SN Computer Science (2023) 4:691
https://doi.org/10.1007/s42979-023-02146-9

ORIGINAL RESEARCH

Reliability and Performance of the Online Literature Database CAMbase after Changing from a Semantic Search to a Score Ranking Algorithm

Sebastian Unger[1] · **Christa K. Raak**[2] · **Thomas Ostermann**[1]

Received: 3 February 2023 / Accepted: 14 July 2023

Abstract
Despite the increase in scientific publications in the field of integrative medicine over the past decades, a valid overview of published evidence remains challenging to get. The online literature database *CAMbase* (available at https://cambase.de) is one of the established databases designed to provide such an overview. In 2020, the database was migrated from a 32-bit to a 64-bit operating system, which resulted in unexpected, technical issues and forced the replacement of the semantic search algorithm with *Solr*, an open-source platform that uses a score ranking algorithm. Although semantic search was replaced, the goal was to create a literature database that is essentially no different from the legacy system. Therefore, a before-after analysis was conducted to compare first the number of retrieved documents and then their titles, while the titles were syntactically compared using two Sentence-Bidirectional Encoder Representations from Transformers (SBERT) models. Analysis with a paired t-test revealed no significant overall differences between the legacy system and the final system in the number of documents ($t = -1.41$, $df = 35$, $p = 0.17$), but an increase in performance ($t = 4.13$, $df = 35$, $p < 0.01$). Analysis with a t-test for independent samples of the values from the models also revealed a high degree of consistency between the retrieved documents. The results show that an equivalent search can be provided by using *Solr*, while improving the performance, making this technical report a viable blueprint for projects with similar contexts.

Keywords Evaluation · Search engine · Information storage and retrieval · Database management systems · Semantics · Deep learning

Introduction

Database technology nowadays is an essential part in everyday life. The constantly changing requirements for storage systems [1] were only one of the important building blocks in the development towards this technology. In particular, in the field of libraries, the requirement of accessing "all the world's literature from a single computer terminal" was a topic of discussion in the late 1960s already before the internet was invented [2]. In this time, first computerized literature databases such as MEDLARS emerged [3]. With the rise of the internet, the number of purpose related database systems rapidly grew. The access to scientific literature was eased and first specialist libraries, e.g., for the field of nuclear magnetic resonance [4], emerged as well. Such specialist literature databases have also been created in the field of complementary and alternative medicine (CAM). A 2009 review counted a total of 45 online accessible databases covering various aspects topics such as phytotherapy, traditional chinese medicine, or music therapy [5].

One of these databases is *CAMbase*. The Chair of Medical Theory and Complementary Medicine at the Witten/Herdecke University initiated the first version in 1998, enabling users to easily find relevant scientific literature on CAM. In 2007, *CAMbase v2.0* arose and was implemented

This article is part of the topical collection "Advances on Knowledge Discovery, Knowledge Engineering and Knowledge Management" guest edited by Joaquim Filipe, Ana Fred, Frans Coenen, Jorge Bernardino and Elio Masciari.

✉ Sebastian Unger
sebastian.unger@uni-wh.de

[1] Faculty of Health, Department of Psychology and Psychotherapy, Witten/Herdecke University, Witten, Germany

[2] Faculty of Health, Center for Integrative Medicine, Witten/Herdecke University, Witten, Germany

Published online: 09 September 2023

using Extensible Markup Language (XML) protocols and interfaces, in accordance with the requirements of the Open Archives Initiative [6]. Most importantly, *CAMbase v2.0* was equipped with a semantic-syntactic search algorithm that benefits users by deconstructing a search query into linguistic (i.e., semantic and grammatical) parts and then transmitting the relevant documents in XML-packaged form [7–9]. At the time *CAMbase v2.0* was released, the proprietary decomposition of search queries was much more detailed than with usual stemming algorithms. Besides forming the word stem, the word order, the word ending, and even umlauts, which are special to the German language, were used for searching in the index and calculating the relevance of documents. Page numbers and stop words such as *"the"*, *"on"*, or *"and"* were removed from the search query in advance, whereas the Boolean operators between each word were still taken into account [7, 8]. A search with relatively similar search queries (e.g., *"treatment of hospital patients"* and *"treatment of patients in hospitals"*) resulted ultimately in different documents by recognizing the linguistic parts. In sum, *CAMbase v2.0* was on the information technological cutting edge for a specialist literature database at the time of its development.

As already mentioned in the early paper of Barraclough [2] and in contrast to conventional opinions, hosting an online literature database is not an easy task in many aspects. In addition to functionality, security must be ensured for the underlying hardware and software. Lifespan of an operating system (OS) increases the likelihood to find existing vulnerabilities, as evidenced by numerous reported vulnerabilities that can potentially cause substantial risks [10, 11]. On the other hand, there is more time to develop patches or better OS versions that are distributed without these vulnerabilities. An example of a decrease in vulnerability risks could be shown using a mean risk factor calculation method for the three versions of Microsoft Windows 7, 8, and 10 [12].

CAMbase v2.0, has been running on the same 32-bit OS since its release. As the database can be accessed publicly, it has been at risk of being attacked, e.g., by denial-of-service attacks (DOS) [13] or intrusions with mostly bad and unethical intentions [14, 15]. Despite the wide spectrum of securing an OS, selecting the right one already shows that it may be used to improve an intrusion-tolerant system [16]. All these facts led to the need to migrate *CAMbase v2.0* to a modern 64-bit OS.

An essential requirement for the migration was to preserve the previous data. According to Haynes' understanding [17], it is important for patient care to follow current and best evidence-based medicine. Even though CAM pursues the approach of addressing the evidence along with emphasizing the patients and their relationship with the practitioner, the evidence remains limited [18], which may itself limit the patient care and shows the need of such data.

