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

Sebastian Unger[1], Christa K. Raak[2] and Thomas Ostermann[1]
[1]*Department for Psychology and Psychotherapy, Faculty of Health, Witten/Herdecke University, Witten, Germany*
[2]*Center for Integrative Medicine, Faculty of Health, Witten/Herdecke University, Witten, Germany*

Keywords: Evaluation, Search Engine, Information Storage and Retrieval, Sentence-Pair Regression, Semantic Similarity, Online Publications.

Abstract: Although there is a steady increase of scientific publications in integrative medicine, it is still difficult to get a valid overview of published evidence. The open accessible bibliographical database CAMbase 3.0 (available at https://cambase.de) hosted by Witten/Herdecke University is one of such established databases in this field. In 2020, CAMbase 2.0 was migrated to a newer 64-bit operating systems, resulting in a variety of issues. A promising solution of keeping and accessing the data of CAMbase 2.0 was to replace the business logic with the open-source platform Solr, which uses a score ranking algorithm instead of a semantic-syntactic interpretation of search queries as in CAMbase 2.0. As a result, the before-after analysis with T-tests showed mainly no significant differences in the equality of the queried titles after applying SBERT, not even in the number of search hits ($t = 1.43$, $df = 35$, $p = 0.17$), but in query times ($t = 4.2$, $df = 35$, $p < 0.01$). While search hits remained stable as the speed increases, the approach with Solr is more efficient, making this technical report a possible blueprint for similar bibliography-based databases projects.

1 INTRODUCTION

1.1 Background

Evidence based practice, according to the early definition of Haynes (1997), is defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individuals". In line with this definition, Integrative medicine is often referred to as an approach that "emphasizes the therapeutic relationship between practitioner and patient, is informed by evidence, and makes use of all appropriate therapies" (Ostermann et al., 2013), including complementary therapies. Despite the development of complementary therapies in the last decades and the steady increase of scientific publications in integrative medicine, e.g., the 45 databases presented by Boehm et al. (2010), it is still difficult to get a valid overview of published evidence, especially on older publications. Thus, *CAMbase v2.0*, an open accessible bibliographical database, was launched in 2007 by a team of Witten/Herdecke University (Ostermann et al., 2007), using *CiXbase,* a semantic search algorithm (Ostermann et al., 2009), to assist the user in its online research.

In the meantime, operating systems (OS) have changed and are more commonly used with a 64-bit architecture instead of a 32-bit system as in *CAMbase v2.0*. To avoid becoming an easy target for various network attacks, e.g., exploits based on discovered vulnerabilities (Alhazmi and Malaiya, 2008) the underlying and outdated 32-bit OS of *CAMbase v2.0* had to be replaced with a newer one in order to keep *CAMbase v2.0* still accessible for the public without disposing any threats. Unfortunately, the semantic search algorithm did not run stable afterwards. Thus, it was decided to replace it with an existing score-based search algorithm, however by ensuring the search reliability and query speed.

The resulting hypothesis thus is that the newly implemented score-based search algorithms statistically does not show significant differences compared to the formerly implemented semantic search algorithm. Especially with respect to the search results, we aimed at answering the question: To what extent do the former semantic search results differ? Secondly, we wanted to explore if those changes had an impact with respect to performance

Unger, S., Raak, C. and 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.
DOI: 10.5220/0011589300003335
In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) - Volume 1: KDIR*, pages 148-157
ISBN: 978-989-758-614-9; ISSN: 2184-3228

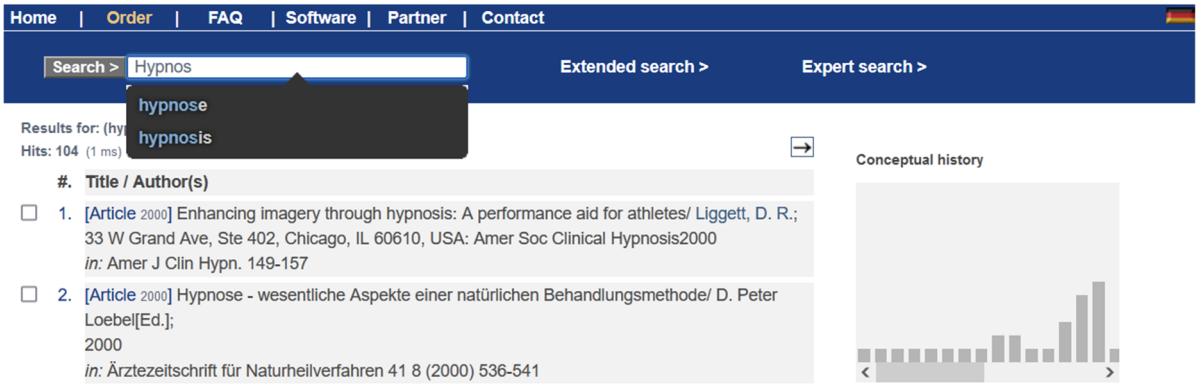

Figure 1: Screenshot of the GUI while typing the word “Hypnosis”.

times. Ideally, the new system is not noticeably different from the legacy system.

