Dimitri Busch[1], PhD,
Dmitry Lande[2], Doctor of Technical Sciences,
Anatolii Feher[2],
Leonard Strashnoy[3]

[1]Fraunhofer Information Center for Planning and Building, Stuttgart
[2]National Technical University of Ukraine - Igor Sikorsky Kyiv Polytechnic Institute
[3] University of California, Los Angeles (UCLA)

Semantic document indexing with Generative AI

Abstract

The possibilities of processing abstract information using generative artificial intelligence systems, in particular ChatGPT, are presented to solve the problems of generating semantic maps, semantic indexing, analysis and visualization, which makes it possible to consider such systems as a useful analytical tools. The ChatGPT system was used to automatically extract basic concepts from documents from a thematic array on the topic of information technology in construction, i.e. perform semantic indexing, as well as build a semantic network for the selected subject area. A working model has been implemented that allows you to find relevant records by clicking on nodes and edges of the constructed semantic network, i.e. navigate through the source information array.

Keywords: Semantic indexing, Semantic network, Artificial General Intelligence, Reference database, BIM, IT

Introduction

Today, another technological revolution is taking place - artificial intelligence is becoming publicly available thanks to systems such as GPT (Generative Pre-trained Transformer), which can generate text content approaching the human level [1]. The use of generative artificial intelligence systems (AGI - Artificial General Intelligence) allows us to take a fresh look at indexing traditional sets of scientific and technical information, since systems of this level have access to procedures for extracting keywords, called entities, etc. from documents, as well as establishing meaningful relations between them [2]. Accordingly, new possibilities for semantic search are opening up, that is, searching not for individual words contained in documents, but for navigating the main essential elements of these documents. Most often, when using AGI systems, entities can be extracted directly from the trained model. But when

abstract information is examined for its semantic indexing, it makes sense to insert the abstracts themselves into queries in AGI systems (prompts) [3].

The purpose of this work is to present new possibilities for processing abstract information using generative artificial intelligence systems, in particular ChatGPT, to solve problems of semantic indexing, generating semantic maps, their analysis and visualization, which allows us to consider such systems as a new level of analytical tool.

In this work, document contents are embedded in ChatGPT prompts in order to identify pairs of related keywords from these documents. According to this approach, prompts are transmitted to the ChatGPT system, which selects from individual documents pairs of concepts that describe their semantic content. These concept pairs are then passed to a graph renderer such as CSV2Graph, which generates a semantic network from them. In the future, it will be shown how this approach can be used for semantic indexing of documents on the topic of information technology in construction, the formation of a semantic network of this subject area, and how the created semantic network can be used for information retrieval and navigation in an abstract database.

Semantic indexing

Let us give an example of the formation of a semantic network based on metadata on the topic of Building Information Modeling, which is a partial topic of the main topic of information technology in construction. Recordings of ArXiV electronic preprints (arxiv.org) were used for the experiment. Below is a fragment of a separate document on the presented topic:

A Geometric-Relational Deep Learning Framework for BIM Object Classification

Interoperability issue is a significant problem in Building Information Modeling (BIM). Object type, as a kind of critical semantic information needed in multiple BIM applications like scan-to-BIM and code compliance checking, also suffers when exchanging BIM data or creating models using software of other domains. It can be supplemented using deep learning. Current deep learning methods mainly learn from the shape information of BIM objects for classification, leaving relational information inherent in the BIM context unused. To address this issue, we introduce a two-branch geometric-relational deep learning framework. It boosts previous geometric classification methods with relational information. We also present a BIM object dataset IFCNet++, which contains both geometric and relational information about the objects. Experiments show that our framework can be flexibly adapted to different geometric methods. And relational features do act as a bonus to general geometric learning methods, obviously improving their classification performance, thus reducing the manual labor of checking models and improving the practical value of enriched BIM models.

Hairong_Luo; Hairong Luo

2022

Computer Vision and Pattern Recognition

http://arxiv.org/abs/2212.00942

As an example, to index individual documents of this information array, the same type of prompts were applied to the AGI system, containing document texts of the following type:

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: A Geometric-Relational Deep Learning Framework for BIM Object Classification…

As you can see, the prompt itself consists of a meaningful request followed by text. At the same time, the query part for various documents is the same everywhere:

„*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:* “

The document text that is included in the prompt contains a title, abstract, and keywords for each entry.