Apart from the preservation of the previous data, a further requirement concerns the effort that must be invested in the migration. For user-friendliness, the key components of *CAMbase v2.0* should be migrated without major changes. These components include the established graphical user interface (GUI) and the approach of generating both the website and the retrieved documents on the client's side using Extensible Stylesheet Language Transformations (XSLT) and XML protocols [6, 8].

This technical report describes the migration process of *CAMbase*, the challenges that had to be solved, and a final evaluation of the system. More precisely, the chapter "Migration Process" gives an overview of the initial system architecture, outlines the issue with a pure migration, and justifies the replacement of an important component of the system architecture, namely the semantic-syntactic algorithm, with a current search engine that uses a score ranking algorithm (final development version: *CAMbase v3.0*). Then, the chapter "Comparison" presents the pre-trained language models and the statistical analysis that largely incorporates these models. This is followed by the chapter "Results", in which it is analyzed whether the new retrieval processes affect the performance of the system by means of speed, accuracy, and reliability. The basic assumption is that current search engines can keep up with the semantic algorithm of 2007 despite using a different algorithm. The following research questions are considered: First, does the system retrieve the same search results after changing to a score ranking algorithm as before? And second, to what extent is the performance of the system affected after this change? The "Lessons learned" chapter revisits the challenges of this migration and the approach that was used to solve them. The advantages and disadvantages of the approach are also highlighted here. The final chapter rounds off this technical report with some concluding remarks.

Migration Process

System Architecture

The architecture, with which *CAMbase v2.0* was built, follows the layered architecture pattern for smaller applications [19]. *CAMbase v2.0* has three main layers with specific roles and responsibilities, namely GUI, business logic, and database (architecture on the left side of Fig. 1). The GUI is used for the graphic presentation. The presentation takes place on the client side after the data (i.e., web elements as well as literature of the database) has been transmitted by the server in XML protocols. The business logic, where the semantically algorithm is located, is responsible for processing user input, search queries, and data retrieval. The last layer is the database, which at the time after migration

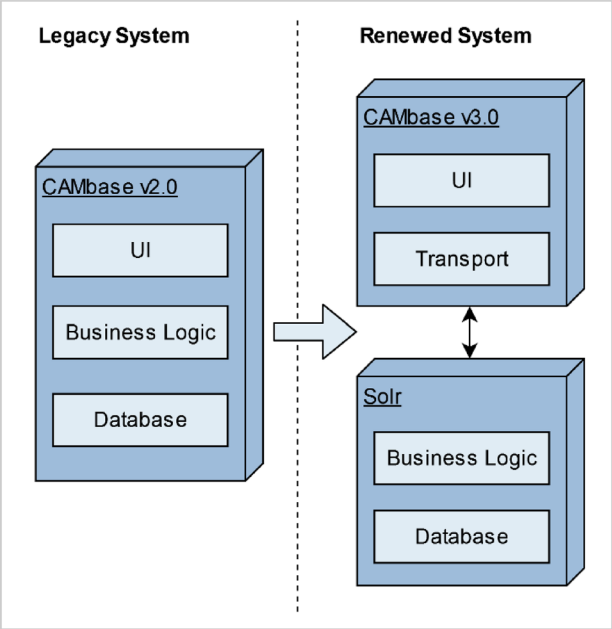

Fig. 1 Model of the system architecture before (*CAMbase v2.0*) and after the migration (*CAMbase v3.0*) taken from [20]

contained 115,355 entries (e.g., books, case reports, clinical studies, or experimental work) from 1906 onwards. With this three-layered architecture, it is intended to give simplicity to develop a new database by just replacing one of the layers.

As already stated, *CAMbase v2.0* has been running on the same 32-bit OS since its release. With the migration of *CAMbase v2.0* on a 64-bit OS, multiple errors occurred (e.g., missing libraries, missing literature references, overlapping GUI elements, or wrong interpreted search queries). Although the pre-defined requirements seemed satisfied, the technical differences of the new architecture have to be carefully taken into account. Otherwise, such migration can even result in software vulnerabilities if the intricacies of this architecture are not considered, showing the complexity of this process [21, 22]. Since there were only binary files and no source code of the legacy search algorithm, a complete inspection or replication was not possible. In order to maintain user acceptance and thus the online literature database itself, the complete search algorithm had to be replaced.

Search Engine Solution

Over the years, especially since the release of *CAMbase v2.0*, search engines have significantly improved by comparing and developing different indexing approaches [23]. In addition, research in this area has compared search engines and rated their search capabilities and functionalities in order to find the most relevant documents [24–27]. Therefore, the legacy search algorithm was replaced with the search engine *Apache Solr*. This fast and poplar search engine is based on *Apache Lucene*, which itself has powerful indexing capability and supports a lot of search features [28, 29]. A comparison with *Xapian* already showed *Solr*'s good performance in searching for the most important documents [30]. Some of the key features of *Solr* are that it is ready-to-deploy, open source, centrally configurable, and allows full-text search and scalable search across multiple servers [29, 31]. In a search, the relevance of a document determines where it appears in the retrieved documents. For this purpose, a document's relevance factor is calculated, which takes into account, among other things, the frequency of words of the search query within a document or even a specific relevance increasing boost [32]. This is by no means a semantic interpretation, but *Solr*'s configuration pool provides a lot of scope for more specific search. For example, queries can be separated by alphanumeric characters or modified by adding stemming or phonetic algorithms.

System Architecture Adjustment

On the 64-bit system, all files related to the legacy algorithm were removed. This also removed the whole business logic from the three-layered architecture. Afterwards, *Solr* (version 8.9.0) was installed. Various routines and preparations (e.g., definition of fields and user roles) followed before importing the cleaned data. Cleaned data here means that, for example, duplicates were removed and types were unified. Since *Solr* is not intended to be used as a stand-alone system, another layer was implemented to allow the communication between GUI and *Solr* (architecture on the right side of Fig. 1). For this, a PHP: Hypertext Preprocessor (PHP) script supported by the PHP Extension Community Library (PECL) was used so that the layer can parse the user inputs as *Solr*-understandable queries and retrieve documents to the users.

