



HAL
open science

A Scalable Document-based Architecture for Text Analysis

Ciprian-Octavian Truică, Jérôme Darmont, Julien Velcin

► **To cite this version:**

Ciprian-Octavian Truică, Jérôme Darmont, Julien Velcin. A Scalable Document-based Architecture for Text Analysis. 12th International Conference on Advanced Data Mining and Applications (ADMA 2016), Dec 2016, Gold Coast, Australia. pp.481-494. hal-01380064

HAL Id: hal-01380064

<https://hal.science/hal-01380064>

Submitted on 19 Dec 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

A Scalable Document-based Architecture for Text Analysis

Ciprian-Octavian Truică¹, Jérôme Darmont², Julien Velcin²

¹ Computer Science and Engineering Department, Faculty of Automatic Control
and Computers, University Politehnica of Bucharest, Bucharest, Romania

`ciprian.truica@cs.pub.ro`

² Univ Lyon, Lumière, ERIC EA 3083, Lyon, France

`{jerome.darmont,julien.velcin}@univ-lyon2.fr`

Abstract. Analyzing textual data is a very challenging task because of the huge volume of data generated daily. Fundamental issues in text analysis include the lack of structure in document datasets, the need for various preprocessing steps and performance and scaling issues. Existing text analysis architectures partly solve these issues, providing restrictive data schemas, addressing only one aspect of text preprocessing and focusing on one single task when dealing with performance optimization. Thus, we propose in this paper a new generic text analysis architecture, where document structure is flexible, many preprocessing techniques are integrated and textual datasets are indexed for efficient access. We implement our conceptual architecture using both a relational and a document-oriented database. Our experiments demonstrate the feasibility of our approach and the superiority of the document-oriented logical and physical implementation.

Keywords: Text Analytics · Indexing Methods · Document-Oriented Databases

1 Introduction

A vast amount of textual data is generated daily and it is really challenging to develop efficient models and systems to enhance processing performance while doing accurate text analysis. The most fundamental challenges when working with large volumes of heterogeneous text datasets include the lack of structure of textual corpora, the various required preprocessing steps, the need for efficient access and the ability to scale up.

Structural issues may be addressed by resorting to textual data warehousing and On-Line Analytical Processing (OLAP). However, such approaches only partially solve the problem because they use a structured schema that falls short when applied to large, heterogeneous volumes of data. Moreover, using a predefined schema makes them extremely dataset-specific.

Moreover, when dealing with textual data, we distinguish different preprocessing levels: quite basic operations (e.g., cleaning HTML tags, tokenization,

language identification); intermediate operations (e.g., stemming, lemmatization, indexing); and advanced operations (e.g., part of speech tagging, named entity recognition, topic modeling). Each complexity layer in this process requires the previous layer and all operations must remain tractable in terms of memory and CPU time. To the best of our knowledge, no text analysis tool implements all layers, nor any processing workflow.

Finally, when working on performance and scaling issues, state-of-the-art research focuses on one aspect of text analysis, e.g., aggregation, top-k keyword extraction and text indexing. However, text processing techniques used in a single application may be many and, as we mention above, interdependent.

Hence, we present in this paper a scalable text analysis architecture that addresses all these issues. More precisely, we deal with the lack of structure by adopting a novel generic, document-oriented data model that allows storing heterogeneous textual corpora with no predefined structure. We also integrate in our framework all the preprocessing methods that are useful for information retrieval, data mining, text analysis and knowledge discovery. We also propose a new compact data structure to minimize index storage space and the response time of create, read, update and delete (CRUD) operations. Such indexes benefit to text preprocessing, querying and further analysis, and adequately contribute to global scaling.

The remainder of this paper is organized as follows. In Section 2, we discuss related works. In Section 3, we present the architecture and implementation of our approach. In Section 4, we experimentally validate our proposal. In Section 5, we finally conclude this paper and hint at future research.

2 Related Works

2.1 Text Cubes and OLAP

Extensive work on information retrieval (IR) and text analysis have been done using OLAP. Most proposals use Text Cubes for OLAPing multidimensional text databases [20]. Lin et al. focus on optimizing query processing and reducing storage costs of Text Cubes [12]. They experimentally show that average query time and storage cost are related to a cube's number of dimensions. Zhang et al. use Text Cubes for topic modeling [20] and experimentally show that their approach is much faster than computing each topic cube from scratch. Finally, Ding et al. address the problem of keyword search and top-k document ranking using Text Cubes [7]. Their algorithms perform well in terms of query response and memory cost when the number of search terms is small.

