

Semantische Dokumentenindexierung mit generativer KI

Dimitri Busch, Dmytro Lande

Zusammenfassung: Ziel dieses Beitrags ist es, neue Verfahren zur semantischen Indexierung von Referenzinformationen mithilfe generativer künstlicher Intelligenz vorzustellen. Ein GPT-Sprachmodell wurde verwendet, um Deskriptoren und Beziehungen zwischen diesen aus Dokumenten zum Thema Architekturgeschichte automatisch zu extrahieren. Aus den extrahierten Deskriptoren und Beziehungen wurde ein semantisches Netzwerk erstellt. Anschließend wurde ein Prototyp entwickelt, der es ermöglicht, mithilfe des semantischen Netzwerkes relevante Dokumente zu finden. Schließlich wird gezeigt, wie sich die Qualität semantischer Netzwerke mithilfe eines Schwarms virtueller Experten verbessern lässt. Der Einsatz generativer künstlicher Intelligenz kann den Arbeitsaufwand und die Kosten der semantischen Indexierung verringern. Die semantische Indexierung und das semantische Netzwerk können ihrerseits zu einer effektiveren Nutzung und Verbreitung wissenschaftlicher Informationen beitragen, indem sie eine semantische Suche, eine einfache Navigation und Benutzerfreundlichkeit ermöglichen.

Schlagwörter: Dokumente, semantische Indexierung, generative künstliche Intelligenz (KI)

Semantic document indexing with generative AI

Abstract: This paper presents new methods for semantic indexing of reference information using generative artificial intelligence. A GPT language model was used to automatically extract descriptors and relationships between them from architectural history documents. A semantic network was created from the extracted descriptors and relationships. A prototype was then developed that made it possible to find relevant documents using the semantic network. Finally, it is shown how the quality of semantic networks can be improved with the help of a swarm of virtual experts. The use of generative artificial intelligence can reduce the workload and costs of semantic indexing. Semantic indexing and the semantic network can contribute to more effective use and dissemination

tion of scientific information by enabling semantic search, easy navigation, and user-friendliness.

Keywords: Documents, semantic indexing, generative artificial intelligence (AI)

DOI: <https://doi.org/10.31263/voebm.v78i1.9251>

Dieses Werk ist – exkl. einzelner Logos und Abbildungen – lizenziert unter einer Creative-Commons-Lizenz Namensnennung 4.0 International-Lizenz

1. Einführung

Aktuell findet eine technologische Revolution statt. Künstliche Intelligenz wird durch große Sprachmodelle (Large-Language-Models [LLM]), z. B. Generative Pretrained Transformer (GPT), allgemein zugänglich. Solche Sprachmodelle ermöglichen die Erstellung von Textinhalten, die dem menschlichen Niveau nahekommen (vgl. Wolfram, 2023).

In diesem Beitrag wird eine Anwendung der großen Sprachmodelle zur semantischen Indexierung von Dokumenten im Bereich Architekturgeschichte dargelegt. Die Dokumente (Datensätze) enthalten Referenzinformationen über Zeitschriftenartikel, z. B. Titel, Kurzreferate (Abstracts) und Deskriptoren (Schlüsselwörter).

Das Ziel dieses Beitrags ist es, neue Verfahren zur Verarbeitung von Referenzinformationen mithilfe generativer Systeme der künstlichen Intelligenz vorzustellen. Die Verfahren werden verwendet, um semantische Indexierung der Dokumente automatisch durchzuführen und ein semantisches Netzwerk zu bilden, das eine Navigation in der Dokumentensammlung ermöglicht.

2. Semantische Indexierung

Die Indexierung ist die Zuordnung von Deskriptoren zu Dokumenten zum Zweck der inhaltlichen Beschreibung und des Wiederauffindens (Retrieval) dieser Dokumente. Bei der semantischen Indexierung werden den Dokumenten nicht nur Deskriptoren, sondern auch Beziehungen zwischen diesen zugeordnet. Im Folgenden ist ein Beispieldokument dargestellt, das semantisch indexiert wurde.