The next stage was to approximate the syntactic search of *CAMbase v2.0*. *Solr* offers many ways to narrow down the search. In the end, a light stemming method was embedded to also search for slightly variant words, which was similarly used by the legacy algorithm. Converting the words (index and query) to low case enables even greater reach in finding relevant documents. In contrast to the legacy algorithm, the words of a query are handled independently, which is done by separating them by blanks. The words still had to be joined with the proper operator for a qualitatively high approximation. The users itself can nevertheless narrow down the query through the execution character supported by *Solr*. A list of synonyms was not utilized, as this would be too time-consuming for migration, but could lead to a more accurate approximation.

Comparison

Pre-Trained Language Model

To compare *CAMbase v2.0 and CAMbase v3.0*, the title of each retrieved document is used for a semantic comparison with two pre-trained language model.

Pre-trained language models can be considered as state-of-the-art in natural language processing and semantic text similarity detection. The models Bidirectional Encoder Representations from Transformers (BERT), A Lite BERT (ALBERT), and Embeddings from Language Models (ELMo) are impressive examples of this, especially when they are fine-tuned [33–35]. There are already numerous optimization approaches in the literature [33, 36–38]. This approach focuses only on Sentence-BERT (SBERT).

SBERT is based on BERT, maintaining the accuracy of BERT but improving the effort and thus the performance needed to compare large numbers of literature titles. In order to compare those titles, SBERT uses siamese and triplet network structures to derive semantically meaningful sentence embeddings [37]. A comparison is then made by taking the cosine similarity between the sentence pairs A and B as the similarity score [39] according to the formula:

$$\cos\theta = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}; A = (A_1, \ldots, A_n), B = (B_1, \ldots, B_n)$$

Technically, values range between -1 and 1 for this approach (see Fig. 2) and thus can be interpreted like correlation coefficients: Values close or equal to 1 means a high correlation while a value close or equal to -1 also means a high correlation, but in the opposite direction.

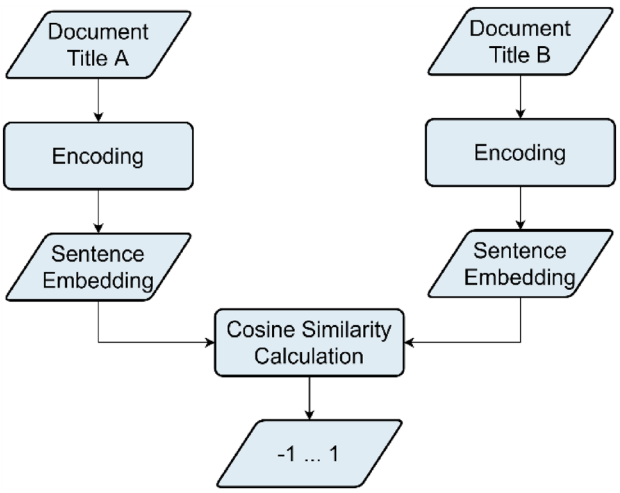

Fig. 2 SBERT approach to compute similarity scores in accordance with [37]

However, value lower than 0 are not expected so that only values between 0 and 1 are considered in the following.

A general-purpose model (all-MiniLM-L6-v2) was selected in a previous comparison [20]. As the name suggests, this model can be applied to many use cases, e.g., comparing literature titles searched within two database systems. However, this model was only trained with English data. This allows an overall comparison in all documents' provided titles but does not consider other languages. Therefore, this search reliability comparison refers additionally to a multi-language model (paraphrase-multilingual-MiniLM-L12-v2), which considers translations of other languages.

Statistical Analysis

In this approach, *CAMbase v2.0*, equipped with a proprietary syntactic algorithm, and *CAMbase v3.0*, equipped with *Solr* and a score ranking algorithm, are compared. The comparison is based on 36 search queries, which were suggested by experts from the field of CAM. They are a combination of terms from the list of Wieland et al. [40] and German key terms (see Table 3 and Fig. 5). The queries were executed in both systems with the four restrictions "All words", "Keywords", "Abstracts", and "Titles". Thus, four dependent pairs of outcome variables are obtained per query, i.e., the retrieved documents, their titles, and the query time that was needed.

Values of the outcome variables then were manually entered into a data sheet for further data analysis. Firstly, the number of retrieved documents was compared. For this purpose, a mean value given by the sum of the documents divided by the search queries executed in them was calculated for each restriction. Secondly, the mean query times were compared analogously to the number of retrieved documents. Finally, the titles of the retrieved documents are compared. Here, SBERT is applied, using the two models described above (all-MiniLM-L6-v2 and paraphrase-multilingual-MiniLM-L12-v2). Similarities calculated by SBERT were based on the titles of the document, where only the most similar title was regarded. The calculation went in both directions, i.e., all titles from the documents of *CAMbase v2.0* are compared with those of *CAMbase v3.0* and vice versa. Mean values then were calculated by summing the SBERT values within the search queries, including all the restrictions, and dividing by the number of retrieved documents. These means represent the reliability of the systems.

Statistical analysis concluded with a t-test for each part with the functions of Microsoft Excel for Windows, considering a level of significance of 5%. Equivalently, for graphical displays, means and their 95% confidence intervals were used.

Results

Quantitative Reliability

In a first case, words of a search query are joint with the operator "OR" in *CAMbase v3.0*. The differences to the legacy systems were tremendous and could easily verified as significant with a t-test on these interim results [41]. This is because a query to the *CAMbase v3.0* retrieves a union of documents with this setting if the search query consists of multiple words. The more words the query has, the larger the union can be. The significant differences are in accordance with some user statements that the content of a search corresponds no longer to the usual, but fortunately without errors. A few users seemed slightly positive about the larger document range, because they might finally obtain more results for their systematic reviews and meta-analyses. Nevertheless, a second case was conducted with the operator "AND". This led to a more comparable number of documents between both systems. Because of the now created intersection, the mean number of documents retrieved in *CAMbase v3.0* is no longer statistically different from those retrieved in *CAMbase v2.0*. This applies to all restrictions and the change was not unnoticed by users either, which stated the content of a search as very accurate. Table 1 contains the t-test results based on the means. In addition, a graphical overview of the means is shown in Fig. 3.