1.2 Requirements

With the migration of *CAMbase v2.0*, the necessary effort of keeping a running database should be as minimal as possible during the process and afterwards. This also includes some specific restrictions. First, the design of the GUI (Graphical User Interface) is like that of the legacy system (see Figure 1). Second, the XML-based (Extensible Markup Language) approach for a client-side construction described in (Ostermann et al., 2004) should also be retained to relieve the server. Third, the modular construction principle for a flexible modification and manageable system maintenance is retained. And finally, which is most important for the field of integrative medicine, the unique data collected over the years can still be accessed by the users.

2 STATE OF THE ART

2.1 Cambase V2.0

After an OS was set up in relation to the internal regulations of Witten/Herdecke University, there was a one-to-one migration of *CAMbase v2.0* from the 32-bit platform to the new 64-bit platform in 2020.

The original *CiXBase* search algorithm uses a semantic-syntactic analysis method to interpret text phrases, which went far beyond simple stemming methods at the time of development. In addition to bibliographic data, full lexical texts can be analyzed and indexed (Ostermann et al., 2015), making the method easily adaptable to other contexts. For that, it processes multilingual database queries in a natural language, using relevant methods of semantic information processing (error-sensitive syntactic parsing, morphology, composite decomposition). The recognized syntactic-semantic connections are not lost through a reduction to Boolean links. Hence, the system can differentiate between similar text phrases, e.g., “teaching in hospitals” and “hospitals in teaching” (Ostermann et al., 2009; Haake et al., 2015). A specially developed ranking algorithm weights, evaluates and relates the query results according to topics, restrictions, and modifications.

Although most requirements were met with the one-to-one migration, it resulted in a variety of issues in search queries in a pre-run, and, in accordance to published experiences (Wressnegger et al., 2017), led to a massive decrease in usability. These issues consist of interpreting umlauts correctly in both search queries and search hits. Further issues occurred in the display of search results, i.e., warning messages that overlapped with elements of the GUI, and HTML coding that could not be rendered correctly. One type of warning message stated that the search could not find some data parts. This could be due to the change from 32-bit to 64-bit system libraries or their handled data, which was stored in a text format with values separated by vertical bars. The file structure can be excluded as reason, as it remained the same during the migration.

Changing the underlying data or linking some libraries to 32-bit versions could not solve the issues, so that the reason for the issues must lie in the algorithm itself. Unfortunately, the files of the search algorithm were stored in binary format on the 32-bit system without its source code. This prohibits a full insight into the program’s code and thus a simple rework of the algorithm was no option. Although, the unique search algorithm of *CAMbase v2.0* could be retained with this one-to-one migration, the steadily occurring issues and the failing attempts to solve these issues forced a complete revision of *CAMbase v2.0* to maintain a positive user acceptance.

2.2 Search Engine Solutions

Since the development of *CAMbase v2.0*, advanced open-source search engines such as *Apache Lucene* or those based on it such as *Elasticsearch* and *Apache Solr* were established. All offer indexing algorithms and ranking processes and are available in many variations in the literature (Hansen et al., 2018; Turnbull & Berryman, 2016). Other search engines offer a completely pre-built GUI like *Open Semantic Search*, minimalizing the effort of building an online database even more. Therefore, five of the most preferred search engines with a brief overview of their key features are presented in the following sections.

2.2.1 Apache Lucene

One of the first search engine projects developed by the Apache Software Foundation (ASF) is *Apache Lucene* written completely in Java ("Apache Lucene - Welcome to Apache Lucene", 2021). Lucene is designed for full text searches, while search terms can be customized, e.g., with the use of wildcards, ranges of years, or text phrases. Search results are ranked with a best-comes-first approach, which uses the implemented powerful retrieval model.

2.2.2 Apache Solr

Another search engine managed by the ASF is *Solr*. This *Lucene*-based search engine is fast, open-source, and, like its predecessor, offers a wide range of customization ("Welcome to Apache Solr - Apache Solr", 2021). This includes scalable indexing, faceting, and hit highlighting, to name just a few. In contrast to the search of *CAMbase v2.0*, the search is based on a scoring algorithm (Kumar, 2015). Its score value involves a number of factors such as word frequency within a database entry or field length used for the search. However, the configuration allows a lot of flexibility with advanced analysis and tokenization features, e.g., separating terms by alphanumeric signs or using different types of stemming and phonetic methods.

2.2.3 Elasticsearch

Elasticsearch is a second search engine based on *Apache Lucene*. It is part of the Elastic Stack (ELK Stack) that features machine learning, security, and reporting, in addition to a set of use cases for analyzing and searching data ("Free and Open Search: The Creators of Elasticsearch, ELK & Kibana | Elastic", 2021). *Elasticsearch* allows to search, analyze, and visualize data of any type and size in real time. Although, it offers better data analytics with the integration in the ELK stack in comparison to *Solr*, both systems can be considered as equal in most use cases (Luburić & Ivanović, 2016).