Ben Kraiem et al. propose a generic multidimensional model for OLAP on tweets [2]. Their experiments show some promising results for knowledge discovery when applying OLAP on a small corpus, but query performance decreases when data volume increases. Bringay et al. propose a data warehouse model to analyze large volumes of tweets [4]. They introduce different operators to identify trends using the top-k most significant words over a period of time for a specific

geographical location, as well as the impact of hierarchies on such operators. Unfortunately, no time performance and storage cost analysis is provided.

In conclusion, research done so far on text analysis and OLAP focuses on small, structured datasets and scaling up is not guaranteed.

2.2 Text Preprocessing and Analysis

Managing morphological variation of search terms in IR has been quite extensively studied [11, 10]. The main successful methods are stemming [14] and lemmatization, which are used to optimize search, minimize the space allocated to inverted indexes (Section 2.3) and, in the case of lemmatization, to add linguistic information. Lemmatization is useful for different types of advanced text analysis, e.g., named entity recognition, automatic domain specific multi-term extraction and part of speech (PoS) tagging. Moreover, lemmatization is easier of use than stemming, saves storage and improves retrieval performance [11].

Topic modeling is a statistical model for discovering hidden themes that occur in a collections of documents. In recent years, it has been extensively studied, showing the usefulness of analyzing latent topics and discovering topic patterns [3, 16]. Popular approaches for topic modeling are latent semantic indexing (LSI) [6], latent Dirichlet allocation (LDA) [3], the non-parametric extension hierarchical Dirichlet process (HDP) [16] and non-negative matrix factorization (NMF) [1].

2.3 Document Indexing

Inverted indexes are data structures used in search engines, whose main purpose is to optimize query response speed. Basic inverted indexes store terms, a list of documents where each term appears and a weight. Weight measures the number of occurrences of the term in a document, e.g., raw term frequency/word co-occurrence (TF), normalized Term Frequency (TF_n), etc. In the various methods for managing inverted indexes, great emphasis is put on storage space reduction. For instance, a pruning algorithm based on term frequency-inverse document frequency ($TF*IDF$) can be used to minimize index size [19]. Yet, updating an inverted index is also a problem, because it is dependent on documents. The index must indeed be updated each time documents are added or deleted.

3 Proposed Approach and Implementation

3.1 Approach Overview

The approach we propose (Figure 1) is subdivided into four steps: 1) clean and preprocess documents using natural language processing (NLP) and store the information in a database; 2) construct indexes; 3) analyze data, e.g., with topic modeling, etc.; 4) query and search data, extract top-k most relevant documents, create visualizations and analyses. We construct the inverted index, vocabulary,

PoS and named entities (NE) indexes during the index construction step. Indexes may be used afterward by data mining, text analysis, search and visualization. The search engine sorts documents based on a ranking function (e.g., TF*IDF, Okapi or BM25) to extract the top-k documents.

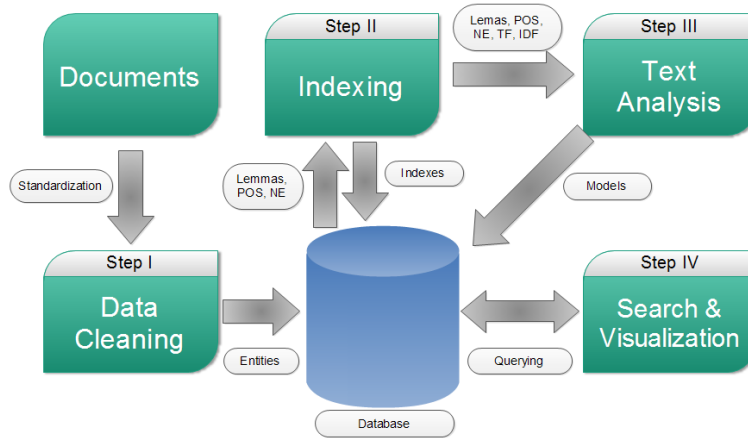


Fig. 1: System architecture

To implement our document-oriented approach, we quite naturally rely on a document-oriented database management system (DODBMS). DODBMSs are a class of NoSQL systems that aim to store, manage and process data using a semi-structured model. DODBMSs encapsulate data in collections of documents [8]. A document can contain other nested documents, which turns out to be very flexible [17].