Semantisch indexiertes Beispieldokument

Titel: 1, 2, 3, 6: Early Gothic Architecture and Perfect Numbers

Abstract: ‘Knowledge of numbers should not be despised [...] We are instructed in number to avoid confusion. Take away number in all things, and they all perish. Take away computation from the world, and all things are encompassed by blind ignorance; people who are ignorant of the knowledge of reckoning cannot be distinguished from the other animals’. Thus wrote Isidore of Sevilla in the 7th century (Book III, 4). Number clearly mattered. Indeed, the **medieval world** seems to have taken to heart the words expressed in the **Book of Wisdom** (11.21): ‘Omnia in mensura et numero et pondere fascistic’. In such a number-obsessed world, one would expect number to play a significant role in the design of medieval architecture, for without number, a building, like everything else, would perish. And indeed, many **medieval descriptions** of buildings display a great concern with **numerical values**. After a short introduction into the issues involved, this paper will concentrate on the use of perfect numbers as a design principle in several select examples.

Quelle: den Hartog, E., 2024, Architectural Histories 2(1), <https://doi.org/10.5334/ah.bu>

Beziehungen (Deskriptorenpaare):

- Gothic Architecture; Perfect Numbers
- Medieval World; Book of Wisdom
- Medieval Descriptions; Numerical Values

Dem genannten Dokument wurden Beziehungen zugeordnet, die durch Deskriptorenpaare repräsentiert sind, z. B. „Gothic Architecture; Perfect Numbers“ und „Medieval World; Book of Wisdom“, wobei die Deskriptoren, die miteinander in Beziehung stehen, durch Semikolons getrennt sind.

Die semantische Indexierung von Dokumenten lässt sich automatisch durchführen, indem die Dokumente in Prompts für große Sprachmodelle (LLM) integriert werden (vgl. Lande/Strashnoy, 2023a und Lande/Strashnoy, 2023b). Im Folgenden wird ein solches Verfahren zur semantischen Indexierung mithilfe eines LLM genauer betrachtet.

3. Ein Verfahren zur semantischen Indexierung mithilfe eines LLM

Das Verfahren zur semantischen Indexierung kann mit folgendem Pseudocode dargestellt werden:

Erstelle eine Dokumentensammlung.

Für jedes Dokument

- Erstelle einen Prompt, um Deskriptorenpaare aus dem Dokument zu extrahieren.
- Führe den Prompt aus.
- Schreibe die Deskriptorenpaare in eine CSV-Datei.

Konvertiere die CSV-Datei in ein semantisches Netzwerk (SVG-Datei).

Nach dem genannten Verfahren wird zuerst eine Dokumentensammlung erstellt. Jedes Dokument derselben wird dann semantisch indiziert. Um ein Dokument zu indizieren, d. h. Paare verwandter Deskriptoren zu erzeugen, wird ein Prompt (Eingabeaufforderung) für ein großes Sprachmodell (LLM), z. B. für GPT-3.5, erstellt. Anschließend wird der Prompt ausgeführt, und die erzeugten Deskriptorenpaare werden in eine CSV-Datei geschrieben. Bei CSV handelt es sich um ein Datenformat, bei dem Datensatzelemente durch Semikolons getrennt werden. Nachdem alle Dokumente indiziert sind, wird die CSV-Datei in ein semantisches Netzwerk im SVG-Format umgewandelt. Dieses Netzwerk kann für die Suche und die Navigation in der Dokumentensammlung verwendet werden. Wie erwähnt, muss ein Prompt erstellt werden, um ein Dokument zu indizieren. Im Folgenden wird die Erstellung des Prompts genauer betrachtet.

4. Erstellung eines Prompts

Der Prompt besteht aus einer Anfrage gefolgt von einem Text. Der Text enthält den Titel, das Abstract und gegebenenfalls die Schlüsselwörter, die dem Dokument bereits zugeordnet wurden. Es wird angenommen, dass dem in Kapitel 2 genannten Dokument noch keine Deskriptorenpaare zugewiesen wurden und dass dieses Dokument indiziert werden soll. Vor der Indexierung sieht das Dokument folgendermaßen aus:

Beispieldokument

Titel: 1, 2, 3, 6: Early Gothic Architecture and Perfect Numbers

Abstract: 'Knowledge of numbers should not be despised [...] We are instructed in number to avoid confusion. Take away number in all things, and they all perish. Take away computation from the world, and all things are encompassed by blind ignorance; people who are ignorant of the knowledge of reckoning cannot be distinguished from the other animals'. Thus wrote Isidore of Sevilla in the 7th century (Book III, 4). Number clearly mattered. Indeed, the medieval world seems to have taken to heart the words expressed in the Book of Wisdom (11.21): 'Omnia in mensura et numero et pondere fascistic'. In such a number-obsessed world, one would expect number to play a significant role in the design of medieval architecture, for without number, a building, like everything else, would perish. And indeed, many medieval descriptions of buildings display a great concern with numerical values. After a short introduction into the issues involved, this paper will concentrate on the use of perfect numbers as a design principle in several select examples.

