DOROTHY AI

Why AI?

And how DorothyAI leverages user expertise to improve searching?

DOROTHY AI

Keeping your business on top ...
by staying on top of your business.

About us

DorothyAI combines next level, artificial intelligence-based search and centuries of patent law experience into a suite of tools and services that compiles data from a variety of sources in minutes that would otherwise take weeks to assemble.

Executive Summary

DorothyAI is built on the understanding that the most valuable asset in every attorney-client relationship is the expertise of the parties. No practice area exemplifies this relationship more than Patent law. The best patent practitioners have not only a deep understanding of the law and regulations associated with the technology in both the jurisdictions in which they practice and around the world, but also a deep understanding of the technology in the areas in which they practice and the businesses built around the technology. Patent practitioners leverage their own expertise with the expertise of inventors to meet business goals, and in the process, add enormous value to the company.

Like the attorney-client relationship, we believe that the DorothyAI platform will provide better results when the user is allowed to use their own expertise to answer the question posed by the search query.

Artificial intelligence provides the opportunity to add another layer to the attorney-client relationship. DorothyAI has focused on understanding how the best searchers in the world find information and why these techniques work. Our unique platform automates many of the tactics patent practitioners use to identify the most relevant disclosures, allowing practitioners to carry out complete, targeted searches in a fraction of the time it takes to get to the same results using keyword searching.

Beyond automation, DorothyAI seeks to allow users to interact with the platform. Rather than tuning our search engine to a theoretical standard, our platform exposes weighting factors to the users, allowing users to import their expertise into the search and tune the search engine to meet your independent needs. Like the attorney-client relationship, we believe that the DorothyAI platform will provide better results when the user is allowed to use their own expertise to answer the question posed by the search query.

What is relevance?

The single most important determinant of whether a search was successful is:

Are the documents returned by the search engine *relevant* to the user?

Relevancy is our mission at DorothyAI. Our customers should view our returned results and think to themselves, “Wow, these results are exactly what I’m looking for,” in other words, these results are *highly relevant* to my search.
Relevance has two definitions in software development:

1. The returned document actually answers the question or solves the problem posed by the search query.
2. The user can easily understand why the search engine retrieved the returned document.

DorothyAI is a patent search company. It’s not surprising that we have focused on answering the question posed by the query, *i.e.* definition #1.

Legacy Searching

The first step in understanding how to answer our customers' questions, is understanding how our customers would answer questions. Every patent practitioner and search professional has methods for making Boolean queries produce the “best” results. In most cases, the final search is a composite of several different methods in which the results from several searches are combined and duplicates are removed. Relevancy is judged primarily by the searcher.

All of the 100+ patent practitioners we spoke to used a version of the workflow presented in the flowchart below. This workflow is nearly universal. The only difference between a novelty search and, for example, a freedom-to-operate search is when the user limits the search to the claims. During a freedom-to-operate