One feature of DODBMSs is that they are often optimized for create and read operations, while offering reduced functionality for update and delete queries. DODBMSs are designed to work with large amounts of data and the main focus is on the efficiency of data storage, access and analysis [13]. Another key feature of DODBMSs is the distribution of data across multiple sites. In particular, DODBMSs can horizontally scale CRUD operations throughput [5]. Moreover, decentralized data stores provide good mechanisms for fail-over, removing the single point of failure, due to their scalability and flexibility [9].

We selected MongoDB as our DODBMS, since it beats the best mean time performances for CRUD operations both in single and distributed environments [17]. Moreover, we also implemented our approach with PostgreSQL, to provide a point of comparison with a well-established, efficient relational database management system (RDBMS) (Section 4).

3.2 Data Models

We design a generic model to store heterogeneous text data using a data warehouse snowflake schema (Figure 2). The central component of the model is the *documents* entity, where we store basic information and metadata about a document, e.g., timestamp, title, raw, clean and lemmatized text, etc. The *document_tags* entity is used to store metadata represented by tags, which can be existing tags, hashtags or at tags. The *vocabulary* entity links documents to information extracted or inferred from the text, which helps enhancing metadata with different weights and tags, e.g., PoS, TF, TF_n , lemmas, etc. The *named_entities* entity stores all the information about entities automatically extracted from the original corpus.

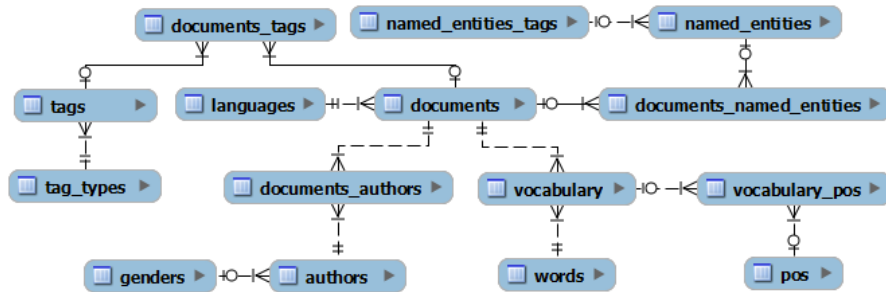


Fig. 2: Conceptual model

The DODBMS schemaless design takes all the information presented in the relational schema and stores it for each document in a record of the collection. Using this design, all one-to-many and many-to-many relationships become either vectors (e.g., *hashtags*, *at tags*) or nested documents (e.g., *words*, *named_entities*). Where the information is not present, these vectors and nested documents may be missing thanks to the flexibility of schemaless database design. A problem that arises is duplication, as multiple records can bear the same metadata, since all the information for a document is stored in one single record. The *vocabulary* entity is constructed as a separate collection. This entity is constructed dynamically, taking user input constraints into account, e.g. date, tags, search words, named-entities.

Interaction with the database is achieved through CRUD operations, aggregation functions and views. We use read operations for information extraction and data visualization. Aggregation functions are used for constructing indexes, searching and preprocessing data for text analysis. We make use of MapReduce for this purpose when using the document-oriented database architecture. Dynamically materialized cubes are constructed using views with aggregation functions, fine-graining query results using different measures, e.g., timestamps, locations, lemmas, tags, named entities.

3.3 Text Preprocessing

The data cleaning module serves three functions: 1) corpus standardization, 2) text preprocessing using NLP to enrich data, and 3) entity creation and information insertion into the database.