...

Der Prompt für dieses Dokument ist im Folgenden dargestellt:

Prompt: Output 20 pairs of the most related entities from the text. Each entity should be described in no more than 3 words. Format of the output: 'entity 1; entity 2'. Each pair is displayed on a new line. Text:

1, 2, 3, 6: Early Gothic Architecture and Perfect Numbers. Knowledge of numbers should not be despised ...

Der Anfrageteil der Prompts ist unabhängig vom Dokument derselbe und sieht wie folgt aus:

Anfrage: Output 20 pairs of the most related entities from the text. Each entity should be described in no more than 3 words. Format of the output: 'entity 1; entity 2'. Each pair is displayed on a new line. Text:

Nach Eingabe der Eingabeaufforderung extrahiert das große Sprachmodell Deskriptorenpaare aus dem Dokument. Diese werden in einer CSV-Datei gespeichert.

Die Deskriptorenpaare im CSV-Format sind im Folgenden dargestellt:

- Gothic Architecture; Perfect Numbers
- Numbers; Computation
- World; Blind Ignorance
- Knowledge; Reckoning
- Isidore of Sevilla; 7th century
- Medieval World; Book of Wisdom
- Design; Architecture
- Number; Building
- Medieval Descriptions; Numerical Values
- Introduction; Design Principle
- Examples; Select Examples
- ...

Nachdem alle Dokumente indexiert sind, wird die CSV-Datei in ein semantisches Netzwerk im SVG-Format umgewandelt. Für die Datenumwandlung von CSV in SVG kann z. B. die Graphenvisualisierungssoftware CSV2Graph (<https://bigsearch.space/uli.html>) verwendet werden, die auf der Grundlage des GraphViz-Systems (Ferreira, 2017) aufgebaut ist. Um die Datenumwandlung mit CSV2Graph durchzuführen, werden die CSV-Daten in das Eingabefeld des CSV2Graphen eingegeben. Anschließend wird der Button ‚Draw‘ angeklickt (Abbildung 1). Daraufhin erscheint im Browser ein semantisches Netzwerk. In Abbildung 2 ist ein Fragment dieses Netzwerkes dargestellt, das Deskriptorenpaare enthält, die aus dem in Kapitel 4 vorgestellten Dokument extrahiert wurden. Jeder Knoten des semantischen Netzes enthält einen Link zu Google mit einer Anfrage für die Suche nach dem entsprechenden Deskriptor. Jede Kante des semantischen Netzwerkes enthält ebenfalls einen Link zu Google mit der Anfrage, nach den entsprechenden Deskriptorenpaaren zu suchen.