On the one hand, the *Solr*-based system offers an increase of documents if manually or automatically a union is built, using the operator "OR". On the other hand, a similar number of documents can be retrieved with an intersection, using the operator "AND". This can also be seen in the bar charts and applies to all four restrictions. For the restriction "All words", the number of documents was slightly lower in *CAMbase v2.0* ($\bar{x} = 193$) than in *CAMbase v3.0* ($\bar{x} = 210$), despite the fact that *CAMbase v2.0* still operated with a semantic-syntactic search algorithm to find more relevant documents.

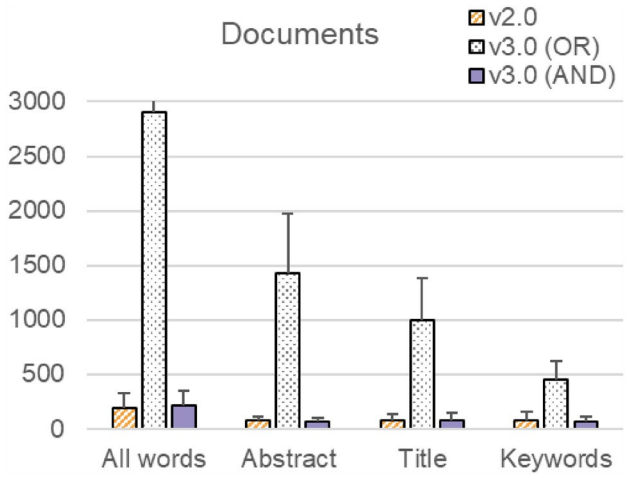

Fig. 3 Means of the number of documents separated into the restriction "All words", "Abstract", "Title", and "Keywords". The striped bar represents *CAMbase v2.0*, the dotted bar represents *CAMbase v3.0* by using the operator "OR", and the filled bar represents *CAMbase v3.0* by using the operator "AND". Means are partially taken from [20]

Performance

Performance was compared via query times like quantitative reliability by setting the operator in *CAMbase v3.0* twice first to "OR" and then to "AND". Both cases showed an improvement in performance compared to the legacy system (see Table 2). Except for the restriction "Title" after setting the operator to "AND", the t-test provides statistical proof of improvement. As this restriction just slightly missed to be significant, it demonstrates that the processing of queries via *Solr* is overall more efficient compared to the algorithm of *CAMbase v2.0*.

Table 1 Results of the paired *t* test, comparing the mean numbers of documents of *CAMbase v2.0* and *CAMbase v3.0*, whereas the operator in *CAMbase v3.0* is first set to "OR" and then to "AND"

Systems	Restriction	| t |	df	*p*
v2.0 vs v3.0 (OR)	All words	4.73	35	< 0.01
	Abstract	4.78	35	< 0.01
	Title	4.41	35	< 0.01
	Keywords	4.02	35	< 0.01
v2.0 vs v3.0 (AND)	All words	1.43	35	0.17
	Abstract	0.45	35	0.66
	Title	1.32	35	0.20
	Keywords	1.6	35	0.12

The values of the bottom row were taken from [20]

Table 2 Results of the paired t-test, comparing the mean query times of *CAMbase v2.0* and *CAMbase v3.0*, whereas the operator in *CAMbase v3.0* is first set to "OR" and then to "AND"

Systems	Restriction	| t |	df	*P*
v2.0 vs v3.0 (OR)	All words	4.43	35	< 0.01
	Abstract	4.00	35	< 0.01
	Title	2.55	28	0.02
	Keywords	3.42	30	< 0.01
v2.0 vs v3.0 (AND)	All words	4.2	35	< 0.01
	Abstract	3.78	33	< 0.01
	Title	2.07	28	0.05
	Keywords	3.17	30	< 0.01

The values of the bottom row were taken from [20]

Figure 4 displays the query times recorded from *CAMbase v2.0* and the two variants of *CAMbase v3.0*. Obviously, *Solr* outperforms the legacy algorithm regardless of the operator. *Solr* also shows, when setting the operator from "AND" to "OR", that its performance is maintained despite the increasing numbers of retried documents.

Search Reliability

As the search in *CAMbase 3.0* with the "AND" operator seemed closer to the legacy system, only this case is analyzed in respect of the search reliability. Regardless of the applied model, the values from SBERT indicate a high level of consistency when comparing the titles retrieved from *CAMbase v2.0* with those retrieved from *CAMbase v3.0* and vice versa. No mean calculated within the 36 search queries was below 0.5 as shown in Fig. 5. With both models, the best result was with the search query "Craniosacral Manipulation" ($N=2$). The exact same documents were retrieved before and after the migration, leading to a SBERT value of 1. The search with "Morita Therapy" ($N=3$) performed the worst with *CAMbase v2.0*. SBERT calculated a value of 0.639 with the general-purpose model and a value of 0.669 with the multi-language model. With the final system, the worst SBERT value was at 0.661 with the general-purpose model and 0.68 with the multi-language model when searching for "Bee Products" ($N=11$). These small values only occur on one side, i.e., a search with these queries in the other system performed much better. The reason for this effect is given by the lower number of documents in comparison to that retrieved from the other system. If there is a higher number of documents in an overall small set than in the other system, the difference obviously cannot be found in the retrieved documents of the system with the smaller set. This can be seen by looking at the equivalent values of 0.968 and 0.939 calculated with the general-purpose model and multi-language model, respectively, for the query "Bee Products" ($N=7$) and the value of 1 calculated with both models for the query "Morita Therapy" ($N=2$). The unilateral effect can also explain lower values if the number of retrieved documents is relatively high in both systems. Again, the difference leads to a lack of equivalent documents, with a high difference amplifying the effect. In *CAMbase v2.0*, for example, the query "Arts therapy" ($N=409$) resulted in a value of 0.995 with both models. However, in *CAMbase v3.0*, the same query ($N=1317$) resulted in a value of 0.727 with the general-purpose model and a value of 0.78 with the multi-language model.

