamese names. It is worth noting that after names are collected, additional methods can be applied to verify whether they are indeed Vietnamese. For instance, leveraging LLMs or checking other fields such as language, workplace, and relationships. However, these post-processing steps fall outside the scope of this paper.

Although this method generates more queries compared to using only unigrams as queries, the returned results exhibit high precision and good recall with minimal post-processing costs. This is because the n-grams produce significantly smaller result sets compared to unigrams.

3.3 Result of the training dataset

Using all 4,000 unigrams as queries to retrieve Vietnamese names would result in a nearly 100% estimated recall. However, the precision of the result set would be extremely low. Our experiment on retrieving the training dataset showed that when using nearly 1,500 unigrams for querying, precision dropped to 60% (**Figure 4a**). It is clear that among the top 10 most popular Vietnamese unigrams, "Le" and "Van" tend to retrieve many false positive cases, particularly European names.

Figure 4b shows the results after using 100 HD-Unigrams for retrieval on the training dataset, which achieves an estimated recall of 86%. We then start using HD-Bigrams to retrieve queries on the experimental dataset. The precision remains at 99.9%, while the recall steadily increases to 98% with 2,300 bigrams.



a. Raw unigram **b. Highly Discriminative N-Grams**

Figure 4. Precision and Recall (%) when using raw unigrams and HDNs for retrieval on the training dataset

By using 500 HDNs for Vietnamese name retrieval, we achieve a recall of 95% and a precision of 99.9%. If we limit the number of queries to 100 (a practical constraint since some systems restrict the number of queries), we can achieve approximately 90% recall and over 99.9% precision with 69 HD-Unigrams and 31 HD-Bigrams.

The experimental setup differences are shown in **Table 4**, facilitates us to determine the optimal number of n-grams, as well as the minimum frequency and confidence thresholds, to balance estimated precision, recall, and the number of queries used for retrieving Vietnamese names.

Table 4. Different experiment setups and results on the training dataset

#HD-Unigrams	#HD-Bigrams	Min Frequency	Min Confidence	Precision	Recall
232	6311	10	0.90	99.762	**99.072**
179	3229	30	0.95	99.808	98.758
111	2334	50	0.99	99.910	98.078
111	1329	100	0.99	99.919	97.437
100	400	100	0.99	99.928	95.181
69	31	100	0.99	**99.943**	89.796

For an analysis of the result, **Table 4** shows a trade-off between precision and recall as the minimum frequency and confidence thresholds are adjusted. As the minimum frequency and confidence increase, precision improves, reaching 99.943% with 69 unigrams and 31 bigrams at a minimum frequency of 100 and confidence of 0.99. However, recall decreases as these thresholds are raised, dropping from 99.072% at a minimum confidence of 0.90 to 89.796% with the same high thresholds. This indicates that while higher frequency and confidence thresholds yield more precise results, they reduce recall by excluding less common but potentially relevant names, highlighting the need to balance precision and recall based on the specific requirements of the name retrieval task.

4. RESULTS ON EVALUATION DATASETS

4.1 Evaluation datasets

The experiment makes use of 2 more datasets (**Table 5**) with different characteristics for evaluating the model's performance:

1. Evaluating Dataset 1 (ED1-career2023): This dataset contains the top cited scientists worldwide, as published by Ioannidis et al. [70, 71]. For the unsure Vietnamese names, we verified the information of scientists based on their homepage or bio on the internet to confirm they are originally Vietnamese. As the number of positive labels is small and we can verify carefully the profile of scientists, we consider this to be the golden set of our application.

2. Evaluating Dataset 2 (ED2-behindthename): Random names were generated from the BehindTheName website. BehindTheName is a website for learning about all aspects of names. All given names from all cultures and periods are eligible to be included in the central name database. There are currently more than 25,000 names in the website database. Given a national name, the website provided the tool to generate random names of that nation. Therefore, we can easily have the correct label for Vietnamese names and names from other countries.