As a result of executing the above prompt, AGI produces for each document pairs of corresponding entities, which become the basis for semantic indexing, for example:

BIM; Object
Interoperability; Problem
BIM; Applications
Scan-to-BIM; Checking
BIM; Date
Models; Software
Deep learning; Classification
Shape information; BIM objects
Relational information; BIM context
Geometric-relational; Learning framework
Geometric methods; Classification performance
BIM object; Dataset
Geometric information; Objects
Relational features; Bonus
General geometric; Learning methods
Classification performance; Manual labor
Checking models; Practical value
Enriched BIM; Models

Computer Vision; Recognition
Pattern Recognition; BIM

By attributing selected entities to the corresponding documents contained in the thematic information array, the procedure for semantic indexing of these documents is actually implemented. In this case, each document is assigned not even individual entities - keywords, but a small semantic network, a set of entities connected by edges, the role of which is played by a special symbol. ";".

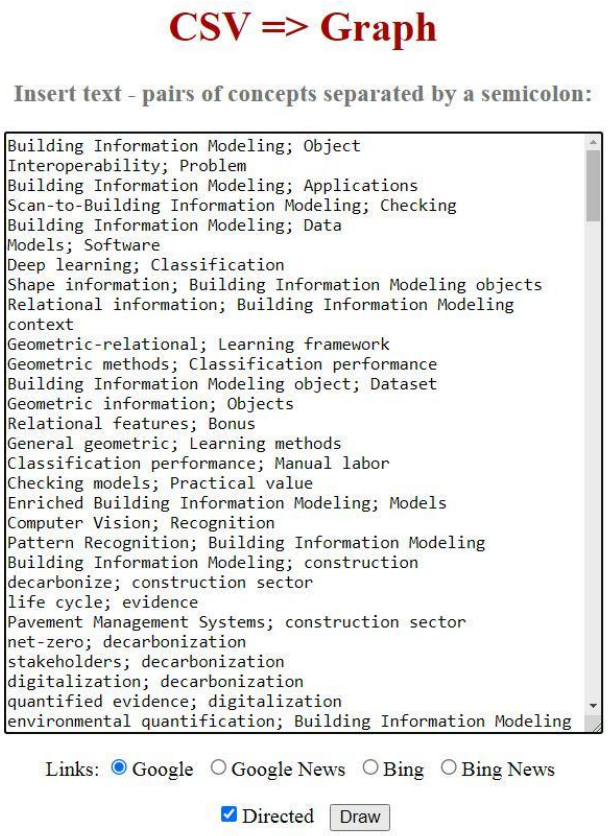
Thus, pairs of meaningfully related concepts were selected from the array of thematic data using the ChatGPT system. In order not to do this manually each time, ChatGPT was called and responses were processed using a program in Java, for which the API capabilities provided by this system were used. As a result, several thousand pairs of concepts were obtained that can serve as the basis for information retrieval.

For further construction of the semantic network, a set of 570 most frequent pairs of concepts was selected, a fragment of which is given below:

Building Information Modeling; Object
Interoperability; Problem
Building Information Modeling; Applications
Scan-to-Building Information Modeling; Checking
Building Information Modeling; Data
Models; Software
Deep learning; Classification
Shape information; Building Information Modeling objects
Relational information; Building Information Modeling context
Geometric-relational; Learning framework
Geometric methods; Classification performance
Building Information Modeling object; Dataset
Geometric information; Objects
Relational features; Bonus
General geometric; Learning methods
Classification performance; Manual labor
Checking models; Practical value
Enriched Building Information Modeling; Models
Computer Vision; Recognition
Pattern Recognition; Building Information Modeling
Building Information Modeling; construction…

It should be noted here that when creating the above set of the most frequent pairs of concepts, the abbreviation BIM was replaced by the full name of the concept, Building Information Modeling.

Then the resulting pairs of concepts were placed in the input field of the graph visualization software CSV2Graph (https://bigsearch.space/uli.html), built on the basis of the GraphViz system [4] (Fig. 1).



Picture 1: Set of concept pairs in CSV2Graph

After filling the input field with data (corresponding to the CSV format), selecting additional options and clicking the Draw CSV2Graph button, a semantic network is formed, one of the fragments of which is shown in Figure 2. The semantic network node contains a link to the Google system with a request to search for the corresponding concept. The edge of the semantic network also contains a link to Google with a request to search for the corresponding pair of concepts.