The entire corpus is standardized by determining all the fields of a document, including metadata and the labels of *documents*. Then, during the preprocessing step, the following techniques are applied: 1) text cleaning by removing HTML/XML tags and scripts; 2) language identification; 3) expanding contractions; 4) extracting features, e.g., PoS, lemmas and named entities; 5) removing stop words and punctuation; 6) computing term weights. We use a multithreading architecture for data cleaning to cope with large data volumes and scale up vertically. At the end of each thread, the information is stored in a dictionary, together with other metadata. We choose to use asynchronous threads because, after a worker thread finishes, a new job can be assigned to it without waiting for the other worker threads to finish. This is made possible because each task is independent. At the end of this step, a record of the *documents* collection is created and inserted into the database. The record contains all labels from the first step and the information extracted using NLP from the second one.

In the DODBMS implementation, a record stores all the information because its attributes are created dynamically. In contrast, the RDBMS architecture can only store predefined fields due to its rigid schema. Thus, undefined fields are omitted.

The RDBMS approach merges the data cleaning step with the index construction step, because many-to-many relationships between entities, translated as bridge tables, are indexes as well. We could not use a multithreading approach here because information could be lost. Multiple threads could indeed check at the same time whether the information is present and receive a negative response. A constraint violation error could appear and the transaction terminate by a *rollback*. If constraints are missing, then duplicate information could appear and this would impact text analysis.

3.4 Index Management

We propose several indexes for document aggregation, search, extraction of the top-k most significant terms and text analysis, e.g., topic modeling, document clustering. These new indexing structures minimize storage costs and maximize the time performance of CRUD operations.

Index construction in the DODBMS architecture is done using the MapReduce framework. Four indexes are created: 1) an *inverted index* that stores, for each term, a list of corresponding documents; 2) a *vocabulary*, a novel inverted index with additional information for each term in the corpus, e.g., list of documents where the term is found, the TF and TF_n of the term for a document and IDF; 3) a PoS index that stores the part of speech of each term; 4) a named-entity index used for storing named entities. There are no integrity constraints

between these collections to improve query response time. Moreover, the structure proposed for the *vocabulary* facilitates query response time, aggregation and search (Figure 3). MapReduce is used to construct all indexes. It is also central in aggregation queries needed by the search algorithms. To improve index construction and query response times, we horizontally scale the database, and by doing so add more MapReduce worker.

In the RDBMS architecture, indexes are the bridge tables translating many-to-many relationships between entities. The *vocabulary* is the bridge table between the *documents* table and the *words* table. The PoS index is the bridge table between the *vocabulary* table and the *pos* table. In this case, the index also contains the TF and IDF of each term.

```
{
  "_id" : 791297891579,
  "word" : "hillary",
  "idf" : 4.39,
  "createdAt" : "2016-03-03 13:09:13.802",
  "docIDs" : [
    { "docID" : 628840524375, "count" : 4, "tf" : 0.64 },
    { "docID" : 628840524457, "count" : 6, "tf" : 0.70 },
    .....
    { "docID" : 628840524430, "count" : 3, "tf" : 0.71 } ] }
```

Fig. 3: Vocabulary index structure

The number of entries in the indexes constructed for the DODBMS is equal to the number of terms in the entire corpus. In the RDBMS, the inverted index has more entries, i.e., $\sum_{d \in D} |t : t \in d|$, where D is the corpus and $|t : t \in d|$ is the number of distinct terms that appear in document d .

Updating indexes in the DODBMS is based on document insertion date. The update method we use constructs an intermediary index for new documents, and then it updates the primary index by appending the new documents' ID and TF to existing labels. Then, the IDF of each term is updated for the whole index. When documents are deleted, we apply a bulk delete operation. In this case, a list of deleted document IDs is stored, which helps update the index structure by removing the deleted documents and then updating the IDF of each term.

Updating indexes in the RDBMS implementation is easier thanks to the database's structure. When documents are added, indexes are automatically updated based on the insertion date of the last added documents. When documents are removed, the corresponding index entries are also removed. For both operations, the IDF of each term must be recalculated.

4 Experimental Validation

In this section, we test each step of our approach and we compare the results achieved by the two instances we developed, i.e., the DODBMS version imple-

mented with MongoDB and the RDBMS version implemented with PostgreSQL. Tests are done using a news corpus consisting of 110,000 articles¹, a corpus of 5,000,000 tweets² and a scientific corpus of 20,000 abstracts from [15]. The size of these corpora, some would argue, is rather small with respect to Big Data. Yet, it is sufficient to illustrate our architecture’s good time performance. Text analysis is indeed not usually done on large corpora. Moreover, other corpora used in the literature are smaller, e.g., 3,000 documents [20], 2,013 records [12], 65,333 tweets [2], 1,801,810 tweets [4].