2.2.4 Meilisearch

An alternative to the mentioned search engines, but not less inefficient, is *MeiliSearch*. This open-source solution was designed to be easily accessible and to meet most use cases, even specific ones ("MeiliSearch", 2021). Therefore, the installation requires little or no configuration, but leaves enough room for the experienced administrator of the engine to customize. To the most relevant features of *MeiliSearch* belong a response time of less than 50 ms (milliseconds), a search in a typo tolerant and natural language, and the use of synonyms to avoid limiting a search to specific words.

2.2.5 Open Semantic Search

An open-source search engine with an advanced GUI for simplifying configurations such as managing lists of synonyms or hyponyms is *Open Semantic Search*. It is based on *Apache Solr* or *Elasticsearch* and can be installed on a Debian or Ubuntu based Linux server ("Open Semantic Search: Your own search engine for documents, images, tables, files, intranet & news", 2021). A search is possible in full text data that is distributed over different data sources or file formats. Powerful search operators can also be used here to increase search hits. A major advantage of *Open Semantic Search* is that the GUI is already pre-built for the end-user, which drastically reduces the effort required for the development of a whole system. However, there is currently no possibility to redesign the GUI as an end-user itself.

3 CONCEPT

3.1 System Architecture

The original architecture of *CAMbase v2.0* can be divided into three main layers (see Figure 2), so that it is easy to develop a new database in principle by replacing one of the layers. On top, there is the presentation layer. It includes an XML-based GUI created on the client side to relieve the server. In the middle, there is the core element. This business layer interprets search queries semantically and syntactically far beyond simple stemming methods developed simultaneously. The database is located at

the bottom of the architecture. The bibliographical data recorded after the migration to the new server includes a total of 115355 bibliographical datasets as evidence of integrative medicine from 1906 onwards.

The promising alternative of keeping the data and making *CAMbase v2.0* still usable was to use one of the already existing search engines of chapter 2.2. This search engine must have the option to import the old data, so that the revision effort remains manageable**.** After reviewing these search engines, *Apache Solr* was chosen as it is well documented and there are library solutions for the communication available, which work with the underlying OS. According to the requirements of chapter 1.2, this kind of change would satisfy everything except the original semantic search algorithm.

3.2 Critical Issues

When the data records were transferred to the *Solr* database system, it turned out that there were some duplicate or incomplete publication references. In addition, fields in the database were represented multiple times for the same entry because the data were collected from different sources that had different labels, e.g., literature title or journal title. Other fields did not match in their type, e.g., in some sources the page number was used as an exact numerical value, while in others the range was given. Instead of transferring the complete data, those references were removed or corrected. Moreover, the data was stored in the ISO-8859-1 standard, so that various characters were not available as plain text but coded. Corrections were also made here so that the text was generally available in the UTF-8 standard. The changes inevitably lead to small differences between the databases, but ensure clean data in *CAMbase v3.0*.

4 IMPLEMENTATION

4.1 Layer Adjustment

In the first step, the binary files of the old search algorithm were removed from the server for the revision of *CAMbase v3.0*. This left the top and bottom layer of the architectural model. Then, *Solr v8.9.0* was installed on the same server followed by the import of the transformed and corrected data. Since the GUI needs to communicate with the search engine, another layer was added to the architecture as shown in Figure 2. This layer is written in PHP (PHP: Hypertext Preprocessor) in combination with an extension of PECL (PHP Extension Community Library) for *Solr*. The layer's task is to interpret and parse the input of users to the search engine and vice versa without interfering the user experience.

The next step was the configuration of *Solr*. As the search shall approximately equalize the semantic-syntactic algorithm of *CAMbase v2.0*, a search query is now interpreted and simplified by a light stemming method. Essentially, this implementation shall cover different spellings and result in a wide range of search hits. The techniques provided by *Solr*, such as searching for entire groups of words by narrowing them down in execution characters, were also applied.

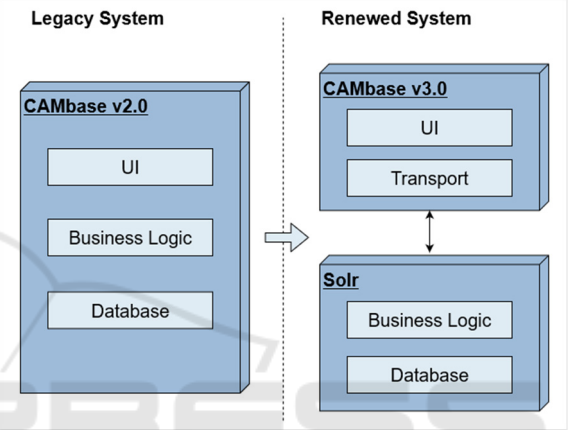

Figure 2: Main layers of the legacy (left) and the renewed system (right) divided into two separate subsystems.

Furthermore, *Solr's* default operator was set to "AND" as this leads to an intersection of all search hits produced by every single word of a search query, expecting more relevant results. A semantical interpretation with the use of synonyms were not implemented as the effort for the creation and maintenance of such a list of synonyms was considered too costly.