Table 5. Evaluating datasets

Evaluation Dataset	Total names	Positive labels
ED1-career2023	252,458	301
ED2-behindthename	30,877	453

We use Elasticsearch for our retrieval queries, a robust, modern, very scalable, and distributed search and analytics engine used everywhere to handle huge volumes efficiently. Elasticsearch provides strong full-text search capabilities and supports complex queries for data analysis in real-time and at minimal latency. It allows us to process structured and unstructured data and index it to retrieve relevant information effectively. In our work, we index an evaluation dataset using Elasticsearch, where each entry is labeled as 1 for Vietnamese names and 0 for non-Vietnamese names. This indexing will organize the data and make it accessible for retrieval tasks.

4.2 Results on evaluation datasets

For evaluation datasets, the HD N-Grams are the combination of both highly discriminative unigrams and bigrams, which sum up to 6’525 HDNs.

Table 6 Query results on evaluation datasets

Dataset	Metric	Raw Unigrams	HD N-Grams
ED1-career2023	Precision	0.0201	0.7851
	Recall	1.0000	0.9867
ED2-behindthename	Precision	0.1770	0.9305
	Recall	1.0000	1.0000

Table 6 illustrates the results of queries on evaluation datasets, comparing raw unigrams and Highly Discriminative N-Grams (HD N-Grams) for querying Vietnamese names. The results show a significant improvement in precision when using HD N-Grams, making this approach far more effective and efficient for real-world applications. For instance, the precision of raw unigrams is notably low, at 0.0201 for the ED1 dataset and 0.1770 for the ED2 dataset. Such low precision rates highlight the inefficiency of using raw unigrams, as they result in a high number of irrelevant matches. While the recall for raw unigrams is perfect (1.0), this comes at the cost of retrieving a large volume of non-Vietnamese names, necessitating additional post-processing to filter the results.

The need for post-processing when using raw unigrams introduces significant overhead. For example, manual filtering to identify Vietnamese names would be time-consuming and prone to human error, while training machine learning models to classify names adds computational cost and requires additional labeled data. Both methods increase the overall cost in terms of time, effort, and resources, making raw unigrams impractical for efficient querying in real-world scenarios.

In contrast, using HD N-Grams substantially improves precision while maintaining high recall. For the ED1 dataset, the precision increases from 0.0201 to 0.7851, representing nearly a 39-fold improvement. Although recall drops slightly from 1.0000 to 0.9867, this means only a negligible number of names (approximately 4) are missed, which is a small trade-off given the substantial gains in precision. The name we missed, e.g. “trac tran” is because the frequency of this bigram is too small in the training data set, therefore we don’t consider it a confident HDN. Similarly, for the ED2 dataset, the precision

rises dramatically from 0.1770 to 0.9305, with recall remaining unchanged at 1.0000. These results demonstrate that HD N-Grams are not only more accurate but also more efficient, as they eliminate the need for costly post-processing steps.

Furthermore, the significant improvement in precision reflects the ability of HD N-Grams to capture the unique structure and token combinations characteristic of Vietnamese names. By focusing on discriminative patterns, HD N-Grams effectively reduce noise and retrieve highly relevant results. This capability is particularly valuable in applications where precision is critical, such as name-matching systems or identity verification tools.

When considering cost optimization, the advantages of HD N-Grams become even more pronounced. The reduced need for additional filtering or classification processes leads to faster query execution and lower operational costs. The evaluation for cost comparison will be discussed in the next part.

4.3 Cost evaluation

The need for cost evaluation is crucial in information retrieval systems. The cost of queries can be assessed in terms of the total number of query results. For instance, when using a search engine like Google, if a query returns 100 results and each page displays only 10 results, the user will need to click through 10 pages to view all the results. This implies that 10 separate API calls are made, which applies to both manual searches and programmatic API queries. Search engines typically have quotas on both the load they can handle and the number of queries they can process. Therefore, it is essential to optimize query keywords to minimize unnecessary API calls, ultimately enhancing the efficiency of the retrieval process.

