Human Interaction and Emerging Technologies (IHIET-AI 2025), Vol. 161, 2025, 1–11
https://doi.org/10.54941/ahfe1005891

Using Compact Retrieval-Augmented Generation for Knowledge Preservation in SMBs

Erik Schönwälder, Martin Hahmann, and Gritt Ott

Technische Universität Dresden, Dresden, 01062, DEU, Germany

ABSTRACT

Knowledge preservation is a critical challenge for small and medium-sized businesses (SMBs). Employee fluctuation and evolving work tasks create a permanent risk of knowledge and experience loss. Therefore, SMBs need effective and efficient strategies for knowledge retention. As most knowledge in companies is primarily encoded as language or text, large language models (LLMs) offer a promising solution for the preservation and utilization of knowledge. However, despite their strengths, their adoption and deployment are challenging. To address this issue, we propose a system based on the Retrieval-Augmented Generation (RAG) concept that combines small, locally run language models with traditional retrieval algorithms to significantly enhance the process of knowledge preservation and utilization by reducing search efforts.

Keywords: Retrieval-augmented generation, Large language models, Knowledge preservation

INTRODUCTION

Information in general, and experiential knowledge in particular, are defining factors in the digitalized working world. Whenever an employee leaves the company, there is a risk that important specialist knowledge and experience will be permanently lost. Additionally, innovations and new technical solutions continuously reshape work activities, requiring employees to identify and process new knowledge and information at a rapid pace. To address these challenges and remain competitive, SMBs need effective and efficient strategies for knowledge retention, especially when expertise is concentrated in only a few individuals. Otherwise, operational continuity is jeopardized, innovation is hindered, and productivity potential remains untapped. The future role of experiential knowledge is currently a topic of expert discussion. While some authors primarily associate the relevance of experiential knowledge with the training and refinement of AI-based assistance systems, other authors – including us – are convinced that experiential knowledge will retain its importance in the long term. It remains essential for managing complex processes and responding to unforeseen events in a volatile, uncertain, complex, and ambiguous (VUCA) world (Richter and Draude, 2023; Bruns et al., 2016).

Until now, strategic decisions for dealing with experience-based knowledge have been based either on a personification strategy (orientation towards personal knowledge exchange) or a codification strategy (technical processing options for explicit experience-based knowledge) (Hinkelmann and Witschel, 2012). In this paper, we demonstrate how both approaches can be integrated into a unified concept using modern language models. We have chosen this technology because company knowledge is predominantly encoded in language or text. It exists in explicit forms, such as manuals and reports, or as tacit knowledge/experience, which must be made explicit by employees through processes like interviews and transcription. Large language models (LLMs) make data stored in text formats more accessible and show great potential in summarization, analysis, and contextual retrieval. They are therefore a promising solution for the preservation and use of knowledge. However, despite their strengths, there are several challenges that hinder the adoption of LLMs in smaller organizations. LLMs require significant computing capacity to operate smoothly, which SMBs generally cannot afford. The on-premise operation would require significant investment in hardware and maintenance. Alternatively, cloud-based services such as ChatGPT or Microsoft Copilot offer external hosting for LLMs. While this saves investment, it raises the issue of data protection and security when working with sensitive business information. In addition, the explainability of how they work is a well-documented problem of LLMs, which makes them unsuitable at first glance for scenarios that require transparent decision-making.

To address these challenges, we propose a system based on the Retrieval-Augmented Generation (RAG) concept. RAG optimizes the output of an LLM by providing company-specific context that is not part of the LLM. Our approach combines small, locally run language models with traditional retrieval algorithms to provide more accurate results. Given a base of data containing company knowledge, such as a folder of PDF manuals, the system enables users to query this data using natural language. Initially, our system uses conventional retrieval approaches to select the documents most relevant to the query. Subsequently, this selection is processed by a small language model to find the most relevant passages within the documents, which are then ranked and presented as answers. Our system produces answers that are on par with the results of LLMs. Due to the efficient retrieval pipeline, our system has significantly lower computation requirements, even though it achieves state-of-the-art performance.

Like most systems based on AI or machine learning models, our system cannot be implemented as an off-the-shelf solution. Rather, it is a "wicked problem" (Pfaffl et al., 2022) due to the specific nature of the company and the resulting many influencing factors. The implementation addresses not only the necessary technical capabilities but also the organization of company processes, employee acceptance, and much more. Therefore, the system we propose should be considered as part of a holistic and integrated development process that targets a human-centric AI application. Deployment of our proposed solution in SMBs significantly reduces search efforts. Thus, greatly simplifying the utilization of existing knowledge and expertise by employees.