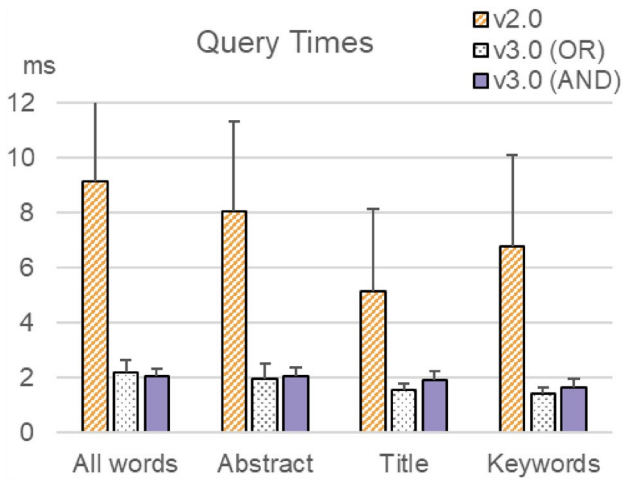

Fig. 4 Means of the query times separated into the restriction "All words", "Abstract", "Title", and "Keywords". The striped bar represents *CAMbase v2.0*, the dotted bar represents *CAMbase v3.0* by using the operator "OR", and the filled bar represents *CAMbase v3.0* by using the operator "AND". Means are partially taken from [20]

It was also observed that a large proportion of values was just below 1, even when the number of documents were equal on both systems. The reason for this is the cleaning process of the data. While, for example, umlauts were coded in *CAMbase v2.0*, all words were indexed in plain text in *CAMbase v3.0*. SBERT in this case also differentiated between titles when additional special characters such as a dot at the end or quotes for highlighting titles in data occurred. For a human user, those titles may be considered identical, but SBERT made a slight difference. The multi-language model seemed to rate these slight differences better. A general worsening or improvement between the models however could not be determined by contrasting the t-values in Table 3. The observation implies that Although SBERT showed differences in both systems, the document comparisons for most search queries are not significantly different. Only five ("Arts Therapy", "Autogenes Training", "Chinese Traditional Medicine", "Krebs", and "Massage") of the 36 search queries led to significant differences between *CAMbase v2.0* and *CAMbase v3.0*. The other queries led to means that are close to each other in both systems displayed in Fig. 5 by the strong slope of means at the right side.

Overall, the comparison of the reliability of the search with SBERT shows a promising result and thus that the legacy and the final system are very similar.

Lessons Learned

In this technical report, the realization and evaluation of a migration of an online literature database from a 32-bit to a 64-bit OS is presented. As the pure migration was unsuccessful, the proprietary search algorithm had to be replaced with *Apache Solr*, which changed the semantic search to a score-based search and required a data migration. By integrating *Solr* to *CAMbase v3.0*, the main goal of providing a useful and functional literature database for CAM could