Table 7 Cost evaluation on evaluation datasets

Dataset	Total queried results (duplicates counted)	
	Raw Unigrams	**HD N-Grams**
ED1-career2023	17,289	730
ED2-behindthename	3,743	1,170

Table 7 compares the total number of query results between using raw unigrams and HD N-Grams for querying Vietnamese names. The results demonstrate a clear improvement in both the significance and efficiency of the query process when using HD N-Grams. For the ED1 dataset, querying with raw unigrams returns over 17,000 results, which is approximately 24 times higher than the total number of results obtained when using HD N-Grams. This stark difference highlights the efficiency of HD N-Grams in narrowing search results to the most relevant matches, thereby reducing unnecessary post-processing to get final relevant results.

For the ED2 dataset, the results are also noteworthy, with raw unigrams returning over 3,700 results, while HD N-Grams yield around 1,170 results. Although the reduction in results is less dramatic compared to ED1, it still demonstrates the efficiency of using HD N-Grams in reducing irrelevant matches. This improvement not only enhances the precision of the results but also suggests that HD N-Grams are capable of efficiently handling queries in datasets with varying levels of complexity and size. The overall trend across both datasets confirms that HD N-Grams provide a more targeted and resource-efficient approach for querying Vietnamese names, which can lead to better performance in real-world applications.

The complete set of highly discriminative n-grams and the evaluation dataset from this research have been made publicly available to the research community via a GitHub repository: https://github.com/toanluu/vietname/

5. CONCLUSION

In this work, we addressed the challenges of querying Vietnamese names from a global database, a task that has become increasingly important for real applications to support Vietnamese communities. As databases constantly grow in size and diversity, the difficulty of distinguishing Vietnamese names from others, especially when dealing with similar naming conventions, increases. This challenge is particularly evident in systems where accurate name matching is critical, such as in demographic research, data mining, and personalized services. We proposed a novel approach that focuses on generating

highly discriminative N-Grams (HDNs), which are optimized for efficiently retrieving Vietnamese names by considering the distribution of N-Grams between Vietnamese and non-Vietnamese names.

The results of applying HDNs in our experiments showed substantial improvements in both precision and cost efficiency. When querying the ED1 dataset, which includes top global scientists, the HDNs approach demonstrated a 39-fold improvement in precision and a 24-fold reduction in cost compared to traditional unigram-based queries. These results underscore the effectiveness of HDNs in improving the retrieval process, making it not only more accurate but also more resource-efficient. By leveraging the distinct distribution patterns of Vietnamese and non-Vietnamese names, our method enables more precise identification and classification in large-scale name-querying systems.

While the proposed approach using Highly Discriminative N-Grams (HDNs) significantly improves the accuracy and efficiency of Vietnamese name retrieval, several limitations remain. The method heavily depends on the quality and representativeness of the training data, which may lead to reduced performance when encountering rare or unconventional name variations. This is reflected in the fact that the query results are still not the most efficient, especially for recall metrics, since some Vietnamese names are not discovered. Additionally, the predefined frequency and confidence score thresholds require careful tuning to balance precision and recall effectively. The current approach is optimized for Vietnamese names, and its applicability to other languages remains an open question, requiring further adaptation. Moreover, while Elasticsearch provides a scalable retrieval framework, exploring alternative indexing and search techniques may further enhance efficiency. We still had some unwanted results in the experimental queries, for example, duplicated highly discriminative bigrams would return non-Vietnamese names. This is because of the match query behavior of Elastic Search. Future work will focus on refining threshold selection strategies, incorporating adaptive learning mechanisms to handle evolving name patterns, and extending the approach to multilingual name retrieval tasks. Additionally, integrating deep learning models for context-aware name disambiguation could complement the HDN approach and improve overall performance.