4.2 Comparisons

The comparison of both databases is designed as a before-after analysis with two time points in which *CAMbase v2.0* served as initial state of the system, in which user inputs are still interpreted syntactically and semantically. The completed migration and revision of *CAMbase v3.0* represents the second state. Because *Solr* has replaced the business logic, a score ranking search operated at this time. For multiple words, an intersection is now returned by using the "AND" operator. In the evaluation period, no further changes or deletions of database entries were made.

In order to evaluate both database systems, 36 common word groups (see Table 2 or Figure 4) from

the field of integrative medicine were used as the basis for the search. This list is derived from the list of Wieland et al. (2011) and is supplemented with additional German key terms suggested by experts from the field. Identical queries then were carried out in each system.

Additionally, each word group was used four times to examine the search hits within four search restrictions ("All words", "Keywords", "Abstract", "Title"). The query time, search hits, and all received titles were taken from the GUI and manually referenced in an Excel sheet for the data analysis.

In a first step of the analysis, the amount of search hits as well as the processing speed produced by both systems were compared for numerical equality. In a second step, the received titles produced by both systems were compared for syntactic equality. For that, a general purpose model (all-MiniLM-L6-v2) of SBERT (Sentence Bidirectional Encoder Representations from Transformers; Reimers and Gurevych, 2019) was used. SBERT is a derivation of the pretrained BERT network (Devlin et al., 2018) and produces semantically meaningful sentence embeddings. For that, SBERT receives at least two sentences, which are encoded and subsequently compared based on cosine similarity. The result is a value between -1 and 1. As values below 0 would indicate an opposite correlation and are not expected, this report uses for better representation only an interval from 0 to 1 that can be expressed as equality, i.e., the higher the value, the better the equality. As SBERT can be considered as state-of-the-art in various sentence classification and sentence-pair regression tasks using an efficient similarity measure method on modern hardware, it is optimal to compare the tens of thousands of received titles of this technical report. The comparison went both ways, i.e., first it was checked whether the received titles of *CAMbase v2.0* occur in the hits of *CAMbase v3.0* and then again in the opposite direction.

All data was analyzed using t-test statistics with a level of significance of 5 %. Comparison results thus are displayed as means with 95 % confidence intervals. Statistical analysis was performed with functions of Microsoft EXCEL for Windows.

5 RESULTS

The top row of Table 1 shows that the intersection of the search hits for the chosen common word groups in *CAMbase v3.0* is not statistically different to those hits produced by the former semantic-syntactic algorithm. However, an overall performance increase could be observed. The T-test provided statistical evidence of those differences between the systems for nearly all search restrictions as stated in the bottom row of Table 1. Only the restriction "Title" slightly missed to be significant.

Table 1: Results of the paired T-test that compares the mean query times and search hits of *CAMbase v2.0* and *CAMbase v3.0*, whereas the default operator used in *CAMbase v3.0* is "AND".

	Restriction	t	df	p
Search hits	All words	1.43	35	0.17
	Abstract	0.45	35	0.66
	Title	1.32	35	0.2
	Keywords	1.6	35	0.12
Query times	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

A graphical overview of the resulting means of the search hits of the first step is shown in Figure 3. On the one hand, the differences in the search hits of *CAMbase v3.0* compared to the legacy system are minimal. The restriction "All words" even resulted in. fewer titles with *CAMbase v2.0* ($\bar{x} = 193$) than with *CAMbase v3.0* ($\bar{x} = 210$). Considering that *CAMbase v2.0* was equipped with a semantic-syntactic search algorithm, the search for synonyms should increase the hits. On the other hand, the search

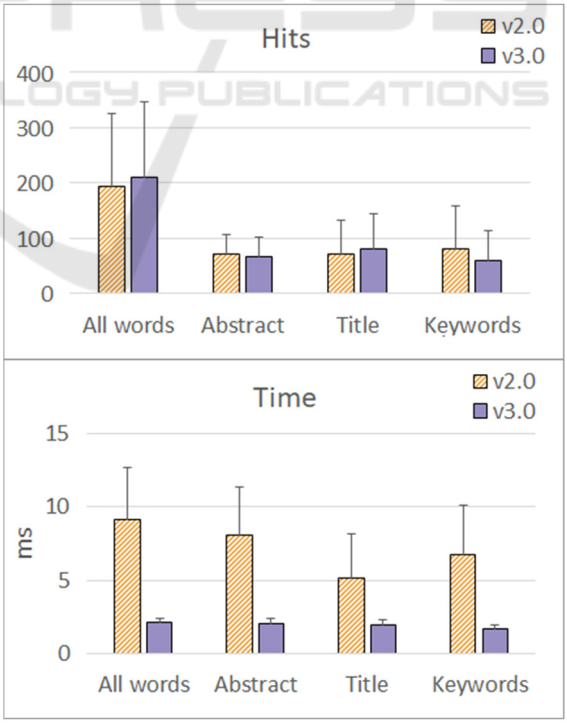

Figure 3: Histograms of means for search hits (top) and query time (bottom) grouped by restrictions. Error bars denote the 95% confidence interval.