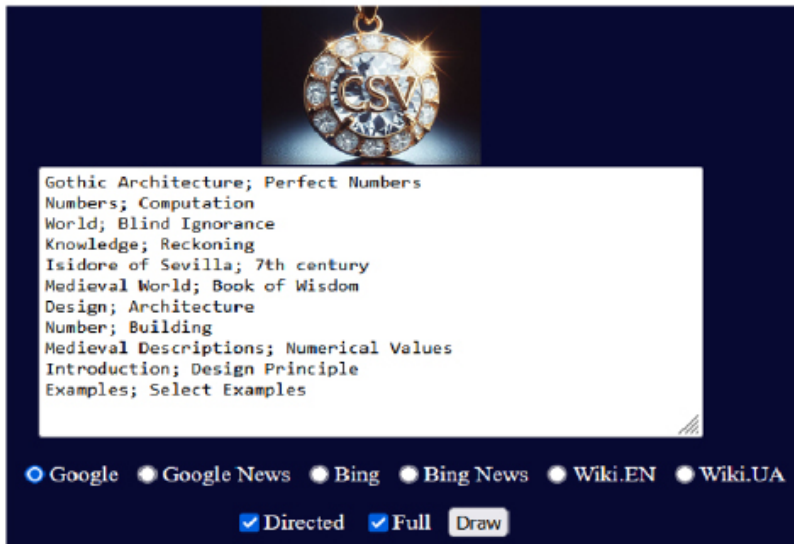


Abbildung 1: Menge der Begriffspaare in CSV2Graph

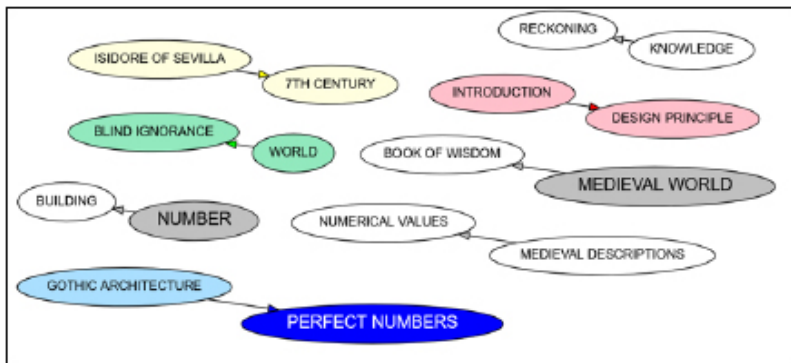


Abbildung 2: Fragment des semantischen Netzwerks

5. Dokumentensammlung

Um das hier beschriebene Indexierungsverfahren umzusetzen, wurde eine Prototyp-Dokumentensammlung mit 39 Dokumenten aus der Onlinezeitschrift *Architectural Histories* gebildet. Dabei handelt es sich um eine Open-Access-Zeitschrift im Bereich Architekturgeschichte

(<https://journal.eahn.org/>). Die Dokumente der Sammlung enthalten nur Referenzinformationen (Titel, Abstracts usw.) und keine Volltexte.

6. Implementierung des Indexierungsverfahrens

Um Deskriptorenpaare aus den Dokumenten zu extrahieren, wurde das große Sprachmodell GPT-3.5 Turbo verwendet, das über FhGenie zugänglich ist. Dies ist ein Chatbot der Fraunhofer-Gesellschaft, der auf Microsoft Azure basiert. Um die Eingabe der Prompts nicht für jedes Dokument manuell durchführen zu müssen, wurde das Sprachmodell von einem Programm in Java aufgerufen. Dieses verwendet die Softwarebibliothek Langchain4j, die eine Schnittstelle zum Sprachmodell darstellt. Die Deskriptorenpaare, die bei der Indexierung extrahiert wurden, wurden in eine CSV-Datei geschrieben und anschließend mithilfe von CSV2Graph in ein semantisches Netzwerk umgewandelt.

Anzumerken ist, dass die Wahl der Modellversion GPT-3.5 Turbo primär aus Kostengründen erfolgte, da der Einsatz eines neueren Modells, wie GPT-4, mit höheren Nutzungskosten verbunden gewesen wäre. Obwohl GPT-3.5 Turbo für die gestellte Aufgabe als angemessen erscheint, ist anzunehmen, dass aktuellere Modelle die Ergebnisse weiter verbessern könnten.

7. Prototyp eines Suchprogramms

Zur Umsetzung eines funktionsfähigen Modells wurde ein Prototypprogramm entwickelt, das nach Anklicken eines Knotens oder einer Kante des semantischen Netzes die entsprechenden Dokumente in der Prototypdatenbank findet und anzeigt. Die Prototypdatenbank enthält momentan 39 Dokumente aus der oben genannten Dokumentensammlung. Die Suche erfolgt nach einer Anfrage, die über den Linkparameter auf einem Knoten oder einer Kante des semantischen Netzes an dieses Programm übermittelt wird. Ein Suchprogramm, das die Apache Lucene-Suchmaschine (<https://lucene.apache.org/>) verwendet, wurde als Java Server Page (JSP, Servlet) unter Apache Tomcat implementiert. Das Servlet ist unter folgender Internetadresse zugänglich: <http://betaindex.de/eahweb>.