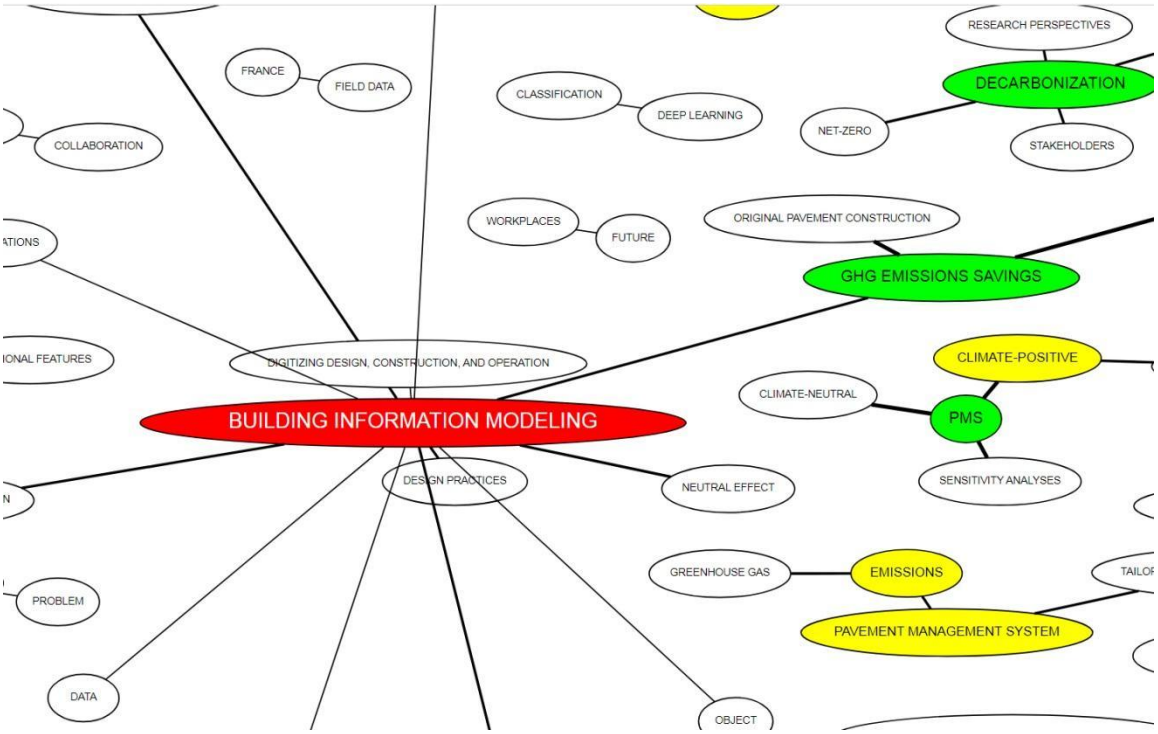


Figure 2: Fragment of the semantic web

Prototype search program

To implement the current model, a prototype program was developed that, after clicking on a node or edge of the semantic network, finds and displays the corresponding records from the prototype database. The search is carried out according to a query that is transmitted to this program through the link parameter on a node or edge of the semantic network. A search program that uses the Apache Lucene search engine (https://lucene.apache.org/), was implemented as a JSP (Java Server Page, servlet) running Apache Tomcat. The prototype database was implemented as a Lucene index [5]. Each entry contains an identifier, title, keywords, and concept pairs (Table 1).

Table 1: Database fields

Designation	Type	Description
id	number	identifier
of	text	title
ab	text	abstract
kw	text, repeated	keyword
rn	text, repeated	a couple of concepts

In order to perform a full-text search, full-text indexing of the title (ti), abstract (ab), keywords (kw) and concept pairs (rn) fields is performed.

Database search queries are written in the Lucene query language. When you click on a semantic network node, you need to find records containing a certain concept. The query corresponding to a semantic network node therefore contains only this concept. Since a concept can consist of several words, it is enclosed in quotation marks in the query.

Example:

“ARCHITECTURE“

“MACHINE LEARNING“

To navigate the database, a constructed semantic network is used, each node of which corresponds to a separate concept (entity), and an edge to a pair of concepts that are connected in a query using the conjunction operator (AND), for example, an edge connecting the concepts “MACHINE LEARNING“ AND “ ARCHITECTURE" is matched with the following query:

“MACHINE LEARNING“ AND “ARCHITECTURE“.

When generating links in an SVG file in the CSV2Graph system, you can select search engines (by default, links to the Google system are generated), for example, for the above request, a link is generated on the nodes:

https://www.google.com/search?q=MACHINE%20LEARNING

https://www.google.com/search?q= ARCHITECTURE

An edge reference would include both concepts:

https://www.google.com/search?q=%22MACHINE%20LEARNING%22+%22ARCHITECTURE%22

The prototype system uses Lucene (https://lucene.apache.org/) like a search engine. Therefore, the original queries must be converted to Lucene queries. There are currently two types of transformations performed: transformation in the SVG file and transformation in the servlet (JSP). In the SVG file, the link to Google, “https://www.google.com/search” is replaced everywhere by http://betaindex.de/chatweb. In a servlet, all concepts are enclosed in upper quotes, and pairs of concepts are joined using the AND operator. Thus, by clicking on the edge between the concepts “MACHINE LEARNING” and “ARCHITECTURE”, a query is generated to the Lucene search engine “MACHINE LEARNING” AND “ARCHITECTURE”, after which records like:

