How search works: advanced

Last Modified on 02/04/2025 1:32 pm EST

Ever wanted to understand how the various pieces of search work together to return results? This section's for you.

Search basics

The goal of search is to get readers to relevant content. There are many things happening behind the scenes to help make this happen. As much as possible, our goal is to provide a search that works really well out-of-the-box while also giving you visibility and control over your search results.

The heart of any search engine is the index, which is where (and how) your content is stored for retrieval. KnowledgeOwl automatically indexes the most important fields of your content for search:

- Title
- Permalink
- Body
- PDFs
- Meta description
- Search phrases

You do not need to add tags or keywords for search to work. Any words you use in the indexable fields will automatically be searchable.

Here's a quick overview of what is and is not indexed for search:

	Indexed / searchable	Optional indexed / searchable	Not indexed / not searchable
Types of pages	Articles Custom content categories Topic display categories	Glossary (*if appropriate settings are selected)	Default categories Blog style categories URL redirect categories
Fields	Title Body Permalink Meta description Search phrases PDFs	Glossary terms The word "Glossary"	Tags (searchable only using tags search) Category / breadcrumbs Glossary term definitions

You can control your search scores and results through a few different methods. See Optimize search for a detailed walkthrough of the strategies we suggest!

Indexing

The knowledge base index is where we store your article content and meta data after we process and optimize it for searching.

When you save an article, we automatically digest and analyze your content for the index. This process involves:

- Stripping out HTML
- Breaking up the text along word boundaries (tokenization)
- Breaking up tokens into small fragments, called N-grams, for partial matches and autosuggestions
- Removing punctuation
- Lowercasing all tokens
- Stemming all tokens based on your chosen primary search language
- Adding the synonyms from your library for matching tokens

Reindexing

The initial index is automatically created when you save an article. If you make a change to your search settings that affects how the data is processed, your knowledge base will need to be reindexed. Reindexing is the process of reanalyzing and redigesting all of your content.

Changes that require a reindex are:

- Changing the primary search language
- Adding, removing, or altering synonyms

During a reindex, full search will be disabled but autosuggest search based off of article titles will still work. Reindexing normally takes less than a minute, but larger knowledge bases or knowledge bases with very lengthy articles could take significantly longer.

If your knowledge base is critical to your business, we recommend waiting until off-hours to run a reindex in case it does take longer, especially the first time you reindex!

Stemming

Stemming is the process of reducing words to their base or root form. For example, "searching", "searches", and "searcher" can all be reduced to the word "search". This is called the stem. In our search, stemming happens automatically in the background as part of the indexing and reindexing processes.

Stemming helps make search smarter. It's how search engines can find the results you want even when you don't type the exact words.

Stemming is a core task of Natural Language Processing (NLP). NLP is a branch of artificial intelligence (AI) that focuses on how computers process and analyze natural human language. It is not related to generative AI like ChatGPT.

Search relevance scoring

When a search is performed, relevance algorithms are used to weight the results. The algorithms assign a score to each matching search result, and the search results are ranked by score.

The following are simplified explanations of the types of algorithms used:

1. **The exact match**
   All words in the search term must appear in a searchable field for it to be considered a match. This algorithm has no tolerance for typos, but it allows for some variation in proximity of the search terms (words don't have to be in the exact order and can have a few words between them). Any matching result gets a significant boost in score because it exactly matches what a reader typed into the search.

2. **The mostly match**
   75% of the search term must appear in the searchable fields. This algorithm has no tolerance for typos. Matching results get a slight boost in score.

3. **The somewhat match**
   50% of the search term must appear in the search fields. This algorithm is tolerant of typos. Matching results receive the lowest boost.

4. **The maybe match**
   Some of the search terms much match. This algorithm is tolerant of typos. No boost is applied.

In order for an article to appear in search results, it must be considered a match by at least 2 of the algorithms. Scores from all algorithms are combined to determine the article ranking.

The algorithms consider three main factors when calculating scores:

1. **Term frequency**
   This is how often the search term appears in an article. The more times it appears, the higher the score.

2. **Inverse document frequency**
   This is how often the search term appears across all your articles. If the search term shows up in many articles, it will have less weight in scoring. If the search term only appears in a few articles, it will result in a higher score.

3. **Field length**
   This is how long the searchable field is. If it's really short, like an article title, matches will have more weight since it's more likely to be significant.

Tags do not impact search weights or relevancy scores.

A note on algorithms

As noted above, search uses several algorithms. Those algorithms are quite complex.

While you can impact your search scores and results by customizing your search weights, optimizing your content, and adding synonyms, it's worth noting that the exact calculations and behavior of the algorithms is not something our support team can fully explain.