speed improves with the searching algorithm of the new system. This is clearly visible in the decrease of the bars displayed in the bottom histogram. It demonstrates that the processing of queries via *Solr* is more efficient in comparison to that of *CAMbase v2.0* with the former syntactic-semantic search algorithm. The analysis with SBERT generally showed a high degree of agreement between the hits of the two systems in both directions. None of the calculated means was below *0.5*. The best result was achieved for the word group "Craniosacral Manipulation". Search hits were exactly the same in the opposing systems, so that an equality of *1* (*N = 2*) was achieved for both. The smallest equality achieved by the legacy system was for the word group "Morita Therapy" with *0.639* (*N = 3*) and by the new system for "Bee Products" with *0.661* (*N = 11*). This small equality is one-sided in each case. The explanation lies in the small set of hits for these word groups, whereas the system with the low equality had a few more hits that obviously cannot be present in the other set of hits. This becomes even clearer when looking in the other direction, using the same word groups. The equality is then *0.968* (*N = 7*, using "Bee Products") with the legacy system and even *1* (*N = 2*, using "Morita Therapy") with the new system.

Next to the smallest one-sided equality that were found, further one-sided outliers can be explained by too big gaps between the sets of hits, when there were relatively high hits like in the word group "Arts therapy". For this word group, the equality is *0.995* (*N = 409*) with the legacy system and, in contrast, only *0,727* (*N = 1317*) with the new system. In general, the differences between the hits are getting higher as the gap gets bigger, i.e., the higher the gap, the more titles are missing in the opposite set of search hits.

Despite outliers, a large proportion of the achieved equality was still just under *1* even if the sets of hits were equal on both sides. Here, the main explanation lies in the slightly different text structure in which the titles were stored. Among others, the legacy system sometimes used a dot at the end of a title or quotes for highlighting titles in data files because data was collected from different sources as described in chapter 3.2. Those textual special characters are missing in the data of the new system, so that two titles could be considered the same by humans, but are slightly differentiated by SBERT. Nevertheless, the most calculated means of both systems for the respective word groups are close to each other. A strong slope of the means at the right side and their approximately equal positions to their counterpart visualize these observations clearly in Figure 4. The associated standard errors in the form of error bars show also that the means do not differ significantly to their counterpart in these cases despite larger distances.

Table 2: Results of the T-test for independent samples. The values of SBERT were averaged for each word group in accordance with the system and used to calculate the shared T-value, which, for simplification reasons, is considered as significant if greater than or equal to 1.96.

Word group	t
Acupressure	0.145
Acupuncture	0.986
Alexander Technique	0.917
Aroma Therapy	0
Arts Therapy	10.626
Autogenes Training	3.613
Ayurvedic Traditional Medicine	0.157
Bach-Blüten-Therapie	0.31
Balneotherapy	0.009
Bee Products	1.165
Biofeedback	0.823
Chinese Traditional Medicine	2.265
Chiropractic	0.077
Color Therapy	0.458
Craniosacral Manipulation	-
Diet Therapy	1.179
Electric Stimulation Therapy	1.717
Electromagnetic Therapy	0.8
Feldenkrais Method	1.006
Grüner Tee	0.643
Herbal Supplements	1.503
Homeopathy	0.105
Hydrotherapy	0.259
Hypnosis	0.351
Kinesiologie	0.292
Krebs	5.915
Light Therapy	0.365
Magnetic Field Therapy	1.579
Massage	4.862
Meditation	0.143
Morita Therapy	0.677
Moxibustion	0.06
Naturopathy	0.086
Osteopathic Manipulation	0.292
Ozone Therapy	0.704
Yoga	1.341

A T-test for independent samples strengthened the observations. The results were mostly not significant, although differences exist as demonstrated with SBERT. There is only significance in the five common word groups "Arts Therapy" (*t = 10.626*), "Autogenes Training" (*t = 3.613*), "Chinese Traditional Medicine" (*t = 2.265*), "Krebs" (*t = 5.915*), and "Massage" (*t = 4.862*). Overall, this indicates a high level of agreement and thus equality between the opposing systems *CAMbase v2.0* and *CAMbase v3.0*.