ID - 014
TI - BIM Hyperreality: Data Synthesis Using BIM and Hyperrealistic Rendering for Deep Learning
AB - Deep learning is expected to offer new opportunities and a new paradigm for the field of architecture. One such opportunity is teaching neural networks to visually understand architectural elements from the built environment. However, the availability of large training datasets is one of the biggest limitations of neural networks. …

KW - Machine Learning
RN - BUILDING INFORMATION MODELING; HYPERREALITY
RN - DATA; SYNTHESIS
RN - BUILDING INFORMATION MODELING; HYPERREALISTIC RENDERING
RN - DEEP LEARNING; ARCHITECTUREAND
RN - NEURAL NETWORKS; VISUAL UNDERSTANDING
RN - TRAINING DATASETS; LIMITATIONS
RN - NEURAL NETWORKS; HUMAN ANNOTATIONS
RN - HYBRID SYSTEM; BUILDING INFORMATION MODELING AND RENDERING
RN - TRAINING DATASET; BUILDING INFORMATION MODELINGRAI
RN - BUILDING INFORMATION MODELING MODEL; RENDERED MODEL
RN - RENDERINGS; DEEP LEARNING MODEL
RN - GENERATIVE ADVERSARIAL NETWORK; GAN MODEL
RN - OUTPUT MODEL; REAL-WORLD PHOTOS
RN - NEURAL NETWORK; SYNTHETIC DATA
RN - PHOTOREALISTIC RENDERINGS; BUILDING INFORMATION MODELING-BASED LABELS
RN - BUILDING OBJECTS; PHOTOS
RN - TRAINING DATA; PHOTOS
RN - FUTURE WORK; BUILDING INFORMATION MODELING MODELS
RN - GENERALIZED MAPPING; PHOTOGRAPHED ENVIRONMENTS
RN - MACHINE LEARNING; ARCHITECTURE

ID - 022
TI - Building Information Modeling and Classification by Visual Learning At A City Scale
AB - In this paper, we provide two case studies to demonstrate how artificial intelligence can empower civil engineering. In the first case, a machine learning-assisted framework, BRAILS, is proposed for city-scale building information modeling. Building information modeling (BIM) is an efficient way of describing buildings, which is essential to architecture, engineering, and construction. Our proposed framework employs deep learning technique to extract visual information of buildings from satellite/street view images. Further, a novel machine learning (ML)-based statistical tool, SURF, is proposed to discover the spatial patterns in building metadata.
KW - Computer Vision and Pattern Recognition
KW - Machine Learning
RN - BUILDING INFORMATION MODELING; CLASSIFICATION
RN - VISUAL LEARNING; ARTIFICIAL INTELLIGENCE
RN - CIVIL ENGINEERING; CASE STUDIES
RN - MACHINE LEARNING; BRAILS
RN - CITY-SCALE; BUILDING INFORMATION MODELING
RN - BUILDING INFORMATION MODELING; ARCHITECTURE
RN - ENGINEERING; CONSTRUCTION
RN - DEEP LEARNING; VISUAL INFORMATION
RN - SATELLITE/STREET VIEW; IMAGES
RN - MACHINE LEARNING; STATISTICAL TOOL
RN - ML; SURF
RN - SPATIAL PATTERNS; BUILDING METADATA

Conclusion

In this work, the ChatGPT system was used to automatically extract basic concepts from documents from an array on the topic of information technology in construction, i.e. perform semantic indexing, as well as build a semantic network for the selected subject area. A working model has been implemented that allows you to find relevant records by clicking on nodes and edges of the constructed semantic network, i.e. navigate through the source information array.

Semantic indexing, the formation of a network of concepts based on AGI technology and the construction of a semantic map can facilitate the use and dissemination of scientific and technical information through the implementation of semantic search, convenient navigation, and user understandability.

Literature

1. Stephen Wolfram. What Is ChatGPT Doing ... and Why Does It Work? Wolfram Media, Inc., March 9, 2023. – 112 p. ISBN-13: 978-1-57955-081-3

2. Lande, Dmitry and Strashnoy, Leonard. Concept Networking Methods Based on ChatGPT & Gephi (April 17, 2023). Available at SSRN: https://ssrn.com/abstract=4420452 or http://dx.doi.org/10.2139/ssrn.4420452

3. Dmytro Lande, Leonard Strashnoy. GPT Semantic Networking: A Dream of the Semantic Web – The Time is Now. – Kyiv: Engineering, 2023. – 168 p. ISBN 978-966-2344-94-3

4. Diogo R. Ferreira. A Primer on Process Mining: Practical Skills with Python and Graphviz. Springer Briefs in Information Systems. – Springer International Publishing. – 2017. 101 p. ISBN 978-3-319-56426-5, 978-3-319-56427-2

5. Atri Sharma. Practical Apache Lucene 8: Uncover the Search Capabilities of Your Application. – Apress. – 2020. – 114 p. ISBN 9781484263440.