Die Prototypdatenbank wurde als Lucene-Index (Sharma, 2020) implementiert. Jeder Eintrag enthält einen Bezeichner, einen Titel,

Schlüsselwörter und Deskriptorenpaare. Die Schlüsselwörter sind Deskriptoren, die nicht vom Sprachmodell, sondern von den Autoren des Dokuments zugeordnet wurden.

Bezeichnung	Typ	Beschreibung
id	Nummer	Kennung
ti	Text	Titel
ab	Text	Abstract
kw	Text, wiederholt	Schlüsselwort
rn	Text, wiederholt	Paar von Deskriptoren

Tabelle 1: Datenbankfelder

Um eine Volltextsuche durchzuführen, wird eine Volltextindexierung der Felder Titel (ti), Abstract (ab), Schlüsselwörter (kw) und Deskriptorenpaare (rn) durchgeführt.

Da der Suchprototyp nicht Google, sondern Lucene als Suchmaschine verwendet und unter <http://betaindex.de/eahweb> zugänglich ist, müssen einige Datenumwandlungen vorgenommen werden, damit der Prototyp ordnungsgemäß funktioniert. Derzeit wurden zwei Arten von Umwandlungen durchgeführt: Umwandlungen in der SVG-Datei und Umwandlungen im Servlet (JSP). In der SVG-Datei wurde der Link zu Google (<https://www.google.com/search>), überall durch <http://betaindex.de/eahweb> ersetzt. Im Servlet wurden Deskriptoren, die ein Deskriptorenpaar bilden, mit dem Operator AND verknüpft. Im Folgenden werden zwei Beispiele der Suche in der Prototypdatenbank dargestellt.

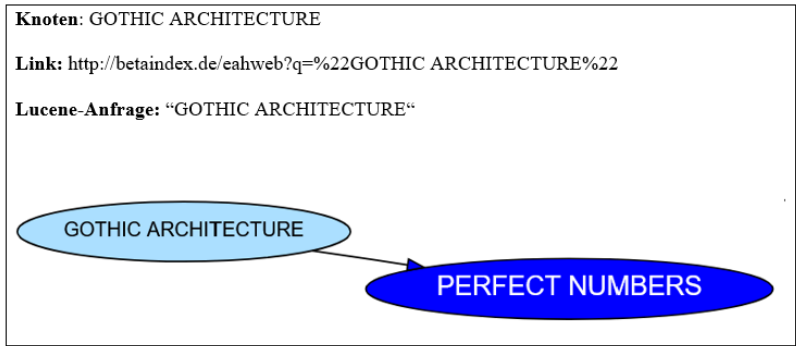


Abbildung 3: Suche nach „GOTHIC ARCHITECTURE“

Im ersten Beispiel sollen Dokumente gefunden werden, die „GOTHIC ARCHITECTURE“ enthalten (Abbildung 3). Dazu muss auf den Knoten „GOTHIC ARCHITECTURE“ des semantischen Netzwerkes geklickt werden. Der Knoten enthält einen Link zum Servlet einschließlich des Anfrage-teils „q=%22GOTHIC ARCHITECTURE%22“. Nach dem Anklicken des Links wird das Servlet angerufen. Dieses wandelt den Anfrageteil des Links in entsprechende Lucene-Anfrage um, führt eine Suche in Lucene durch, und findet entsprechende Dokumente:

ID - 005

TI - 1, 2, 3, 6: Early **Gothic Architecture** and Perfect Numbers

AB - 'Knowledge of numbers should not be despised [...] We are instructed in number to avoid confusion. Take away number in all things, and they all perish. Take away computation from the world, and all things are encompassed by blind ignorance; people who are ignorant of the knowledge of reckoning cannot be distinguished from the other animals?. Thus wrote Isidore of Sevilla in the 7th century (Book III, 4). Number clearly mattered. Indeed, the medieval world seems to have taken to heart the words expressed in the Book of Wisdom (11.21): 'Omnia in mensura et numero et pondere fecisti?. In such a number-obsessed world, one would expect number to play a significant role in the design of medieval architecture, for without number, a building, like everything else, would perish. And indeed, many medieval descriptions of buildings display a great concern with numerical values. After a short introduction into the issues involved, this paper will concentrate on the use of perfect numbers as a design principle in several select examples.

RN - **GOTHIC ARCHITECTURE**; PERFECT NUMBERS

RN - NUMBERS; COMPUTATION

RN - WORLD; BLIND IGNORANCE

...