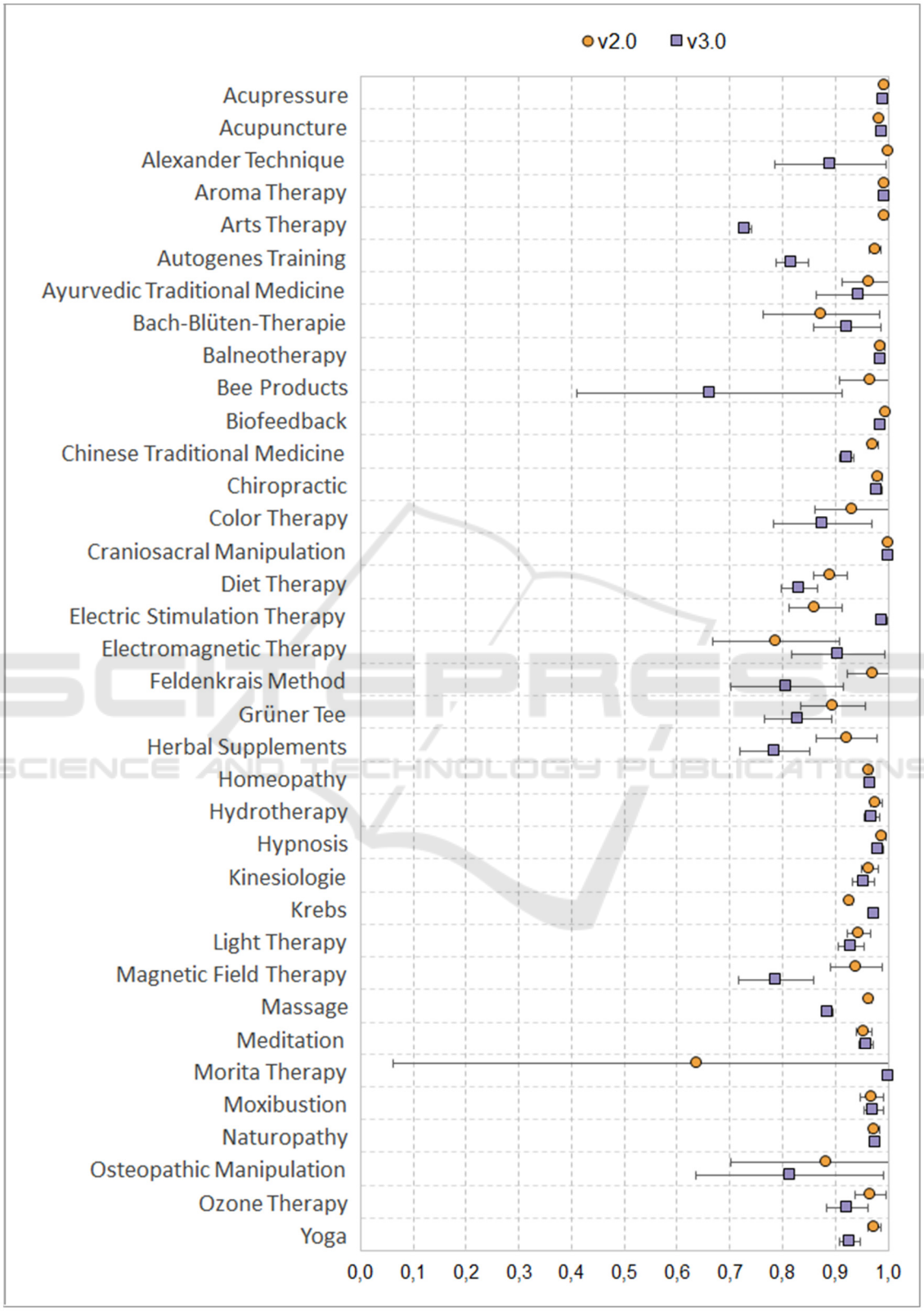

Figure 4: Summery of the syntactic analysis of the received titles using SBERT. Circles (v2.0) and squares (v3.0) mark the mean value of the calculated equality within the 36 common word phrases. Error bars denote the 95% confidence interval.

6 DISCUSSION

Quality management in the field of database migration is an important issue, which might affect usability in multiple dimensions, i.e., with respect to data accessibility, search accuracy, and performance (Zamzami et al., 2011). In particular, this is challenging when bibliographical metadata is concerned (Van Kleeck et al., 2016) and automated methods have been suggested for calculating metadata quality metrics (Zavalina et al., 2018).

In the adaption of *Solr* to *CAMbase v3.0*, the important goal, making the historically grown database accessible to the public, was satisfactorily achieved. Quality management was performed on the basis of a derivation of the BERT network, which itself has been introduced as a noteworthy method for detecting similarities in textual or bibliographical data. The numerous projects, e.g., the comparative evaluation of biomedical similar article recommendation (Zhang et al., 2022), the evaluation of similarity in clinical texts (Kades et al., 2021), or the query matching (Xu et al., 2019) that is most similar to our approach, can confirm this. Although the model of SBERT used here was trained only with English data, it could handle a mixture of English and German titles including some umlauts, showing its robustness. The use of one model was sufficient for the purpose of this report, as a comparison of the sets of search hits should be performed as a whole. Nevertheless, it is recommended to use a corresponding trained model for data of different languages. Apart from the use of SBERT, our approach described here also meets other requirements from chapter 1.2 to a good degree. The GUI, which is already comfortable for its users, and the modular construction principle could be retained next to the unique database. Furthermore, the construction of the GUI for relieving the server remains on the client side. Even if the search is no longer semantic-syntactic, this approach is preferable.