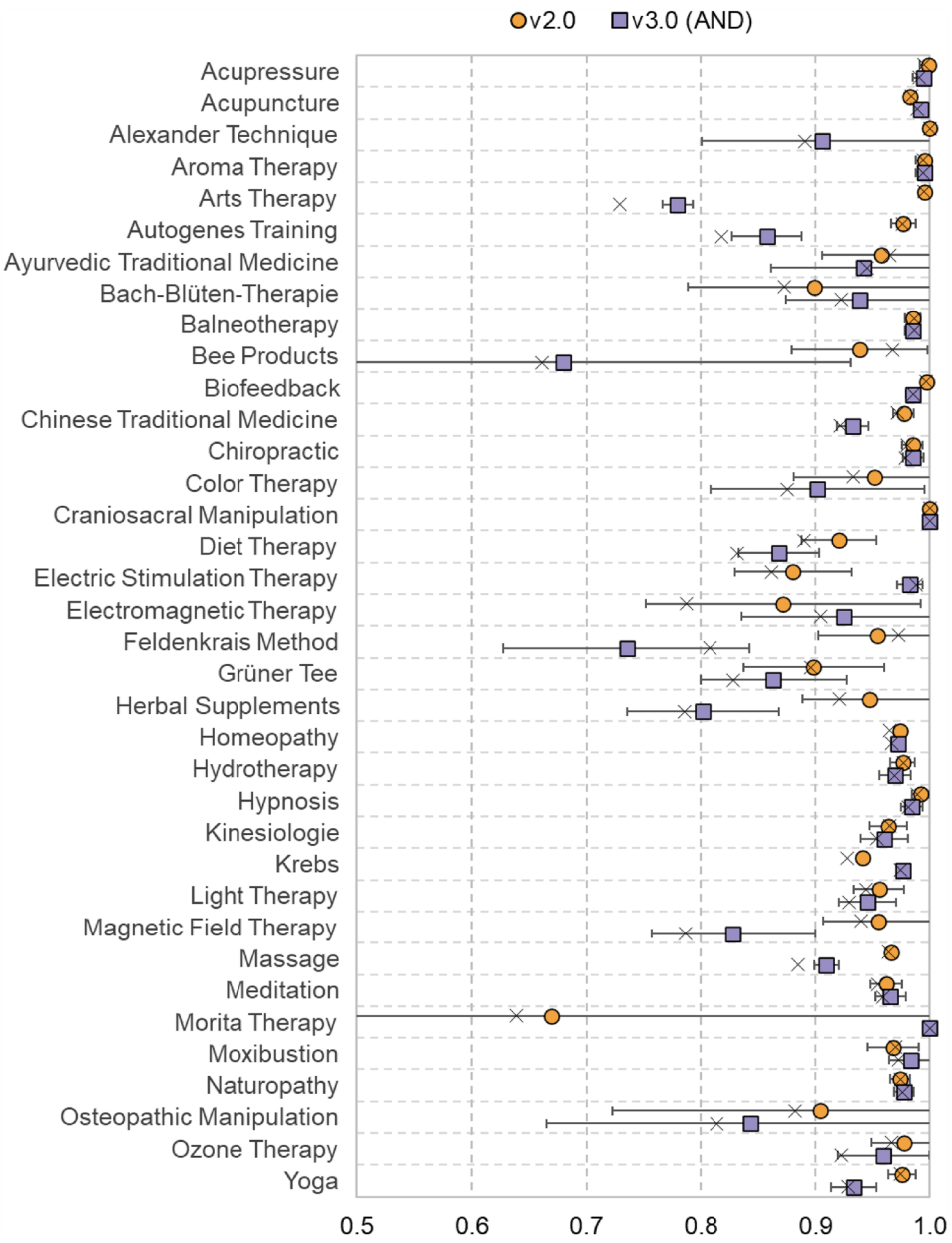

Fig. 5 Overview of means that are calculated from the syntactic analysis via the multi-language model of SBERT. Calculations were performed for *CAMbase v2.0* (circles) and *CAMbase v3.0* (squares) using the "AND" operator accordingly and separately for search queries. Error bars denote the 95% confidence interval and crosses donate the equivalent means of the general-purpose model from [20]

be achieved. The approach of implementing a ready-made search engine solution has been shown to be a good solution to provide similar search results for users without abandoning the graphical user interface and the modular structure given by the historically grown database.

Compared to the release date of *CAMbase v2.0*, there are notable good open-source search engine solutions available on the Internet today. *Solr* was chosen because there is a large community behind it that drives development. In addition, *Solr* 's documentation is quite extensive, covering different use cases which clearly helped with the installation, the configuration of the project, the import of the literature data into the project, and with even linking *Solr* to the GUI through an existing library solution. Despite the time investment, *Solr* remains a flexible solution that has even led to an increase in performance after replacing the legacy algorithm, which could be useful in similar projects.

A methodological limitation is given by the fact that this technical report omitted the calculation of sensitivity and precision of relevant and irrelevant retrieved documents suggested by Lefebvre et al. [42]. Although the analysis was not intended for this type of evaluation, it could be considered in further analysis or in an analysis of similar projects. Instead, SBERT, a derivation of the language representation model BERT [43], was used to ensure the quality aspects like data accuracy and data accessibility, which could be affected by the migration [44]. BERT itself has already proven to be a remarkable method for detecting similarities in textual or

Table 3 Alphabetically ordered list of search queries, which is derived from the list of Wieland et al. [40] and extended with German key terms, and the results of the *t* test for independent samples, whereas the t_1 values correspond to the general-purpose model and the t_2 values to the multi-language model

Search query	$\lvert t_1 \rvert$	$\lvert t_2 \rvert$
Acupressure	0.145	0.211
Acupuncture	0.986	1.72
Alexander technique	0.917	0.833
Aroma therapy	0.0	0.0
Arts therapy	10.626	9.44
Autogenes training	3.613	2.994
Ayurvedic traditional medicine	0.157	0.123
Bach-blüten-therapy	0.31	0.26
Balneotherapy	0.009	0.002
Bee products	1.165	9.995
Biofeedback	0.823	0.93
Chinese traditional medicine	2.265	2.203
Chiropractic	0.077	0.024
Color therapy	0.458	0.462
Craniosacral manipulation	–	–
Diet therapy	1.179	1.188
Electric stimulation therapy	1.717	1.508
Electromagnetic therapy	0.8	0.499
Feldenkrais method	1.006	1.177
Grüner tee	0.643	0.364
Herbal supplements	1.503	1.759
Homeopathy	0.105	0.169
Hydrotherapy	0.259	0.206
Hypnosis	0.351	0.385
Kinesiologie	0.292	0.096
Krebs	5.915	4.907
Light therapy	0.365	0.272
Magnetic field therapy	1.579	1.459
Massage	4.862	3.813
Meditation	0.143	0.147
Morita therapy	0.677	0.649
Moxibustion	0.06	0.353
Naturopathy	0.086	0.164
Osteopathic manipulation	0.292	0.29
Ozone therapy	0.707	0.374
Yoga	1.341	1.254

A value greater than or equal to 1.96 is considered as significant system difference. The t_1 values are taken from [20]

bibliographic data in similar contexts [45–47]. The results with SBERT were generally sufficient. According to these reults, the documents retrieved through the 36 specific search queries showed an overall high equality between *CAMbase v2.0* and *CAMbase v3.0*. The two chosen language models had some minor issues with additional punctuation marks or coded umlauts in German language, which were a bit higher in the general-purpose model (all-MiniLM-L6-v2) as in the multi-language model (paraphrase-multilingual-MiniLM-L12-v2). Nevertheless, both demonstrated similar, good quality results, which indicates their accuracy and robustness. The general-purpose model stands out a bit, as it was trained on English data and could still handle the mixture of English and German documents. Which model is better, depends on the purpose of the use case. In our case neither of the two models fits perfectly. The more optimal model should be a mixture of the two, trained in both languages without considering translations. As a recommendation, even if the results were sufficient, a model should be trained appropriately for its specific use case, e.g., by fine-tuning, which can lead to better results [48].