Furthermore, the application of HDNs extends beyond retrieval systems and can be effectively utilized in other tasks, such as statistical demographic research or name-ethnicity classification. For example, in classification tasks, compared to traditional machine learning methods, HDNs offer a low-cost solution with fewer parameters, making them an attractive alternative for applications that require high performance but are constrained by computational resources. In future work, we plan to refine and expand our approach to handle larger, more complex datasets, aiming to further improve precision and broaden the applicability of HDNs in various domains. This ongoing research will contribute to more efficient, scalable, and accurate systems for querying and analyzing names in global databases.

REFERENCES

1. Wikipedia. Available from: https://en.wikipedia.org/wiki/Overseas_Vietnamese.
2. Tống, D.T., *Viết thuật ngữ khoa học, tên người và địa danh.* VNU Journal of Science: Social Sciences and Humanities, 2013. **29**(2).
3. Brunelle, M., *Vietnamese (Tiếng Việt).* The handbook of Austroasiatic languages, 2015. **2**: p. 909-953.
4. Le, T.M.T. and N.T. Pham, *English and Vietnamese Female Names: A Comparison and Contrast.* Journal of Educational and Social Research, 2020. **10**.
5. KHƯƠNG, P.H., *ĐÔI NÉT VỀ ĐẶC ĐIỂM HỌ TÊN CỦA NGƯỜI TRUNG QUỐC VÀ NGƯỜI VIỆT NAM.*
6. Nhung, Đ.T., *HỌ TÊN NGƯỜI NGA VÀ NGƯỜI VIỆT.* LỜI NGƯỜI HƯỚNG DẪN KHOA HỌC: p. 97.
7. Sanderson, M. and W.B. Croft, *The history of information retrieval research.* Proceedings of the IEEE, 2012. **100**(Special Centennial Issue): p. 1444-1451.
8. Hambarde, K.A. and H. Proenca, *Information retrieval: recent advances and beyond.* IEEE Access, 2023. **11**: p. 76581-76604.
9. Xiong, X. and M. Zheng, *Merging mixture of experts and retrieval augmented generation for enhanced information retrieval and reasoning.* 2024.
10. Esteva, A., et al., *COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization.* NPJ digital medicine, 2021. **4**(1): p. 68.
11. Jain, S., K. Seeja, and R. Jindal, *A fuzzy ontology framework in information retrieval using semantic query expansion.* International Journal of Information Management Data Insights, 2021. **1**(1): p. 100009.

