Generative AI for Information Retrieval and Taxonomies

Photo: Summit Art Creations, Shutterstock.com

Heather Hedden

It's been established that Generative AI (GenAI) cannot create acceptable back-of-the-book indexes, but there are other tools for information findability and retrieval that indexers and information professionals may create and which are subject to AI generation. These include periodical/database indexes, thesauri, and taxonomies. Creating these other information tools and systems, especially taxonomies, has greater potential than book indexes for involving AI.

AI in Periodical/Database Indexing

Periodical/database indexing differs significantly from back-of-the-book indexing in several ways, which makes it more suitable for applying traditional AI methods, if not GenAI. Periodical/database indexing is sometimes referred to as "open indexing," since it involves articles or other documents that continuously get added and thus is ongoing, and the index constantly evolves over time. This contrasts with book indexing, which in this context is called "closed indexing," since the index is created for one book and then is done.

Periodical/database indexing uses an existing thesaurus or taxonomy of terms, rather than creating terms (main entries and subentries) at the time of indexing. Thus, indexing is reduced to merely applying terms and does not include creating terms or cross-references between terms (although indexers may suggest additions). So, this kind of indexing is a more straightforward task that AI can more easily manage.

Additionally, instead of page locators and page flow, with the issue of an indexable topic spanning several pages even if it is not explicitly named on each page, in periodical/database indexing, the entire document is the target of an index entry. Thus, it is easier for AI to identify the suitability of indexing a document rather than trying to identify page ranges.

In periodical/database indexing, there is also no recommended limit to the number of document records per index term, in contrast to five to six page locators in a book index. Most often, there aren't even any subentries in periodical/database indexing these days, which greatly simplifies the task for AI indexing.

Traditional (nongenerative) AI methods have been used in article/database indexing (not exclusively periodicals) for over thirty years, and these methods are commonly used by both publishers and companies/organizations with large volumes of internally created documents. The technologies involved (often in combination) include statistical-based machine learning, rules-based automated indexing, natural language processing (NLP), and named entity recognition. Image recognition technologies have advanced more recently and are now often used for the automated indexing of images.

The need for professional periodical indexers has significantly declined as a consequence of the widespread adoption of auto-indexing, although manual tagging continues to be done in some cases, especially when it's just a small task as part of a larger editorial or content management role.

The relative success of these traditional AI methods means that the newer GenAI technologies have not been needed, and thus, for the most part, have not been adopted for basic auto-indexing in periodical/database indexing.

AI in Thesaurus Creation

Periodical/database indexing relies on the use of a preexisting controlled vocabulary, often a thesaurus, of prescribed terms for indexing. This ensures consistent indexing over time on articles written by different authors and indexed by different indexers. Thesauri have been created for specific subject areas or disciplines, and published thesauri are available for reuse.

The main features of a thesaurus are hierarchical (Broader/Narrower) relationships and associative (Related) relationships between terms, and equivalence (Use/Used for) relationships that redirect from an entry/nonpreferred term to a preferred term. Related-term relationships function the same as a *See also* cross-reference, and the equivalence relationship (Use/Used for) is just like a *See* cross-reference.

GenAI is not so suitable for generating thesauri for several reasons. There are fewer published examples of thesauri for large language models (LLMs) to get trained on, in contrast with taxonomies, which exist on many public websites, so it's difficult to get good results. Furthermore, GenAI is more likely to create something that resembles a taxonomy more than a thesaurus, with emphasis on hierarchy.

In particular for thesauri, the use of the associative (Related Term) relationship is very subjective. It is like the *See also* reference, which indexing tests of GenAI have shown to be especially inaccurate or inappropriate. GenAI does a reasonable job of suggesting Use (*See*) relationships (cross-references), though, so thesaurus creators might use GenAI for obtaining suggestions for nonpreferred terms (Used from) relations.

AI in Taxonomy Creation

Taxonomies are similar to thesauri but tend to be simpler and are often smaller. As such, it is possible to at least create a partially acceptable taxonomy with the aid of GenAI.

Unlike thesauri, most taxonomies lack the associative (Related Term) relationships between terms, which is too difficult for AI to generate. The focus of a taxonomy is, instead, its hierarchical relationships, which GenAI is reasonably good at. Alternative entry terms (for Use/*See* cross-references), while not required, are a common feature of taxonomies, and these can also be AI generated.

Unlike thesauri, taxonomies are less often created by subject matter experts, but rather by taxonomist-generalists, marketers, and information technology professionals. People in these roles are more likely to turn to GenAI tools to help with taxonomy creation. There are also numerous examples of taxonomies, unlike thesauri, on websites, which GenAI can use for training examples.

Taxonomies are more common in business and industry applications for information retrieval. Inclusion of taxonomies (with a limited number of hierarchies and no associative relationships) is supported in various information systems: web content management systems, enterprise content management systems, document management systems, digital asset management systems, etc. Not just the creators of taxonomies but also the users of taxonomies are more likely to embrace AI technologies.

Using GenAI to help create taxonomies is now becoming mainstream, as enterprise taxonomy management software is starting to incorporate GenAI features, in combination with an LLM subscription, to assist users with the creation of taxonomies. These include Graph Modeling from Graphwise, Synaptica Graphite from Squirro, and Semaphore from Progress. The software vendors which incorporate LLMs into their products to assist with taxonomy generation require the user to approve GenAI suggestions.