The addition of small qualitative surveys of user statements helped to ensure and improve the data quality as well. At first, *CAMbase v3.0* had significant change in retrieved documents when the operator was set to "OR", which users immediately noticed. Users could no longer find their usual literature but were delighted with the wide range of documents available, although it takes longer to find the right literature. However, it does not correspond to the goal of an equivalent online database. Therefore, the operator in *CAMbase v3.0* was finally set to "AND". Now, users state the literature as more accurate, which is in accordance with the former analysis, and have a much better experience [20]. The fact that *CAMbase v3*.0 is a new system was hardly noticed, which could be due to the remained GUI. In contrast to that, users miss the functionality of easily narrowing down their search with words from a thematic landscape [8]. This functionality has only been partially implemented. Instead, the search can now be manually influenced by Boolean operators. However, users will need a certain training period to use these new functions. This highlights the need of an online tutorial, a feature for further development. Yet, all statements came from only a few supportive users. A larger sample could reveal more critical and detailed statements, which can be collected in a more systematic, qualitative study.

Conclusions

The assessment of various parameters, e.g., after a data migration, is important for quality management of bibliographical data [49], especially for sensitive or confidential data such as in the medical field. Possible data changes could be measured and categorized to support the data quality [50]. User statements and a semantic textual analysis evaluated the data of this report. The combination of both resulted in a well-accepted final system.

In sum, this technical report may serve as blueprint for similar projects. If the implementation is followed carefully,

Solr can be considered to some extent as an alternative or replacement to a search engine that uses a semantic algorithm. In particular, the semantic text analysis via SBERT has proven to be a promising tool for quality management, which therefore is highly recommended and should be used and investigated in further analyses.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42979-023-02146-9.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The authors confirm that the data, supporting the findings of this study, are available within the supplementary materials.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Morris RJT, Truskowski BJ. The evolution of storage systems. IBM Syst J. 2003;42(2):205–17. https://doi.org/10.1147/sj.422.0205.
2. Barraclough ED. On-line searching in information retrieval. J Doc. 1977;33(3):220–38. https://doi.org/10.1108/eb026643.
3. Rogers FB. The development of MEDLARS. Bull Med Libr Assoc. 1964;52(1):150–1.
4. Ulrich EL, Markley JL, Kyogoku Y. Creation of a nuclear magnetic resonance data repository and literature database. Protein Seq Data Anal. 1989;2(1):23–37.
5. Boehm K, Raak C, Vollmar HC, Ostermann T. An overview of 45 published database resources for complementary and alternative medicine. Health Info Libr J. 2010;27(2):93–105. https://doi.org/10.1111/j.1471-1842.2010.00888.x.
6. Ostermann T, Zillmann H, Matthiessen PF. CAMbase–the realisation of an XML-based bibliographical database system for complementary and alternative medicine. Z Arztl Fortbild Qualitatssich. 2004;98(6):501–7.
7. Zillmann H. Information retrieval and search engines in full-text databases. LIBER Q. 2000;10(3):335–41. https://doi.org/10.18352/lq.7605.
8. Ostermann T, Raak CK, Matthiessen PF, Büssing A, Zillmann H. Linguistic processing and classification of semi structured bibliographic data on complementary medicine. Cancer Informatics. 2009;7:159–69.
9. Haake E, Blenkle M, Ellis R, Zillmann H. Nur die ersten Drei zählen! optimierung der rankingverfahren über popularitätsfaktoren bei der elektronischen bibliothek bremen (E-LIB). Obib. 2015;2(2):33–42. https://doi.org/10.5282/o-bib/2015H2S33-42.
10. Alhazmi OH, Malaiya YK. Quantitative vulnerability assessment of systems software. In: Alhazmi OH, editor. Annual reliability and maintainability symposium. Cham: Alexandria IEEE; 2005.
11. Alhazmi OH, Malaiya YK. Application of vulnerability discovery models to major operating systems. IEEE Transact Reliab. 2008;57(1):14–22. https://doi.org/10.1109/TR.2008.916872.
12. Kaluarachchilage PKH, Attanayake C, Rajasooriya S, Tsokos CP. An analytical approach to assess and compare the vulnerability risk of operating systems. Int J Comp Net Inform Sec. 2020;12(2):1–10. https://doi.org/10.5815/ijcnis.2020.02.0.
13. Elleithy KM, Blagovic D, Cheng WK, Sideleau P. Denial of service attack techniques: analysis, implementation and comparison. J Syst, Cybernet Inform. 2006;3(1):66–71.
14. Sundaram A. An introduction to intrusion detection. Crossroads. 1996;2(4):3–7.
15. Rao UH, Nayak U. Intrusion detection and prevention systems the InfoSec handbook. Berkeley. 2014. https://doi.org/10.1007/978-1-4302-6383-8_11.
16. Garcia M, Bessani A, Gashi I, Neves N, Obelheiro R. Analysis of operating system diversity for intrusion tolerance. Soft Pract Exper. 2014;44(6):735–70.
17. Haynes RB, Sackett DL, Richardson WS, Rosenberg W, Langley GR. Evidence-based medicine: how to practice & teach EBM. Can Med Assoc J. 1997;157(6):788.
18. MacPherson H, Peters D, Zollman C. Closing the evidence gap in integrative medicine. British Med J. 2009;339:335. https://doi.org/10.1136/bmj.b3335.
19. Richards M (2015) Software architecture patterns. Sebastopol: O'Reilly Media, Incorporated
20. Unger S, Raak C, Ostermann T. Search reliability comparison of two text-based search algorithms in an online literature database for integrative medicine: a technical report on a 32-bit to 64-bit Migration. In: Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. Valletta: SCITEPRESS 2022. pp. 148–57. Doi: https://doi.org/10.5220/0011589300003335
21. Yamaguchi F, Maier A, Rieck K. 64-Bit migration vulnerabilities. Inform Technol. 2017;59(2):73–81. https://doi.org/10.1515/itit-2016-0041.
22. Chang H, Karne R, Wijesinha A (2016) Migrating a bare PC web server to a multi-core architecture In: 2016 IEEE 40th annual computer software and applications conference. IEEE. Doi: https://doi.org/10.1109/COMPSAC.2016.15
23. Mathew AB, Pattnaik P, Madhu Kumar SD. Efficient information retrieval using Lucene, LIndex and HIndex in Hadoop. In: 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications. IEEE; 2014. pp. 333–40. Doi: https://doi.org/10.1109/AICCSA.2014.7073217
24. Hansen J, Porter K, Shalaginov A, Franke K. Comparing open source search engine functionality, efficiency and effectiveness with respect to digital forensic search. NISK 2018; (108)
25. Berryman J, Turnbull D. Relevant search: With applications for Solr and Elasticsearch. Shelter Island: Manning Publications Co; 2016.
26. Kılıç U, Karabey I. Comparison of solr and elasticsearch among popular full text search engines and their security analysis. In: future internet of things and cloud workshops, 2015 6th International Conference on. IEEE; 2016. pp. 163–68.
27. Luburić N, Ivanović D. Comparing apache solr and elasticsearch search servers. In: Proceedings of the 6th International Conference on Information Society and Technology. 2016. pp. 287–91.