12. Gupta, V. and A. Dixit, *Recent query reformulation approaches for information retrieval system-a survey*. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 2023. **16**(1): p. 94-107.
13. Ibrihich, S., et al., *A Review on recent research in information retrieval.* Procedia Computer Science, 2022. **201**: p. 777-782.
14. Khemmarat, S. and L. Gao, *Predictive and personalized drug query system.* IEEE Journal of Biomedical and Health Informatics, 2016. **21**(4): p. 1146-1155.
15. Cafarella, M.J. and O. Etzioni. *A search engine for natural language applications*. in *Proceedings of the 14th international conference on World Wide Web*. 2005.
16. Seo, D.-W., et al., *Cumulative query method for influenza surveillance using search engine data.* Journal of medical Internet research, 2014. **16**(12): p. e289.
17. Thelwall, M., *Quantitative comparisons of search engine results.* Journal of the American Society for Information Science and Technology, 2008. **59**(11): p. 1702-1710.
18. Bradlow, E.T. and D.C. Schmittlein, *The little engines that could: Modeling the performance of World Wide Web search engines.* Marketing Science, 2000. **19**(1): p. 43-62.
19. Google. *Google Search Console API Usage Limits* Available from: developers.google.com/webmaster-tools/limits.
20. LinkedIn. *LinkedIN API Terms of Use*. Available from: linkedin.com/legal/l/api-terms-of-use.
21. Marchiori, M., *The quest for correct information on the Web: Hyper search engines.* Computer Networks and ISDN Systems, 1997. **29**(8-13): p. 1225-1235.
22. Madhu, G., D.A. Govardhan, and D.T. Rajinikanth, *Intelligent semantic web search engines: a brief survey.* arXiv preprint arXiv:1102.0831, 2011.
23. Lewandowski, D., S. Sünkler, and N. Yagci. *The influence of search engine optimization on Google's results: A multi-dimensional approach for detecting SEO*. in *Proceedings of the 13th ACM Web Science Conference 2021*. 2021.
24. Almukhtar, F., N. Mahmoodd, and S. Kareem, *Search engine optimization: a review.* Applied computer science, 2021. **17**(1): p. 70-80.
25. Järvelin, A., A. Järvelin, and K. Järvelin, *s-grams: Defining generalized n-grams for information retrieval.* Information Processing & Management, 2007. **43**(4): p. 1005-1019.
26. Wang, X., et al., *Efficient and effective knn sequence search with approximate n-grams.* Proceedings of the VLDB Endowment, 2013. **7**(1): p. 1-12.
27. Robertson, A.M. and P. Willett, *Applications of n-grams in textual information systems.* Journal of Documentation, 1998. **54**(1): p. 48-67.
28. Bai, X., et al. *Scalable query n-gram embedding for improving matching and relevance in sponsored search*. in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2018.
29. Wang, X., A. McCallum, and X. Wei. *Topical n-grams: Phrase and topic discovery, with an application to information retrieval*. in *Seventh IEEE international conference on data mining (ICDM 2007)*. 2007. IEEE.
30. Dai, Z., et al. *Convolutional neural networks for soft-matching n-grams in ad-hoc search*. in *Proceedings of the eleventh ACM international conference on web search and data mining*. 2018.
31. Chelba, C., M. Norouzi, and S. Bengio, *N-gram language modeling using recurrent neural network estimation.* arXiv preprint arXiv:1703.10724, 2017.
32. Podnar, I., et al. *Scalable peer-to-peer web retrieval with highly discriminative keys*. in *2007 IEEE 23rd International Conference on Data Engineering*. 2006. IEEE.
33. Buttcher, S., C.L. Clarke, and G.V. Cormack, *Information retrieval: Implementing and evaluating search engines*. 2016: Mit Press.
34. Lal, N., S. Qamar, and S. Shiwani, *Information retrieval system and challenges with dataspace.* International Journal of Computer Applications, 2016. **147**(8).
35. Hambarde, K.A. and H. Proenca, *Information retrieval: recent advances and beyond.* IEEE Access, 2023.
36. Salton, G., E.A. Fox, and H. Wu, *Extended boolean information retrieval.* Communications of the ACM, 1983. **26**(11): p. 1022-1036.
37. Kolda, T.G. and D.P. O'leary, *A semidiscrete matrix decomposition for latent semantic indexing information retrieval.* ACM Transactions on Information Systems (TOIS), 1998. **16**(4): p. 322-346.
38. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation.* Journal of machine Learning research, 2003. **3**(Jan): p. 993-1022.