The results found in our analysis are in accordance with some user statements. After a short familiarization phase, users perceived the results as very accurate and did not notice any changed content. However, they stated that *CAMbase v2.0* included the option to narrow down the search results with a thematic landscape mapping of “keywords” (Ostermann et al., 2009), which was also implemented in other bibliographical databases such as *ArtheData* (Elbing et al., 2009). Although this indeed was a very innovative tool for both researchers and practitioners, it was only partially implemented. Nevertheless, the new possibility to influence the search by manually entering Boolean operators and other modifications, which was not supported in *CAMbase v2.0*, was well accepted by them.

Currently, there are many promising search engine solutions that are open-source and freely available. Some have been improved since the development of *CAMbase v2.0*, others have been newly added. A general statement about which of the search engines is the best cannot be made without any elaboration in this technical report, as this depends too much on the purpose of the application and the environment conditions. However, an advantage of *Solr* is its great popularity, so there is a large community behind it that drives development. It also offers a lot of configuration opportunities, which can be time costly, but increases flexibility in use cases. Flexibility has been demonstrated in importing the old data from a text format to a new but identical data structure in this migration process. Additionally, the preliminary queries could easily be parsed into a *Solr*-understandable syntax with the help of the PECL extension. However, some other search engines offer these possibilities, too. The choice of a database solution thus depends more on the administrator's knowledge or specific aspects.

7 CONCLUSIONS

This article describes the analysis of search reliability and query speed of the new architecture and the content of *CAMbase v3.0*, an open accessible bibliographical database on complementary therapies hosted by Witten/Herdecke University. Its underlying system architecture was completely renewed in 2020 and adapted to *Solr*, an open-source platform based on Apache Lucene for score ranking searches. The search results were evaluated in comparison to the former semantic-search algorithm. A limitation of this technical report is that the calculation of sensitivity and precision of relevant and irrelevant search hits suggested by Lefebvre et al. (2017) was omitted. However, as this analysis does not address such an evaluation of search hits, the limitation is negligible, but worth mentioning for similar projects.

As the intention was an equal system, the intersections of *CAMbase v3.0* and the search hits of the former algorithm were compared statistically, syntactically, and with user experiences. On the one hand, the T-test showed high equality in search hits and SBERT’s sentence embeddings of the received titles, which resulted in a consistent user experience in this aspect. On the other hand, there was a performance increase in *CAMbase v3.0*. The adaption

of *Solr* into *CAMbase v2.0* has therefore not significantly changed the search results of a literature database for integrative medicine but improved its efficiency.

Although, the hits were generally stated as accurate by users, they needed a certain time of familiarization with the new features. To make it easier, the provision of a tutorial for operating with the new system and its added features is currently prepared. We therefore intend to collect further statements in a more systematic, qualitative study.

In a nutshell, the analysis applied in this technical report may serve as a possible blueprint for quality assurance of similar projects in the field of bibliographical databases. This is not intended to be a guarantee of success, but rather implies that SBERT is a promising tool that should be used and explored in further analyses.

ACKNOWLEDGEMENTS

We would like to thank all our colleagues who have supported us with the migration and the subsequent verification of the cleaned and transferred data.

REFERENCES

Alhazmi, O. H., & Malaiya, Y. K. (2008). Application of vulnerability discovery models to major operating systems. *IEEE Transactions on Reliability, 57*(1), 14–22. https://doi.org/10.1109/TR.2008.916872

Apache Lucene - Welcome to Apache Lucene. (2021, September 8). Retrieved from http://lucene.apache.org/

Free and Open Search: The Creators of Elasticsearch, ELK & Kibana | Elastic. (2021, September 8). Retrieved from https://www.elastic.co/

Boehm, K., Raak, C., Vollmar, H. C. & Ostermann, T. (2010), An overview of 45 published database resources for complementary and alternative medicine. *Health Information & Libraries Journal, 27*, 93–105. https://doi.org/10.1111/j.1471-1842.2010.00888.x

Elbing, U., Schulze, C., Zillmann, H., Raak, C. K., & Ostermann, T. (2009). Arthedata—An online database of scientific references on art therapy. European Journal of Integrative Medicine, 1(1), 39-42. https://doi.org/10.1016/j.eujim.2009.01.001

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805

Haake, E., Blenkle, M., Ellis, R., & Zillmann, H. (2015). Nur die ersten Drei zählen! Optimierung der Rankingverfahren über Popularitätsfaktoren bei der Elektronischen Bibliothek Bremen (E-LIB). *o-bib. Das offene Bibliotheksjournal, 2*(2), 33–42. https://doi.org/10.5282/o-bib/2015H2S33-42

Hansen, J., Porter, K., Shalaginov, A., & Franke, K. (2018). Comparing Open Source Search Engine Functionality, Efficiency and Effectiveness with Respect to Digital Forensic Search. *NISK Journal, 12*(2019), 1893-6563. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2584227

Haynes, B. R., Sackett, D. L., Richardson, S. W., Rosenberg, W., & Langley, R. G. (1997). Evidence-based medicine: How to practice & teach EBM. *Canadian Medical Association. Journal (CMAJ), 157*(6), 788. Conference Proceedings

Kades, K., Sellner, J., Koehler, G., Full, P. M., Lai, T. E., Kleesiek, J., & Maier-Hein, K. H. (2021). Adapting bidirectional encoder representations from transformers (BERT) to assess clinical semantic textual similarity: algorithm development and validation study. JMIR medical informatics, 9(2), e22795.