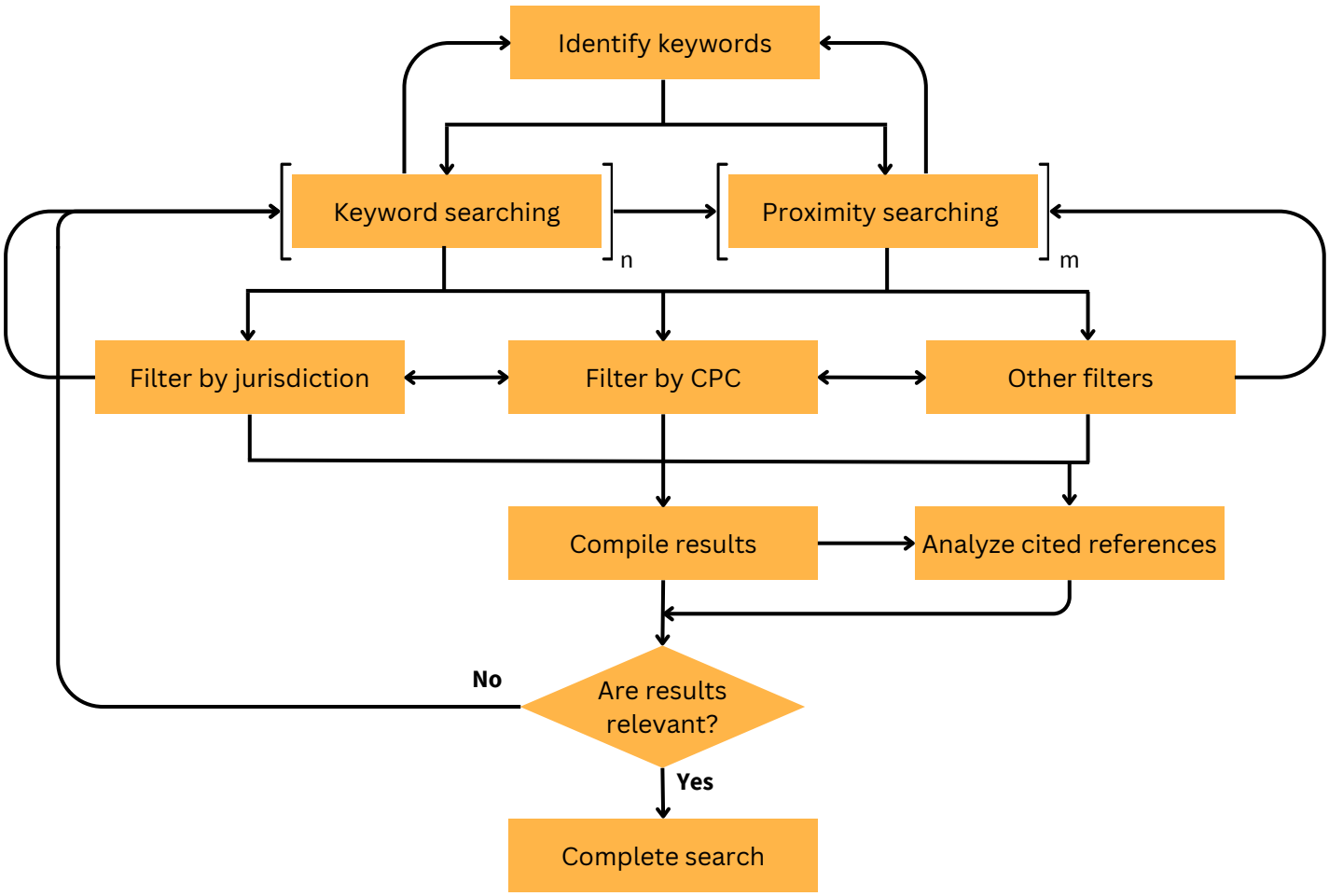

FIG. 1 Workflow for keyword searching

Legacy Searching cont.

search, searchers typically limit the search to the claims (and sometimes issued claims) from the beginning of the project. Novelty searches are often limited to claims to reduce the number of returned results.

The industry relies on Boolean/keyword searching; therefore, the first step is invariably identifying the key terms in an invention disclosure, product description, or claims to be invalidated.

The terms extracted from the source document are used to initiate a keyword search. Keyword searching is an iterative process. Thousands of results are returned by Boolean search engines. To ensure that the best results are identified, a search is performed, results are reviewed, additional terms are added to the search query, and a new search is carried out with the additional terms. Fewer total results may be returned by introducing additional terms, particularly when the “AND” Boolean operator is used, requiring both terms flanking the operator, rather than the “OR” operator, which returns results containing one or both of the terms flanking the operator.

Like the “AND” operator, the additional steps in the workflow are primarily used to reduce the number of results returned by the search engine. Filtering by CPC code and limiting the search to the claims of the documents being searched or to a specific jurisdiction are examples of search strategies that are primarily used to limit the number of results returned by the search engine. These strategies assuredly cause the user to miss references that are relevant, but this is offset by the reduced time required to review the returned results.

Word collocations on the other hand can effectively reduce the number of returned results while moving more relevant documents to the top of the results list. Word collocations use proximity operators to boost the relevancy score of documents if they contain terms near each other. For example, in Google Patent, “NEAR” or “NEARx” means matches are a maximum of x words away. “WITH” means 20 words

Legacy Searching cont.

away, and “SAME” means 200 words away, any order. Collocations allow the user to find documents that use keywords in the same sentence or paragraph. Boolean based search engines infer that documents that have keywords near each other are more relevant and score these documents more highly.