28. The Apache Software Foundation. Apache Lucene - Welcome to Apache Lucene. https://lucene.apache.org/. Accessed 12 Jan 2023.
29. The Apache Software Foundation. Welcome to Apache Solr - Apache Solr. https://solr.apache.org/. Accessed 12 Jan 2023.
30. Glauner PO, Iwaszkiewicz J, Le Meur J-Y, Simko T. Use of Solr and Xapian in the Invenio document repository software. arXiv 2013; Doi: https://doi.org/10.48550/arXiv.1310.0250.
31. Grainger T, Potter T. Solr in action. Manning Publications Co; 2014.
32. Kumar J. Apache Solr search patterns. Birmingham: Packt Publishing; 2015.
33. Choi H, Kim J, Joe S, Gwon Y. Evaluation of BERT and ALBERT sentence embedding performance on downstream NLP Tasks. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE; 2021.
34. Peters ME, Ruder S, Smith NA. To tune or not to tune? Adapting pretrained representations to diverse tasks. arXiv 2019; Doi: https://doi.org/10.48550/arXiv.1903.05987.
35. Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-tuning bidirectional encoder representations from transformers (bert)-based models on large-scale electronic health record notes: an empirical study. JMIR Med Inform. 2019;7(3):e14830. https://doi.org/10.2196/14830.
36. Zhang Y, He R, Liu Z, Lim KH, Bing L. An unsupervised sentence embedding method by mutual information maximization. arXiv 2021; Doi: https://doi.org/10.48550/arXiv.2009.12061
37. Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. arXiv 2019; Doi: https://doi.org/10.48550/arXiv.1908.10084
38. Jahan MS, Khan HU, Akbar S, Farooq MU, Gul S, Amjad A. Bidirectional language modeling: a systematic literature review. Sci Program. 2021;2021:1–15. https://doi.org/10.1155/2021/6641832.
39. Wang B, Kuo C-CJ, Sbert WK. A Sentence embedding method by dissecting BERT-based word models. IEEE/ACM Trans Audio Speech Lang Process. 2020;28:2146–57.
40. Wieland LS, Manheimer E, Berman BM. Development and classification of an operational definition of complementary and alternative medicine for the Cochrane collaboration. Altern Ther Health Med. 2011;17(2):50–9.
41. Unger S, Ostermann T, Raak C. Comparison of two text-based search algorithms in an online literature database for integrative medicine – first results. 66. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS), 12. Jahreskongress der Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V. (TMF) 2021.
42. Lefebvre C, Glanville J, Beale S, Boachie C, Duffy S, Fraser C, Harbour J, McCool R, Smith L. Assessing the performance of methodological search filters to improve the efficiency of evidence information retrieval: five literature reviews and a qualitative study. Health Technol Assess. 2017;21(69):48. https://doi.org/10.3310/hta21690.
43. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv 2018; Doi: https://doi.org/10.48550/arXiv.1810.04805
44. Zamzami IF, Fatani HAA, Zammarah NAH. Data migration challenges: The impact of data quality—Case study of University Putra Malaysia UPM. In: 2011 International Conference on Research and Innovation in Information Systems. Malaysia: IEEE; 2011. pp. 1–5. Doi: https://doi.org/10.1109/ICRIIS.2011.6125732.
45. Kades K, Sellner J, Koehler G, Full PM, Lai TE, Kleesiek J, Maier-Hein KH. Adapting bidirectional encoder representations from transformers (BERT) to assess clinical semantic textual similarity: algorithm development and validation study. JMIR Med Infor. 2021;9(2):2279. https://doi.org/10.2196/22795.
46. Xu Y, Liu Q, Zhang D, Li S, Zhou G. Many vs. Many Query Matching with Hierarchical BERT and Transformer. In: CCF International Conference on Natural Language Processing and Chinese Computing. Dunhuang: Springer International Publishing; 2019. pp. 155–67.
47. Zhang L, Lu W, Chen H, Huang Y, Cheng Q. A comparative evaluation of biomedical similar article recommendation. J Biomed Inform. 2022;134:104106. https://doi.org/10.1016/j.jbi.2022.104106.
48. Rizzo J. Evaluating pre-trained language models on partially unlabeled multilingual economic corpora. Munich: 2022.
49. Van Kleeck D, Langford G, Lundgren J, Nakano H, O'Dell AJ, Shelton T. Managing bibliographic data quality in a consortial academic library: a case study. Catal Classi Quart. 2016;54(7):452–67. https://doi.org/10.1080/01639374.2016.1210709.
50. Zavalina OL, Shakeri S, Kizhakkethil P, Phillips ME. Uncovering Hidden Insights for Information Management: Examination and Modeling of Change in Digital Collection Metadata. In: International Conference on Information. Sheffield: Springer International Publishing; 2018. p. 645–51.