Since our approach is based on the processing of unstructured textual data, the focus of data collection can shift from labor-intensive preparation and cleansing to continuously gathering meaningful information and experience-based knowledge. This enhances the process of knowledge preservation and curation, while also supporting the work-integrated learning process of employees. Overall, this leads to improvements in company flexibility, performance, and workplace attractiveness.

BACKGROUND

This section provides an overview of the key concepts and algorithms used in our proposed system. We begin with a general introduction to information retrieval systems, followed by an outline of strategies for adjusting the ranking of retrieved documents. Next, we describe the incorporation of language models into these systems, before finally addressing the human-centered approach for deploying our system in SMBs.

Information Retrieval – BM25: The problem of retrieving relevant information based on a user query is typically referred to as Information Retrieval (IR). Search engines like Google or Bing implement IR systems to find websites that match a user's query. Over time, BM25 (Robertson and Walker, 1994; Robertson et al., 1994), a retrieval algorithm, has dominated the landscape of IR approaches. By representing each document and query as a vector within a vector space, BM25 ranks documents according to their similarity to the given query (Zhu et al., 2024). These vectors are determined by a combination of term frequency (TF) – how often a term appears in a document –, inverse document frequency (IDF) – how rare a term is across the entire corpus of documents – and a normalization factor based on document length. Using these measures, a document is considered more relevant if it contains more occurrences of the query terms, especially those that are rare across the corpus (Robertson et al., 2004). While BM25 is highly efficient and effective for term-based retrieval, it faces challenges in understanding the nuances of natural language, such as synonyms, paraphrases, and context beyond simple keyword matches.

Reranking Documents – Cross Encoder: To address the shortcomings of traditional IR systems like BM25, reranking systems aim to refine the initial ordering of the retrieved documents. With a specific emphasis on the quality of document ranking, more complex and computationally intensive methods are utilized to incorporate deeper semantic understanding, beyond the plain structural information of the documents (Zhu et al., 2024). The cross-encoder architecture (Humeau et al., 2020), a common choice for reranking systems, considers documents and queries pairwise rather than independently. More precisely, it encodes the query and each document together into a single vector and predicts a relevance score for the pair. Unlike BM25, which computes vectors for the query and each document independently, cross-encoders allow for direct interaction between the query and documents, enabling them to capture complex semantic relationships. This makes cross-encoders particularly powerful for reranking, as they can incorporate details like contextual relevance, that traditional systems like

BM25 cannot encode. However, due to the cross-encoder's complexity and computational effort, it is usually applied as a reranking system on a subset of relevant documents rather than on the entire document corpus.

Large Language Models (LLMs) – RAG: The combination of traditional retrieval systems and rerankers retrieves documents relevant to a given query but remains limited in its ability to represent information. While users expect the system to respond in a human-like manner with fully worded answers, it instead retrieves relevant but unprocessed documents. To address this issue, the retrieval-augmented generation (RAG) approach incorporates LLMs, such as GPT-3 (Brown et al., 2020), which can generate text based on input, enhancing the system with fully worded responses. Pre-trained on vast volumes of text data, LLMs achieve remarkable results in natural language processing tasks, including text generation, and classification. Due to their ability to capture contextual information within texts, they can understand natural language, including its semantic patterns. In the RAG approach, the input context of an LLM is enriched with relevant documents retrieved by a traditional retrieval system. These documents serve as an external knowledge base for generating a response, improving the quality and accuracy of the LLM's output, and expanding its knowledge beyond its pre-training data (Fan et al., 2024). By combining the strengths of retrieval-based systems and generative models, RAG provides relevant documents as factual grounding while ensuring grammatically correct, coherent, and fully worded answers.

Human-centered solution development: This holistic approach integrates technical, organizational, and personnel considerations, emphasizing the importance of stakeholder engagement – especially from the intended users. Such engagement is critical to gaining acceptance and ensuring the success of innovative technical solutions. A key feature of a human-centered solution is designing work tasks that are comprehensive and support learning, while also avoiding overburdening or underutilizing employees.

To achieve this, our approach focuses on distributing functions between employees and the system in a complementary manner (Huchler, 2022). Tasks are assigned by evaluating who – human or system – is best suited for each task, while ensuring that ultimate decision-making authority always resides with the human.