Kumar, J. (2015). *Apache Solr Search Patterns*. Birmingham: Packt Publishing Ltd.

Lefebvre, C., Glanville, J., Beale, S., Boachie, C., Duffy, S., Fraser, C., Harbour, J., McCool, R., & Smith, L. (2017). 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*, *21*(69). https://doi.org/10.3310/hta21690

Luburić, N., & Ivanović, D. (2016). Comparing apache solr and elasticsearch search servers. In M. Zdravković, M. Trajanović & Z. Konjović (Eds.). *6th International Conference on Information Society and Technology: ICIST 2016* (pp. 287–291). http://www.eventiotic.com/eventiotic/files/papers/url/icist2016_54.pdf

MeiliSearch. (2021, September 8). Retrieved from https://www.meilisearch.com/

Open Semantic Search: Your own search engine for documents, images, tables, files, intranet & news. (2021, September 8). Retrieved from https://www.opensemanticsearch.org/

Ostermann, T., Beer, A.-M., Bankova, V., & Michalsen, A. (2013). Whole-Systems Research in Integrative Inpatient Treatment. *Evidence-Based Complementary and Alternative Medicine (ECAM), 2013*(962729). https://doi.org/10.1155/2013/962729

Ostermann, T., Malik, M., & Raak, C. (2015). The Use of Extensible Markup Language (XML) to Analyse Medical Full Text Repositories: An Example from Homeopathy. In M. Helfert (Ed.), *Proceedings of the 4th International Conference on Data Management Technologies and Applications* (pp. 219–224). SCITEPRESS. https://doi.org/10.5220/0005484002190224

Ostermann, T., Raak, C. K., Matthiessen, P. F., Büssing, A., & Zillmann, H. (2009). Linguistic processing and classification of semi structured bibliographic data on complementary medicine. *Cancer Informatics, 7*, 159–169. https://doi.org/10.4137/cin.s1182

Ostermann, T., Zillmann, H., & Matthiessen, P. F. (2004). Cambase: The realisation of an xml-based

bibliographical database system for complementary and alternative medicine. *Zeitschrift fur Arztliche Fortbildung und Qualitatssicherung, 98*(6), 501–507. https://europepmc.org/article/med/15527194

Ostermann, T., Zillmann, H., Raak, C. K., Buessing, A., & Matthiessen, P. F. (2007). CAMbase: A XML-based bibliographical database on Complementary and Alternative Medicine (CAM). *Biomedical Digital Libraries, 4*(2). https://doi.org/10.1186/1742-5581-4-2

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*. https://doi.org/10.48550/arXiv.1908.10084

Turnbull, D., & Berryman, J. (2016). *Relevant search: With applications for Solr and Elasticsearch*. Retrieved from http://proquest.tech.safaribooksonline.de/9781617292774

Van Kleeck, D., Langford, G., Lundgren, J., Nakano, H., O'Dell, A. J., & Shelton, T. (2016). Managing Bibliographic Data Quality in a Consortial Academic Library: A Case Study. Cataloging & Classification Quarterly, 54(7), 452-467. https://doi.orPg/10.1080/01639374.2016.1210709

Welcome to Apache Solr - Apache Solr. (2021, September 8). Retrieved from https://solr.apache.org/

Wieland, L. S., Manheimer, E., & Berman, B. M. (2011). Development and classification of an operational definition of complementary and alternative medicine for the cochrane collaboration. *Alternative Therapies in Health and Medicine, 17*(2), 50–59.

Wressnegger, C., Yamaguchi, F., Maier, A., & Rieck, K. (2017). 64-bit migration vulnerabilities. *it - Information Technology, 59*(2), 73–81. https://doi.org/10.1515/itit-2016-0041

Xu, Y., Liu, Q., Zhang, D., Li, S., & Zhou, G. (2019). Many vs. many query matching with hierarchical bert and transformer. In CCF International Conference on Natural Language Processing and Chinese Computing (pp. 155-167). Springer, Cham. https://doi.org/10.1007/978-3-030-32233-5_13

Zamzami, I. F., Fatani, H. A. A., & Zammarah, N. A. H. (2011, November). 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 (pp. 1-5). https://doi.org/10.1109/ICRIIS.2011.6125732

Zavalina, O. L., Shakeri, S., Kizhakkethil, P., & Phillips, M. E. (2018). Uncovering Hidden Insights for Information Management: Examination and Modeling of Change in Digital Collection Metadata. In International Conference on Information (pp. 645-651). Springer, Cham. https://doi.org/10.1007/978-3-319-78105-1_74

Zhang, L., Lu, W., Chen, H., Huang, Y., & Cheng, Q. (2022). A comparative evaluation of biomedical similar article recommendation. Journal of Biomedical Informatics. https://doi.org/10.1016/j.jbi.2022.104106