Nevertheless, taxonomies are more complicated and nuanced than they might seem. Taxonomies should be customized for the content and data for which they will be used, the end users' needs and expectations, the use cases or purposes they will serve, and the front-end application requirements. A GenAI tool would need very detailed and lengthy prompt instructions just to get started and then require a lot of refinement. Using GenAI to create an entire taxonomy results in a taxonomy that is not sufficiently customized, taxonomy terms and relationships that are even incorrect for the specific context, and taxonomy terms and relationships that are inconsistent with each other since they are pulled from different sources. It may even happen that a GenAI prompt to create a taxonomy will extract from copyrighted taxonomies.

Taxonomists have found that it is much more practical and successful to use GenAI for incremental *parts* of the taxonomy creation process, rather than an entire taxonomy at once. Unlike creating an index, it is entirely feasible to create a taxonomy in steps or to create parts of a taxonomy separately.

Taxonomy Creation Tasks Suitable for GenAI

The best approach is to use GenAI for a taxonomy selectively. Certain branches of a taxonomy may be more suitable for AI generation than others. A taxonomy is structured into distinct hierarchies, which are often based on types of terms or aspects, also called facets, which may serve as refinement filters. For example, GenAI could be used to suggest taxonomy hierarchies for industries, geographic locations, technologies, regulations, and trending topics, but not for an organization's own products, services, departments, office locations, or customers.

GenAI is also suitable for various subtasks of taxonomy creation. These include suggesting narrower terms for a concept, converting a flat list of terms into a suggested hierarchy, suggesting alternative labels (synonyms or **Used from** terms) for a concept, generating definitions for concepts, or explaining the relationship of two technical terms to each other (broader/narrower inclusive, related and overlapping, or synonymous). GenAI thus can assist a taxonomist working in an unfamiliar subject area. In building hierarchies, GenAI can also assist a subject matter expert who is not familiar with taxonomy standards or best practices.

Using GenAI for subtasks has other benefits. By working on specific taxonomy tasks, it's easier to write specific prompts for GenAI applications. Working with smaller portions of the taxonomy, the taxonomy editor can more easily review GenAI results, treating them as suggestions, and accept or reject individual suggestions. It's not so much a problem of incorrectly generated results, but rather of a generated taxonomy that is not suitable or appropriate for the particular context. A GenAI tool may suggest a greater number of narrower terms or alternative labels than needed, but then one can select from among the AI suggestions.

GenAI can generate a full multilevel hierarchical taxonomy, but it's recommended that this be done only for suggestions of a "starter" taxonomy or as a source for ideas and inspiration without adopting most of it. In any case, the specialized purpose of a custom taxonomy should always involve "human-in-the-loop" interaction, instructions, review, and editing.

Taxonomies for Supporting GenAI

Not only can GenAI support taxonomy creation, but taxonomies can also support the use of GenAI in information retrieval. Although GenAI usually does not utilize taxonomies, similar to web search engines, GenAI using LLMs that are combined with taxonomies in an internal enterprise implementation yields better results.

GenAI and LLMs are used instead of simple search engines in information retrieval that involves more complex search queries and those that involve accessing data and not just full documents, such as for corporate knowledge management and business intelligence purposes. When querying internal enterprise information or data, a higher level of accuracy is also expected and needed. LLMs can reference an enterprise knowledge base when Retrieval Augmented Generation (RAG) is implemented. A RAG system retrieves relevant facts from an "external" knowledge base (namely an enterprise's own data and content) before generating an answer. This reduces hallucinations, contradictions, and inconsistencies.

A taxonomy can enhance the results of RAG, since even internal information may have inconsistencies, such as the different use of terminology across different departments. Synonyms need to be made explicit. Relationships between concepts (broader, narrower, and related) in the taxonomy can extend the retrieval. The hierarchy feature of a taxonomy also serves LLMs by providing context and thus more specific meaning for concepts through their hierarchical relationships. Furthermore, there may be terminology used uniquely to the enterprise, which an LLM wouldn't know but added definitions in the taxonomy can support.

Data-heavy organizations that implement LLMs internally combined with a custom taxonomy at the enterprise level usually opt to go a step further and implement what is called GraphRAG. GraphRAG combines LLMs with a knowledge graph, which comprises a taxonomy, ontology, and linked instance data in a graph database. This way, the LLMs can make use of the explicit semantic relationships in the knowledge graph, which support complex, multicomponent queries. Because of their support for GenAI and LLMs using RAG or GraphRAG, taxonomies may be growing in popularity.

Conclusions

Using GenAI has altered how many people look for information. However, when information seekers want to explore, browse, and refine topics under their own control, they will still use indexes and displayed taxonomies. GenAI cannot generate adequate indexes or thesauri, although traditional AI methods are often used for auto-indexing documents or other database content using a thesaurus or taxonomy. GenAI can, however, be used effectively to assist in tasks for creating a taxonomy (not a thesaurus) due to the simpler nature of taxonomies and the flexible workflow in creating them. Human involvement is still needed, though, to select and refine the GenAI suggestions. In the other direction, taxonomies are also starting to be used by enterprises to support the use of GenAI as a part of Retrieval Augmented Generation (RAG) implementations. ■

Heather Hedden is an information management professional who specializes in taxonomies, ontologies, metadata, and indexing. She started her business Hedden Information Management in 2004 and has alternated between being self-employed and employed. She has offered various services over the years including back-of-book indexing, website A–Z index creation and training, taxonomy and controlled vocabulary design and creation, and taxonomy training. She is currently providing independent consulting in taxonomies/ontologies full-time.