The implementation of this approach is company-specific and is detailed in the scenarios presented later in the section *Adaption to Specific Businesses*.

SYSTEM REQUIREMENTS

To make a company's knowledge stored in texts accessible, several constraints, especially when considering SMBs, need to be adhered to. The following outlines the key requirements and their corresponding consequences for the design of a retrieval system:

R1: Due to the limited budgets of SMBs the retrieval system must incur only minimal additional costs and investments. Considering the significant computational effort required by sophisticated retrieval techniques or LLMs, the system needs to be designed efficiently, using lightweight methods, so that typical hardware resources provided by an SMB are sufficient and no further

costs are incurred. In particular, this means the entire system must be able to run locally, without the need for additional compute nodes or cloud services.

R2: Taking the sensitive business information a company handles into account, the retrieval system must guarantee full data privacy, avoiding any risk of data leaks. In synergy with system requirement R1, this results in a fully locally runnable system without relying on external services that could compromise data privacy.

R3: With regard to the SMB's employees, the usability of the system and the quality of the results are crucial. Generally, no assumption can be made about the availability of IT experts within SMBs, so both the deployment of the system and its maintenance must be as straightforward as possible. Additionally, the system's user interface needs to be intuitive.

R4: Along with system requirement R3, the system must be highly extensible. Considering the lack of professionals within SMBs, unsupported file formats cannot easily be converted into supported ones. Therefore, the system's code must be well-written, allowing an external developer to easily add support for additional file formats, so the system can be customized to meet the SMB's specific needs.

SYSTEM ARCHITECTURE

The abstract architecture of our proposed retrieval system consists of four main stages, as shown in Figure 1. Below, we describe each stage in detail and relate them to the established system requirements.

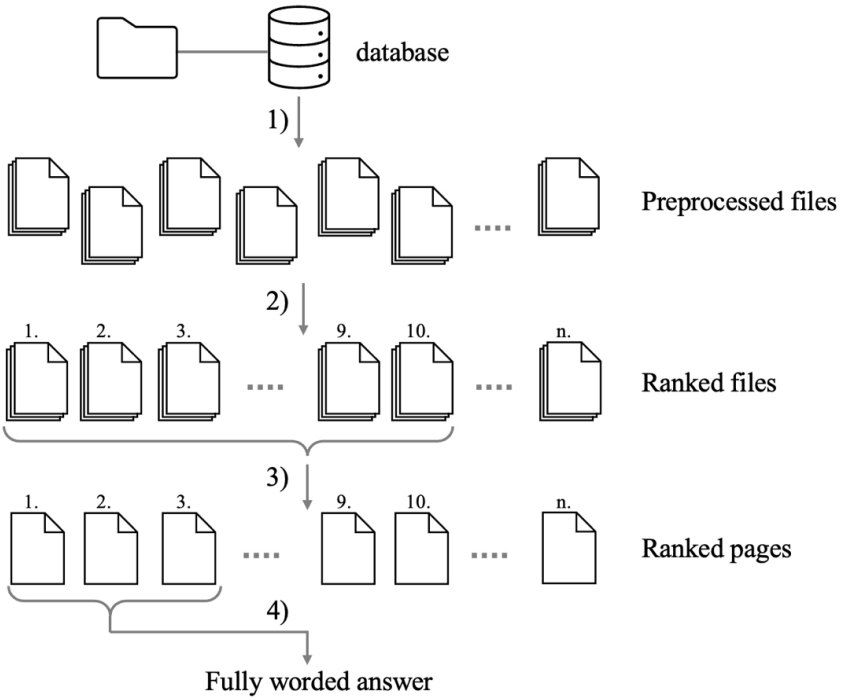

Figure 1: Abstract architecture.

Stage 1 – Preprocessing: To ensure our system functions properly, it assumes that all data – i.e., the company's knowledge – that should be made searchable is placed within a central folder or folder structure, referred to as a *database* in Figure 1. At system startup, the texts in all PDF files within this folder are parsed. Note that we focus on PDF files here because we have determined that this file format is the most common in SMBs for storing textual data and that all tools typically used within SMBs, such as Microsoft Word or Microsoft PowerPoint, offer PDF exports. Moreover, conversions from formats like *.docx* or *.pptx* can easily be automated without manual intervention.