Our architecture can be deployed in a cloud environment if all the requirements are met, i.e., if Python packages, PostgreSQL, and MongoDB are available. Tests are done on machines that reside in an OpenStack private cloud platform. We purposely selected this hardware architecture and dataset sizes to show that our architecture can achieve good performance even on end-user workstations, as it is sometimes not desirable to send data online due to privacy issues. Moreover, end-users presumably cannot afford very powerful, parallel computers.

4.1 News Articles Corpus Experiments

The first set of experiments are done using two computers with the same hardware configuration: 4 GB RAM and 1 CPU with two 2.2 GHz cores. We choose this hardware architecture to show that our method gives good results on simple computers. Using the initial news articles corpus, seven corpora are created consisting of 100 to 110,000 documents. They are referred to as Corpus $i, i \in \{1, 2, \dots, 7\}$. For comparison reasons, experiments are done using a single-thread approach.

Figure 4a presents the average time (in seconds) for populating the databases. Duplicate documents are removed in this step. This is done by checking whether an article already exists in the database based on its title. If the document does not exist, then a new record is added. Otherwise, tags are verified so that metadata are not omitted, as the same article could have more tags for different instances found in the corpus. The second set of tests evaluates the efficiency of text cleaning and index construction (Figure 4b).

Figure 4c shows the total storage space (in MB) for all corpora. To respect database normalization in PostgreSQL, bridge tables materializing many-to-many relations have to be added. In MongoDB, such relationships translate into vectors or nested documents inside collections. For example, the *documents* collection contains the authors table as an array of nested documents and the tags table as an array. This brings the issue of duplicates, as we may have the same tags for different document that would be stored in each element of the collection. However, it is a small cost to pay as, using this structure, joins are removed, whereas join is the costliest operation in RDBMSs.

Experimental results show that MongoDB efficiently stores the data, minimizing storage space by 30% with respect to PostgreSQL. Moreover, based on

¹ <http://www.corpora.heliohost.org>

² Collected with Twitter’s tools at <https://dev.twitter.com>

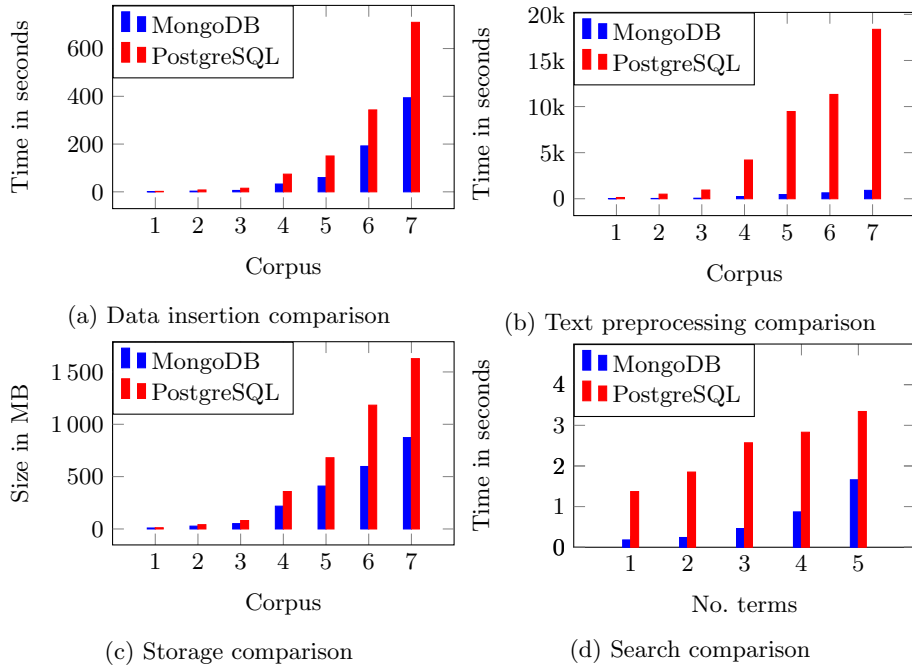


Fig. 4: Performance comparison

the number of records in each collection, from a computational point of view, a select operation performed on a smaller entity shows faster response times than one performed on an entity with a lot of records. For example, it is faster to query the *vocabulary* collection than to interrogate the *vocabulary* table, because the table contains more records than the collection.