ID - 021

TI - Reviews Winter 2020

AB - Bergdoll, B. A Review of Miriam Paeslack, Constructing Imperial Berlin: Photography and the Metropolis. Minneapolis: University of Minnesota Press, 2019. Fozi, S. A Review of Robert Bork, **Late Gothic Architecture**: Its Evolution, Extinction, and Reception. Turnhout, Belgium: Brepols, 2018. Higgott, A. A Review of Eva Branscome, Hans Hollein and Postmodernism: Art and Architecture in Austria 1958–1985. London and New York: Routledge, 2018. Vassallo, J. A Review of Martino Stierli, Montage and the Metropolis: Architecture, Modernity, and the Representation of Space. New Haven; London: Yale University Press, 2018; and Craig Buckley, Graphic Assembly: Montage, Media, and Experimental Architecture in the 1960s. Minneapolis: University of Minnesota Press, 2019. Wheeler, K. A Review of Conor Lucey, Building Reputations. Architecture and the Artisan, 1750–1830. Manchester: Manchester University Press, 2018.

RN - BERGDOLL; MIRIAM PAESLACK

RN - BORK; LATE **GOTHIC ARCHITECTURE**

RN - HIGGOTT; EVA BRANSCOME

...

Kante: zwischen „GOTHIC ARCHITECTURE“ und „PERFECT NUMBERS“

Link: <http://betaindex.de/eahweb?q=%22GOTHIC ARCHITECTURE%22+%22PERFECT NUMBERS%22>

Lucene Anfrage: „GOTHIC ARCHITECTURE“ AND „PERFECT NUMBERS“

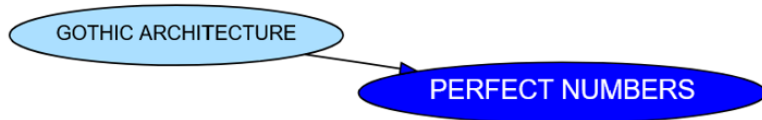


Abbildung 4: Suche nach „GOTHIC ARCHITECTURE“ und „PERFECT NUMBERS“

Im zweiten Beispiel sollen Dokumente gefunden werden, die das Deskriptorenpaar „GOTHIC ARCHITECTURE; PERFECT NUMBERS“ enthalten (Abbildung 4). Hierzu wird auf die Kante geklickt, die „GOTHIC ARCHITECTURE“ und „PERFECT NUMBERS“ verbindet. Die Kante enthält einen Link zum Servlet einschließlich des Anfrageteils „q=%22GOTHIC ARCHITECTURE%22+%22PERFECT NUMBERS%22“. Nach dem Anklicken des Links wird das Servlet angerufen. Dieses wandelt den Anfrageteil des Links in eine entsprechende Lucene-Anfrage um, führt eine Suche in Lucene durch und findet ein entsprechendes Dokument:

ID - 005

TI - 1, 2, 3, 6: Early **Gothic Architecture** and **Perfect Numbers**

AB - 'Knowledge of numbers should not be despised [...] We are instructed in number to avoid confusion. Take away number in all things, and they all perish. Take away computation from the world, and all things are encompassed by blind ignorance; people who are ignorant of the knowledge of reckoning cannot be distinguished from the other animals?. Thus wrote Isidore of Sevilla in the 7th century (Book III, 4). Number clearly mattered. Indeed, the medieval world seems to have taken to heart the words expressed in the Book of Wisdom (11.21): 'Omnia in mensura et numero et pondere fecisti?. In such a number-obsessed world, one would expect number to play a significant role in the design of medieval architecture, for without number, a building, like everything else, would perish. And indeed, many medieval descriptions of buildings display a great concern with numerical values. After a short introduction into the issues involved, this paper will concentrate on the use of **perfect numbers** as a design principle in several select examples.

RN - **GOTHIC ARCHITECTURE; PERFECT NUMBERS**

RN - NUMBERS; COMPUTATION

RN - WORLD; BLIND IGNORANCE

...