Once all files are parsed, the texts are preprocessed by removing stopwords and applying stemming. Stopwords are words that occur frequently but carry little information, such as *the*, *a*, or *to*. Removing these words improves a retrieval system's accuracy while also enhancing its execution speed (Kaur and Buttar, 2018). Similarly, stemming is a technique that reduces words to their base form. For example, *runner*, *running*, and *runs* are all reduced to the stem *run*, which decreases the number of distinct terms and leads to greater efficiency and accuracy in the retrieval system (Flores et al., 2010).

Stage 2 – Traditional retrieval: Once all documents are parsed, the system is ready to accept user queries and provide answers by leveraging the given database. When a user submits a query, it is processed in the same way as the documents in the database, i.e., stemming and stopword removal are applied. Subsequently, the lightweight BM25 algorithm is utilized to convert both the query and the documents into vectors and compute a score for each document according to its relevance to the query. As a result, a ranked list of documents is produced, indicating the relevance of each document to the user's query.

It should be noted that the creation of document vectors occurs at system startup and only again when the database is updated. This way, only the query vector needs to be computed during retrieval, making the system more efficient.

Stage 3 – Reranking: Given the ranked list of documents from BM25, the first ten documents are split into their individual pages and are processed, along with the user query, by a lightweight cross-encoder, specifically the *svalabs/cross-electra-ms-marco-german-uncased* model. During our experiments, we determined that the first ten documents are very likely to contain the information requested by the user query. Since even the chosen lightweight cross-encoder requires significant computational effort, we argue that processing ten documents is a good trade-off between computational cost and ensuring sufficient information to properly answer a user's query. To achieve a fine-grained result where the necessary information can be directly located, we split the PDF files into pages. This way, the new ranking list consists of the most relevant pages, rather than the most relevant documents (i.e., PDF files), making the searched information accessible without needing to search through the entire retrieved documents.

Stage 4 (optional) – Answer generation: Given the most relevant pages retrieved by the cross-encoder, the fourth stage of our system uses an LLM to format the content of these pages into a fully worded answer. Leveraging the RAG concept, we pass the three most relevant pages as context, along with

the user query, to the LLM to generate a fully worded answer. Furthermore, we prompt the LLM to cite which page was used to produce each generated phrase. Regarding the possibility of LLM hallucination, this approach ensures that the user can easily verify the content of the generated answer.

Considering the system requirements of R1 (minimal additional costs) and R2 (data privacy protection), the fourth stage can be activated depending on the provided hardware resources of the SMB and the data being handled. Unlike the first three stages, an LLM requires substantial hardware resources to be deployed and to respond in a reasonable amount of time. If an SMB provides appropriate resources, the LLM can be run locally. Otherwise, an external LLM, such as GPT-3, needs to be queried using an API or when dealing with sensitive data, the fourth stage can be deactivated.

Frontend: Along with the described backend, we provide an easy-to-use web interface that enables users to submit queries, as shown in Figure 2. Once users have entered their query, it is processed, and both the ten most relevant pages and the fully worded answer (if stage four is activated) are displayed as a result set.

In addition to allowing users to enter queries and view relevant results, the web interface also incorporates a feedback mechanism. This mechanism enables users to assess the relevance of each retrieved page and the fully worded answer, as well as to write comments on the results. Combined with a logging function, this feedback allows for the iterative refinement of the system by analysing user feedback along with logged queries and answers.