Figure 4d presents the mean time for extracting the top-k documents. Tests are performed on Corpus 7 with $k = 20$. After each search, the database cache and buffers are cleared so that the comparison is accurate. MongoDB is from 86% faster than PostgreSQL for one term-search to over 50% faster for five terms.

Table 1a presents mean text cleaning and index construction times, as index construction is done separately in MongoDB. MapReduce functions were developed to further improve performance. Our results show that text cleaning and index creation is improved by 94% with MongoDB (Figure 4b). Moreover, index update is an important feature in a system where new documents are added or deleted. We use new corpora of 500 to 5,000 articles from Corpus 5 to test this feature in MongoDB. For comparison purposes, for each operation, we tested the performance of updating and rebuilding the entire index. Updating the inverted index and the PoS index (Table 1b) works fast if the number of added documents is small, but time performance shifts for bigger corpora. Then, it is better to rebuild the entire index. If documents are deleted, it is faster to rebuild the inverted index (Table 1c). Little improvement is seen between updating and

Table 1: Index building and updating in MongoDB (seconds)

(a) Text cleaning and index construction

Corpora	Clean Text	PoS Index	Inverted Index	Vocabulary
Corpus 1	25.36	4.95	5.94	6.57
Corpus 2	103.74	12.63	16.15	19.26
Corpus 3	204.91	20.10	25.36	32.95
Corpus 4	934.56	57.74	81.46	117.15
Corpus 5	1 805.27	103.38	141.81	209.14
Corpus 6	2 524.33	148.69	212.16	312.61
Corpus 7	3 564.07	216.34	311.88	461.88

(b) After new documents are added

No. Documents	Inverted Index		Vocabulary		PoS Index	
	Update	Rebuild	Update	Rebuild	Update	Rebuild
500	37.65	123.63	1 163.95	189.62	33.53	91.52
1 000	76.54	126.16	1 214.20	208.47	67.79	93.08
2 500	144.56	126.07	1 303.89	204.18	129.44	95.24
5 000	201.26	130.16	1 395.10	201.51	179.90	97.72

(c) After documents are removed

No. Documents	Inverted Index		Vocabulary		PoS Index	
	Update	Rebuild	Update	Rebuild	Update	Rebuild
500	1.08	122.36	1 129.35	198.35	91.33	91.91
1000	1.38	123.42	1 135.08	199.20	92.00	92.88
2500	1.51	126.64	1 160.26	195.09	93.15	94.52
5000	1.60	129.76	1 202.71	203.01	96.03	97.40

rebuilding the PoS index (Table 1c) when documents are deleted. Concerning vocabulary, it is faster to rebuild the entire index than to update it, because the IDF must be recomputed for each element in the collection (Tables 1b and 1c).

4.2 Twitter Corpus Experiments

This set of experiments is carried out using one machine with the following hardware configuration: 12 GB RAM and 3 CPU with 4 2.6 GHz cores. We choose this hardware configuration to prove that our architecture does not require specialized hardware to have good time performance. We work on 5,000,000 tweets in these experiments.

Figure 5a presents the results obtained when using a multithreading architecture. The improvement obtained from switching from a single thread to a 12-thread implementation is 90%, lowering preprocessing time by a factor of 10. We can observe that the number of nodes used by MongoDB directly impacts performance and enhances response time, especially for large numbers of tweets.