39. Wang, J., et al., *Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges.* ACM Computing Surveys, 2024. **56**(7): p. 1-33.
40. Robertson, S. and H. Zaragoza, *The probabilistic relevance framework: BM25 and beyond.* Foundations and Trends® in Information Retrieval, 2009. **3**(4): p. 333-389.
41. Pang, L., et al. *Deeprank: A new deep architecture for relevance ranking in information retrieval*. in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2017.
42. Guo, J., et al., *A deep look into neural ranking models for information retrieval.* Information Processing & Management, 2020. **57**(6): p. 102067.
43. Zhu, Y., et al., *Large language models for information retrieval: A survey.* arXiv preprint arXiv:2308.07107, 2023.
44. Baek, I., et al., *Crafting the path: Robust query rewriting for information retrieval.* arXiv preprint arXiv:2407.12529, 2024.
45. Carpineto, C. and G. Romano, *A survey of automatic query expansion in information retrieval.* Acm Computing Surveys (CSUR), 2012. **44**(1): p. 1-50.
46. Su, Z., et al. *Modeling intent graph for search result diversification*. in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021.
47. Zhou, Y., et al. *PSSL: self-supervised learning for personalized search with contrastive sampling*. in *Proceedings of the 30th ACM international conference on information & knowledge management*. 2021.
48. Nakano, R., et al., *Webgpt: Browser-assisted question-answering with human feedback, 2021.* URL https://arxiv. org/abs/2112.09332, 2021.
49. Yates, A., R. Nogueira, and J. Lin. *Pretrained transformers for text ranking: BERT and beyond*. in *Proceedings of the 14th ACM International Conference on web search and data mining*. 2021.
50. Peng, W., et al. *Large language model based long-tail query rewriting in taobao search*. in *Companion Proceedings of the ACM on Web Conference 2024*. 2024.
51. Weller, O., et al., *When do generative query and document expansions fail? a comprehensive study across methods, retrievers, and datasets.* arXiv preprint arXiv:2309.08541, 2023.
52. Bai, J., et al. *Using query contexts in information retrieval*. in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 2007.
53. FAGAN, J., *Experiments in automatic phrase indexing for document retrieval: A comparison of syntactic and non-syntactic methods(Ph. D. Thesis)*. 1988.
54. Mitra, M., et al. *An analysis of statistical and syntactic phrases*. in *RIAO*. 1997.
55. Berger, A. and J. Lafferty. *Information retrieval as statistical translation*. in *ACM SIGIR Forum*. 2017. ACM New York, NY, USA.
56. Kowalski, G., *Information retrieval architecture and algorithms*. 2010: Springer Science & Business Media.
57. Cavnar, W.B. and J.M. Trenkle. *N-gram-based text categorization*. in *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*. 1994. Ann Arbor, Michigan.
58. Treeratpituk, P. and C.L. Giles. *Name-ethnicity classification and ethnicity-sensitive name matching*. in *Proceedings of the AAAI conference on artificial intelligence*. 2012.
59. Lee, J., et al. *Name Nationality Classification with Recurrent Neural Networks*. in *IJCAI*. 2017.
60. Chen, Y., et al. *Identifying language origin of person names with n-grams of different units*. in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. 2006. IEEE.
61. Schnell, R., et al., *A new name-based sampling method for migrants using n-grams.* German Record Linkage Center, Working Paper Series, No. WP-GRLC-2013-04, 2013.
62. Aliwy, A., A. Abbas, and A. Alkhayyat, *NERWS: Towards improving information retrieval of digital library management system using named entity recognition and word sense.* Big Data and Cognitive Computing, 2021. **5**(4): p. 59.
63. da Fonseca, F.P.C. and L.A. Digiampietri. *Improving researcher's area of expertise identification using TF-IDF Characters N-grams*. in *Proceedings of the XVII Brazilian Symposium on Information Systems*. 2021.
64. Georgieva-Trifonova, T. and M. Duraku. *Research on N-grams feature selection methods for text classification*. in *IOP Conference Series: Materials Science and Engineering*. 2021. IOP Publishing.
65. Ojo, O., et al., *Performance study of n-grams in the analysis of sentiments.* Journal of the Nigerian Society of Physical Sciences, 2021: p. 477-483.
66. Yalcin, K., I. Cicekli, and G. Ercan, *An external plagiarism detection system based on part-of-speech (POS) tag n-grams and word embedding.* Expert Systems with Applications, 2022. **197**: p. 116677.
67. El Hadi, A., et al., *Finding Relevant Documents in a Search Engine Using N-Grams Model and Reinforcement Learning.* Journal of Information Technology Research (JITR), 2022. **15**(1): p. 1-17.

68. Zipf, G.K., *the Principle of Least Effort*. 1949: Cambridge: Addisor-Wesley.
69. Zipf, G.K., *The psycho-biology of language: An introduction to dynamic philology*. 2013: Routledge.
70. Ioannidis, J.P., et al., *A standardized citation metrics author database annotated for scientific field.* PLoS biology, 2019. **17**(8): p. e3000384.
71. Ioannidis, J.P., K.W. Boyack, and J. Baas, *Updated science-wide author databases of standardized citation indicators.* PLoS biology, 2020. **18**(10): p. e3000918.