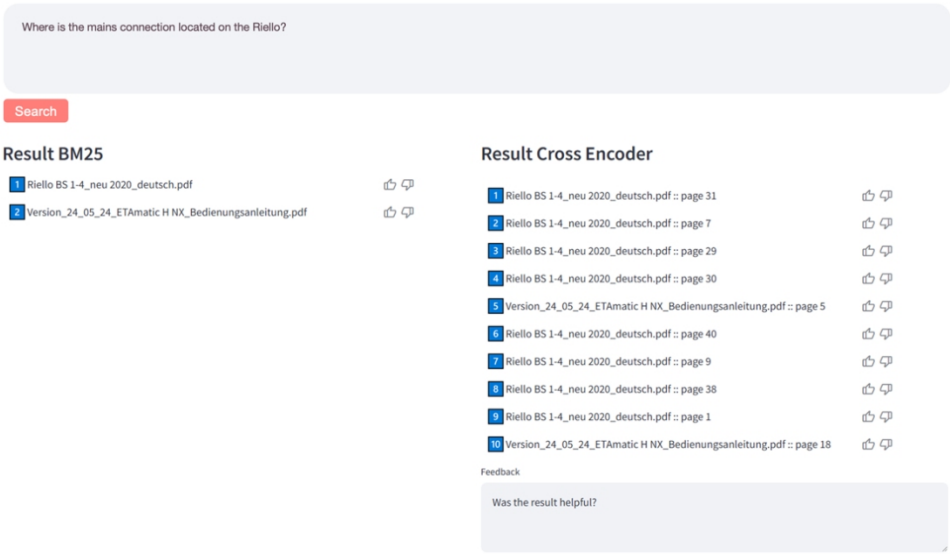

Figure 2: Web-interface.

As the system is designed efficiently and applies only lightweight methods, it can be run locally (R1 and R2) with low hardware requirements (R1). In cases where the company uses special file formats that cannot easily be converted into PDFs, the code to preprocess data is highly extensible,

making the system customizable (R4). To launch the system, only a folder or folder structure needs to be specified, and all preprocessing steps are applied automatically (R3). Once launched, the web interface provides an easy way to interact with and assess the system (R3).

ADAPTION TO SPECIFIC BUSINESSES

Currently, we are applying our described system in several real-world scenarios in cooperation with different SMBs. For each, we followed an adaptation process that implements the concept of human-centered solution development.

Our process always starts with interviews and workshops to gain an understanding of the business, its employees, and its leadership. Based on this input, we identify needs, opportunities, and requirements, from which we derive the steps necessary to adapt our system architecture.

Next, we identify and collect structured and unstructured data sources. For explicit digital data, this is mostly straightforward. We examine existing company software systems, such as ERP systems, and extract information that can be utilized in our retrieval system. While many digital data sources exist, we often encounter challenges regarding data access, which requires the involvement of software vendors or IT departments. Furthermore, we observe the widespread use of supplementary images. Since our system cannot work with them directly, they must be textually described or tagged to be fully utilized.

The most challenging part of this step is the provision of implicit experiential knowledge. In cases where information exists only in people's minds, we employ so-called tandem discussions, where the knowledge of highly experienced, often older employees (one partner of the tandem) is retrospectively queried and documented. The second tandem partner acts as a corrective and ensures that the recorded information is complete and comprehensible. The transcripts of these sessions are then roughly cleaned up and adapted to serve as an additional data source in the retrieval system, making them available to all system users.

A particular challenge when working with transcripts is company- or domain-specific vocabulary. Often, specialized terms are rarely used but highly relevant. In addition, everyday language terms may carry a different meaning in the business context of an SMB. This poses a challenge for traditional retrieval systems as well as language model-based cross-encoders, which may fail to identify documents as relevant when domain-specific query terms are not in general training datasets or have different meanings there. To make domain-specific content searchable, a synonym list must be provided, listing each specialized term along with its corresponding synonyms. With this, user queries can be expanded with appropriate synonyms, allowing the retrieval system to be adapted to the specific domain.

The final preparation and refinement of the transcripts require little additional effort, as supplementary sources – such as technical documentation and manuals – are also included in the retrieval database.

This allows the retrieval system to use all available sources to identify relevant information for a given query.

With this described procedure, we create the initial database for our system. However, to ensure the ongoing expansion of the knowledge base, we also provide concepts for establishing procedures for continuous and digitized documentation. We conclude our adaptation process by configuring our system according to the identified requirements and data sources.

IMPACT ON WORK ACTIVITIES

As mentioned in the last section, we are currently applying our system in cooperation with several SMBs and are still gathering data to evaluate the impact of our system from various perspectives. In particular, we are collecting data through the system's feedback and logging component, which records queries, retrieved pages, generated answers, and the user's feedback, if provided. Using this data, we can, for instance, test semantic correctness by having experts check the logged results for technical accuracy, comprehensibility, or clarity. Furthermore, we use questionnaires at various points in time to measure user acceptance, taking into account individual affinity for technology (Franke et al., 2019).

While our data-based system evaluation is still ongoing, we have observed the introduction and initial application of our approach in different businesses. Based on these observations, we postulate three assumptions about the impact on work activities, which we aim to strengthen in the future using the data being collected.

The *first assumption* is an increase in employee productivity and well-being. Our approach offers context- and situation-adapted research, supporting, for instance, service technicians to work directly at the client's location. This improves the efficiency of order processing and task completion, reduces stress, and supports work-integrated learning processes for individual employees.