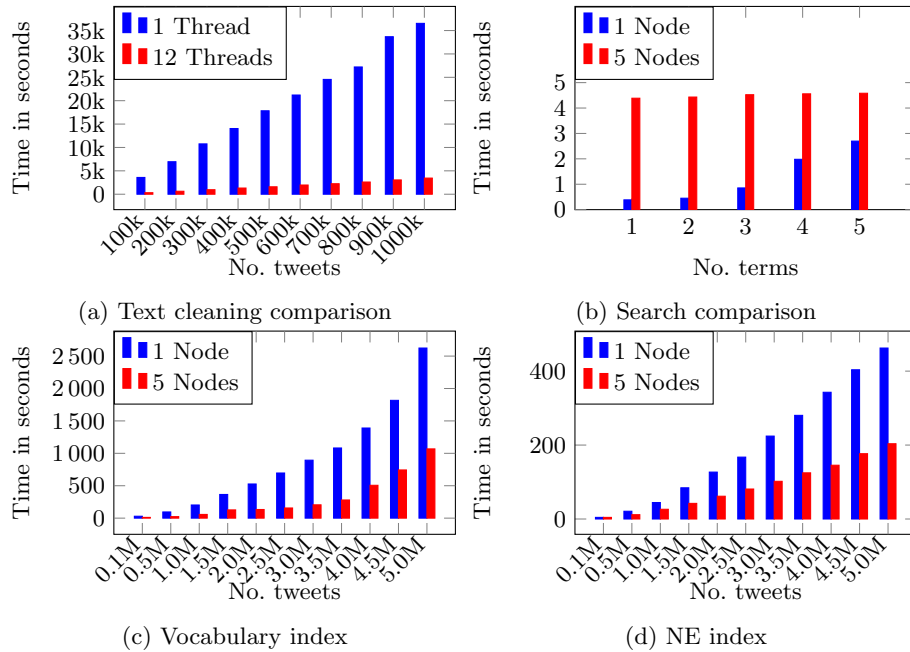


Fig. 5: Index construction comparison

The construction time of the vocabulary index improves significantly, by over 59% (Figure 5c). The same happens with the named entities index, with an improvement over 40% (Figure 5d). Keyword search performance remains constant when we scale the database horizontally (Figure 5b).

4.3 Scientific Articles Corpus Experiments

This set of experiments uses the scientific corpus and is carried out using the same hardware configuration as in Section 4.2. These experiments are designed to test the time performance for constructing the vectorization matrices and extracting topics. Figure 6a displays construction time for four different vectorization matrices, namely TF, TF_n , TF*IDF and Okapi BM25. The best performance is obtained by the TF_n vectorization matrix because all the information exists in the *vocabulary* index. TF*IDF and Okapi BM25 vectorizations are slower because they must be computed for each element during matrix construction. The second set of tests presents the performance time of extracting topics from the entire corpus (6b). LSI is faster than LDA and HDP by a factor of 21 and 13, respectively. NMF achieves the best performance.

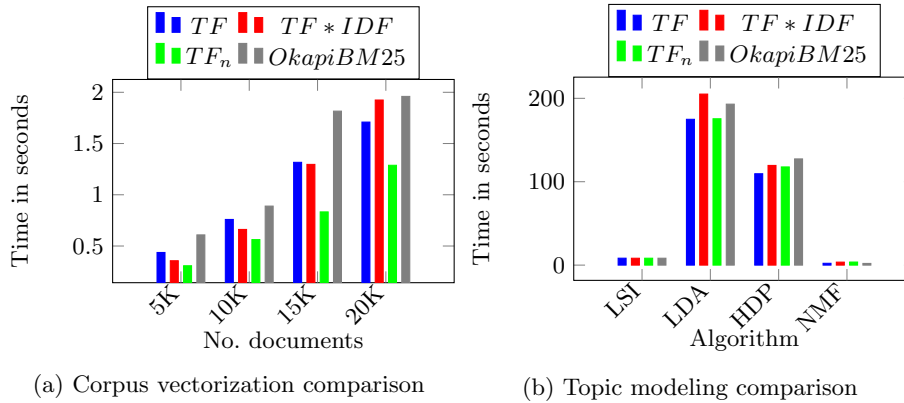


Fig. 6: Topic modeling comparison

5 Conclusion

In this paper, we present a new, complete architecture for text analysis that improves search performance, minimizes storage cost through efficient document-oriented storage, and scales up horizontally and vertically. Moreover, by exploiting MapReduce to parallelize index construction and by designing new structures for indexing and decreasing the number of records stored in the database, we minimize the number of CRUD operations and further enhance performance. Finally, the algorithm we propose for extracting top-k documents for a given search phrase also considerably improves query response time.