8. Schwarm virtueller Experten

Deskriptorenpaare, die mithilfe eines einzigen Prompts extrahiert werden, können Deskriptoren von geringer Bedeutung enthalten oder unscharfe Beziehungen darstellen. Um die Qualität der Deskriptoren und der Beziehungen zu erhöhen, schlagen Lande und Strashnoy ein fortgeschrittenes Indexierungsverfahren vor, das sie als „Swarm of Virtual Experts“ bezeichnen (Lande/Strashnoy, 2024). Bei diesem verbesserten Verfahren wird ein großes Sprachmodell mehrmals mit unterschiedlichen Prompts aufgerufen. Im Folgenden wird der Schwarm virtueller Experten genauer betrachtet. Zuerst werden einige wesentliche Begriffe definiert.

Expertenrolle

Eine Expertenrolle ist die Berufsbezeichnung einer Fachperson mit Expertise in einem bestimmten Fachgebiet. Eine Expertenrolle im Bereich der Architekturgeschichte kann beispielsweise „Architekt:in“ sein. Expertenrollen können entweder von einer menschlichen Fachkraft oder automatisch mithilfe eines großen Sprachmodells ermittelt werden. Um Expertenrollen für die Domäne „Architectural History“ zu ermitteln, kann beispielsweise folgender Prompt verwendet werden:

Prompt: Specify up to 5 the roles of experts involved in research in the area of Architectural History, for example “Architect” and “Archaeologist”. Each role of experts is displayed on a new line. Output only names of the roles in the singular.

Virtueller Experte

Ein virtueller Experte (*virtual expert*) ist ein Prompt, der für eine einzelne Expertenrolle erstellt wird und zum Extrahieren von Deskriptorenpaaren aus einem Dokument dient. Der virtuelle Experte für die Expertenrolle „Architekt:in“ sieht z. B. folgendermaßen aus:

Prompt: Output 20 pairs of the most related entities from the text from the perspective of an **Architect**. Each entity should be described in no more than 3 words. Format of the output: ‘entity 1; entity 2’. Each pair is displayed on a new line. Text: 1, 2, 3, 6: Early Gothic Architecture and Perfect Numbers. Knowledge of numbers should not be despised ...

Das Verfahren zur Indexierung eines Dokuments mithilfe virtueller Experten lässt sich mit folgendem Pseudocode darstellen:

Ermittle Expertenrollen.

Für jede Expertenrolle

- Erstelle einen virtuellen Experten, um Deskriptorenpaare zu extrahieren.
- Führe den virtuellen Experten aus.
- Speichere Deskriptorenpaare, z. B. in ein Array.

Wähle Deskriptorenpaare aus, die mindestens θ -mal vorkommen.

Speichere die ausgewählten Deskriptorenpaare in eine CSV-Datei.

Nach dem dargestellten Verfahren werden zuerst Expertenrollen ermittelt. Anschließend wird für jede Expertenrolle ein virtueller Experte (Prompt) erstellt und ausgeführt. Schließlich werden Deskriptoren ausgewählt, die mindestens θ -mal vorkommen. Der Grenzwert θ kann u. a. nach mathematisch-statistischen Methoden ermittelt werden (vgl. Lande/Strashnoy, 2024).

Das oben genannte Verfahren wurde in Java implementiert und auf das in Kapitel 4 vorgestellte Beispieldokument angewendet. Vor der Indexierung des Dokuments wurden mithilfe von FhGenie folgende Expertenrollen für die Domäne „Architectural History“ gefunden:

1. Architect
2. Building Historian
3. Conservation Specialist
4. Archaeologist
5. Art Historian

Für die Expertenrollen wurden virtuelle Experten automatisch erstellt und ausgeführt. Aus den gefundenen extrahierten Deskriptorenpaaren wurden anschließend nur Paare ausgewählt, die mindestens dreimal vorkommen:

- Number; Building
- Gothic Architecture; Perfect Numbers
- Medieval Descriptions; Numerical Values
- Design principle; Perfect numbers
- Number; Design
- Number; Principle
- Number; Design principle

Aus den Deskriptorenpaaren wurde mithilfe von CSV2Graph ein semantisches Netzwerk erstellt, das in Abbildung 5 dargestellt ist. Wenn der Grenzwert θ auf 1 gesetzt und das Graphenvisualisierungstool Gephi (gephi.org) verwendet wird, lässt sich ein semantisches Netzwerk erzeugen,

das in Abbildung 6 dargestellt ist. Es ist ersichtlich, dass die mithilfe der virtuellen Experten aufgebauten semantische Netzwerke mehr Beziehungen zwischen Deskriptoren enthalten als das Netzwerk aus Abbildung 2, wodurch sie besser strukturiert sind.