The *second assumption* is a need for continuous effort to update and expand the knowledge database. As previously mentioned, numerous employees within a company are required to consistently provide information for the knowledge database. While processing this data technically poses only a minor challenge for the system, it represents an additional workload for the involved employees. Curating the knowledge base–i.e., ensuring the technical accuracy, timeliness, and validity of the stored data–currently requires additional human resources, supported by the aforementioned feedback system. This can only be justified by the creation of the benefits stated in our first assumption.

Our *third assumption* is an increase in technical effort required by the admin to ensure the system runs properly. While setting up the system poses few challenges – the installation is straightforward, and end users can access it via a simple web browser – formatting the data can sometimes be cumbersome, especially when working with image data. A suitable description needs to be provided if one is not already available.

Our assumptions suggest that our system offers benefits but also comes at a cost. In our future work, we aim to evaluate these assumptions using real-world data to further develop and optimize our system approach.

CONCLUSION

Our concept has generated great interest among companies. Together with employees, we are continuously refining the conditions of use based on tests conducted within the companies and gaining extensive experience regarding requirements for document structuring, text clarity, and documentation needs. The next planned step is a study on employees' acceptance of the system.

ACKNOWLEDGMENT

The project is funded by the Federal Ministry of Education and Research under the funding code 02L19C301 Project duration: 01.11.2021 – 31.10.2026.

REFERENCES

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

Bruns, W., Eschenbach, S., Maier, E., 2016. Erfahrung – der unsichtbare Erfolgsfaktor in Wirtschaftsunternehmen: Dokumentation der Ergebnisse einer Befragung von Führungskräften in der Schweiz, Österreich und Deutschland.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., Li, Q., 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Presented at the KDD '24: The 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, Barcelona Spain, pp. 6491–6501. https://doi.org/10.1145/3637528.3671470

Flores, F. N., Moreira, V. P., Heuser, C. A., 2010. Assessing the Impact of Stemming Accuracy on Information Retrieval, in: International Conference on Computational Processing of the Portuguese Language.

Franke, T., Attig, C., Wessel, D., 2019. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. Int. J. Human–Computer Interact. 35, 456–467.

Hinkelmann, K., Witschel, H. F., 2012. Auswahl der richtigen Wissensmanagement-Methoden. Blickpkt. KMU. https://hdl.handle.net/11654/8989

Huchler, N., 2022. Komplementäre arbeitsgestaltung. grundrisse eines konzepts zur humanisierung der arbeit mit KI. Z. Für Arbeitswissenschaft 76, 158–175.

Humeau, S., Shuster, K., Lachaux, M.-A., Weston, J., 2020. Poly-encoders: architectures and pre-training strategies for fast and accurate multi-sentence scoring.

Kaur, J., Buttar, P., 2018. A Systematic Review on Stopword Removal Algorithms 4, 207–210.

Pfaffl, C., Czernich, N., Falck, O., Zimmermann, V., Demary, V., Goecke, H., Schönert, S., Hess, T., Egle, C., Krcmar, H., 2022. Digitale Transformation– wie kann Deutschland zu den führenden Nationen aufschließen? Ifo Schnelld. 75, 03–23.

Richter, C., Draude, C., 2023. Erfahrungswissen in der Pflege – Chancen partizipativer Aktionsforschung und diskriminierungssensibler Technikentwicklung. Gr. Interakt. Organ. Z. Für Angew. Organ. GIO 54, 1–10. https://doi.org/10.1007/s11612-023-00672-x

Robertson, S., Walker, S., 1994. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval, Proceedings of the 17th ACM Conference on Research and Development in Information Retrieval (SIGIR), Dublin, Ireland. https://doi.org/10.1007/978-1-4471-2099-5_24

Robertson, S., Zaragoza, H., Taylor, M., 2004. Simple BM25 extension to multiple weighted fields, in: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. Presented at the CIKM04: Conference on Information and Knowledge Management, ACM, Washington D. C. USA, pp. 42–49. https://doi.org/10.1145/1031171.1031181

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M., 1994. Okapi at TREC-3, in: Harman, D. K. (Ed.), Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994, NIST Special Publication. National Institute of Standards and Technology (NIST), pp. 109–126.

Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Liu, Z., Dou, Z., Wen, J.-R., 2024. Large Language Models for Information Retrieval: A Survey.