Our experimental results show that a document-oriented architecture is best-suited and improves performances when working with large volumes of text when adding documents into the database, cleaning text and constructing indexes. For all test cases, the mean time for populating the DODBMS is half that of the RDBMS. Cleaning texts and constructing inverted indexes is also faster when using a DODBMS. Although duplicates can be found inside a DODBMS, storage costs are significantly lower than with a RDBMS. A demo application that further shows the capabilities of this architecture is presented in [18].

In future work, we plan to add new features to our framework, such as automatic domain specific multiterm extraction, cross-language IR, word embedding and new topic models, e.g., dynamic topic modeling. From an architectural point of view, we also want to parallelize the algorithms and use a GPU for computations.

References

1. Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., Zhu, M.: A practical algorithm for topic modeling with provable guarantees. In: International Conference on Machine Learning. pp. 939–947 (2013)

2. Ben Kraiem, M., Feki, J., Khrouf, K., Ravat, F., Teste, O.: OLAP of the tweets: From modeling toward exploitation. In: International Conference on Research Challenges in Information Science. pp. 1–10 (2014)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022 (2003)
4. Bringay, S., Béchet, N., Bouillot, F., Poncelet, P., Roche, M., Teisseire, M.: Towards an On-Line Analysis of Tweets Processing. In: International Conference on Database and Expert Systems Applications. pp. 154–161 (2011)
5. Cattell, R.: Scalable SQL and NoSQL data stores. *ACM SIGMOD Record* 39(4), 12–27 (2011)
6. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6), 391–407 (1990)
7. Ding, B., Zhao, B., Lin, C.X., Han, J., Zhai, C., Srivastava, A., Oza, N.C.: Efficient keyword-based search for top-k cells in text cube. *Transactions on Knowledge and Data Engineering* 23(12), 1795–1810 (2011)
8. Han, J., Haihong, E., Le, G., Du, J.: Survey on NoSQL database. In: International Conference on Pervasive Computing and Applications. pp. 363–366 (2011)
9. Hecht, R., Jablonski, S.: NoSQL Evaluation: A Use Case Oriented Survey. In: International Conference on Cloud and Service Computing. pp. 336–341 (2011)
10. Jivani, A.G.: A comparative study of stemming algorithms. *International Journal of Computer Technology and Applications* 2, 1930–1938 (2011)
11. Kettunen, K., Kunttu, T., Järvelin, K.: To stem or lemmatize a highly inflectional language in a probabilistic IR environment? *Journal of Documentation* 61(4), 476–496 (2005)
12. Lin, C.X., Ding, B., Han, J., Zhu, F., Zhao, B.: Text cube: Computing IR measures for multidimensional text database analysis. In: International Conference on Data Mining. pp. 905–910 (2008)
13. Redmond, E., Wilson, J.R.: Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement. *The Pragmatic Bookshelf* (2012)
14. Sharma, D.: Stemming Algorithms: A Comparative Study and their Analysis. *International Journal of Applied Information Systems* 4, 7–12 (2012)
15. Tang, J., Wu, S., Sun, J., Su, H.: Cross-domain Collaboration Recommendation. In: ACM SIGKDD. pp. 1285–1293 (2012)
16. Teha, Y.W., Jordana, M.I., Beala, M.J., Bleia, D.M.: Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* 101(476), 1566–1581 (2012)
17. Truică, C.O., Boicea, A., Rădulescu, F., Bucur, I.: Performance evaluation for CRUD operations in asynchronously replicated document oriented database. In: International Conference on Control Systems and Computer Science. pp. 191–196 (2015)
18. Truică, C.O., Guille, A., Gauthier, M.: CATS: Collection and Analysis of Tweets Made Simple. In: ACM Conference on Computer-Supported Cooperative Work and Social Computing. pp. 41–44 (2016)
19. Vishwakarma, S.K., Lakhtaria, K.I., Bhatnagar, D., Sharma, A.K.: An Efficient Approach for Inverted Index Pruning Based on Document Relevance. In: International Conference on Communication Systems and Network Technologies. pp. 487–490 (2014)
20. Zhang, D., Zhai, C.X., Han, J., Srivastava, A., Oza, N.: Topic cube: Topic modeling for OLAP on multidimensional text databases. In: SIAM International Conference on Data Mining. pp. 1124–1135 (2009)