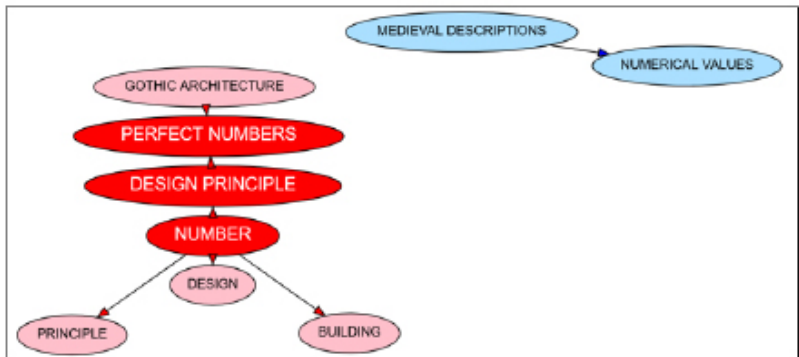


Abbildung 5: Semantisches Netzwerk, von virtuellen Experten ($\theta = 3$) erstellt



Abbildung 6: Semantisches Netzwerk, von virtuellen Experten ($\theta = 1$) erstellt

9. Zusammenfassung

In dieser Arbeit wurde ein großes Sprachmodell (LLM) verwendet, um Dokumente einer Dokumentensammlung im Bereich Architekturgeschichte semantisch zu indexieren. Dazu wurden aus Dokumenten Deskriptorenpaare extrahiert, aus denen anschließend ein semantisches Netzwerk gebildet wurde. Dieses wurde für die Suche von Dokumenten verwendet, die semantisch indexiert wurden.

Es wurde ein funktionsfähiger Prototyp entwickelt, der es ermöglicht, durch Anklicken von Knoten und Kanten des semantischen Netzwerks relevante Dokumente zu finden und in der Dokumentensammlung zu navigieren. Darüber hinaus wurde ein verbessertes Indexierungsverfahren mit virtuellen Experten dargestellt.

Der Einsatz generativer Technologien der künstlichen Intelligenz kann den Arbeitsaufwand und die Kosten der semantischen Indexierung verringern. Die semantische Indexierung und das semantische Netzwerk können ihrerseits zu einer effektiveren Nutzung und Verbreitung wissenschaftlicher Informationen beitragen, indem sie eine semantische Suche, eine einfache Navigation und Benutzerfreundlichkeit ermöglichen.

Dr. Dimitri Busch

ORCID iD: <https://orcid.org/0009-0008-6875-9853>

Fraunhofer-Informationszentrum Raum und Bau (IRB),
Stuttgart, Deutschland

dimitri.busch@irb.fraunhofer.de

Prof. Dr. Dmytro Lande

ORCID iD: <https://orcid.org/0000-0003-3945-1178>

Nationale Technische Universität „Kiewer Polytechnisches Institut
Ihor Sikorskyj“,

Kiew, Ukraine

dwlande@gmail.com

Literatur

Ferreira, Diogo R. (2017). *A Primer on Process Mining: Practical Skills with Python and Graphviz*. Cham: Springer.

Lande, Dmytro; Strashnoy, Leonard (2024). *Swarm of Virtual Experts in the Implementation of semantic Networking*. ResearchGate Preprint, <https://doi.org/10.13140/RG.2.2.16686.11845>.

Lande, Dmytro; Strashnoy, Leonard (2023a). *Concept Networking Methods Based on ChatGPT & Gephi*. SSRN Preprint, 28 April, <https://doi.org/10.2139/ssrn.4420452>.

Lande, Dmytro; Strashnoy, Leonard (2023b). *GPT Semantic Networking: A Dream of the Semantic Web – The Time is Now*. Kyiv: Engineering.

Sharma, Atri (2020). *Practical Apache Lucene 8: Uncover the Search Capabilities of Your Application*. New York, NY: Apress.

Wolfram, Stephen (2023). *What Is ChatGPT Doing ... and Why Does It Work?* Champaign, IL: Wolfram Media.