The documents cited during prosecution of particularly relevant patent publications, i.e. “cited documents” or “back and forth references” can provide a wealth of relevant information. In most cases, these references were identified by a human, either a patent examiner or the applicant for the patent, as being relevant to the claims at issue in the patent publication. However, the usefulness of these cited documents can be limited. For example, cited references include secondary references used to make an obviousness rejection that may describe components of the invention described in the patent publication that are not part of the invention being searched, or these may have been cited by the applicant out of caution, having only a tangential relation to the patent publication. Cited by references

Search	Description of Query	Reviewed Resutls	Minutes/ Result	Total Time (hrs)	Cost*
1	Guess Keywords	100	0.5	0.83	$83.33
2	Add/Remove keywords	100	0.5	0.83	$83.33
3	Revise keywords	100	0.5	0.83	$83.33
4	Guess synonyms of keywords	100	1.0	1.67	$166.67
5	Revise synonyms of keywords	100	1.5	2.5	$250.00
			Total	**6.67**	**$666.67**

* "Cost" is based on the cost to the firm of $100/hr per associate encompassing salary and overhead expenses . Cost is not based on fees billable to the client.

Legacy Searching cont.

suffer from the same issues. Moreover, reviewing cited documents is often limited to only the most relevant returned results from the other searches. There is no assurance that the other results do not include relevant cited results.

After more than 6 hours of searching using these various tactics and tools, the searchers we interviewed typically have a relatively good set of results. In some cases, the project is considered complete and the searcher can move on to another project. In other cases, the results may be spotty or underwhelming, prompting the searcher to continue doing iterative searches, using these tools, and looking for better results.

Artificial Intelligence and Context

Context is lacking in the process described above, providing artificial intelligence, in particular, natural language processing ("NLP"), with a distinct advantage over keyword/Boolean based search engines. Many of the, "Where the heck did that come from," moments experienced by searches are a result of terms being identified in a reference without context. NLP engines infer context by scoring the search query as a whole and identifying sentences, paragraphs, and sections of patent publications that have been similarly scored during indexing. This process is often referred to as "creating embeddings." References that include all of the sentences and paragraphs with similar scores describe similar concepts and are returned at the top of the results list.

At best, proximity operators offer a mechanical means for providing context. However, whether search terms are within 3 words of each other or not is limited in its usefulness. What if the author used a synonym of one of the words that is not in the query, or described the component rather than naming it. NLP based search engines overcome these difficulties and return the result as relevant, by tokenizing keywords, linking them to a thesaurus of synonyms of the term or phrase, and by similarly scoring the word and its description.

Let's say you are searching for a kitchen knife that cuts food using a laser. The query [laser "kitchen knife"] in Google Patents primarily results in U.S. patent publications describing methods for making kitchen knives using laser welding, laser engraving, laser annealing, etc., and robotic cookers that use lasers to scan the surface of food.
A modified query [laser NEAR "kitchen knife"] returns U.S. patent publications describing laser assemblies used in machining operations and surgical tools. The term "kitchen knife" does not appear in any of the first 5 results. These results seem further from our target.

Modifying our query to, laser NEAR knife kitchen, results in U.S. patent publications describing cooking apparatuses that use lasers to cook, not cut, food. We are getting somewhat better results, but the search seems to be directed more to cooking than cutting.
Query kitchen laser NEAR knife results in similar cooking apparatuses. Here's the description of the provided by Google patents as the most relevant result:

> *800 , the robotic hands 72 execute the mini-manipulation 770 of cracking an egg with a knife, where the optimal way to execute each movement in the cracking an egg operation 772 , the holding a knife operation 774 , the striking the egg with a knife operation 776 , and opening the cracked egg operation*

This is not what we are looking for.

Artificial Intelligence and Context cont.

Reviewing the results reveals that terms “laser cooking” and “cooking” are now highlighted. In attempting to put the query in context, Google Patent took the presence of the term “kitchen” in the query to infer that the search relates to “cooking” and returned results that combine the concepts of lasers and cooking as highly relevant. In this case, tokenization, i.e. replacing “kitchen” in the query with the synonym “cooking” caused the search engine to produce spurious results. Notably, the phrase “kitchen knife” was not identified as a cutting utensil for food in the previous searches. The terms “knife” and “kitchen” were taken completely out of the context of cutting food.

With this in mind, we modify the search to, food cutting laser NEAR knife. Finally, we get results that look promising. Notably however, the most relevant result according to Google Patents is a publication entitled “Method and system for more accurately determining nutritional values and reducing waste of food items,” with the description:

> *Lend the concession stand personnel a warmer for the evening so the concessions can keep their pizzas warm throughout the game. Offer to include or recommend that the fundraiser purchase a roller blade cutter to cut the pizzas. Even though the equal slice cutting tool will cut equal slices at the…*

At least we’re seeing synonyms for “knife” (“blade”) in the results.

Our natural language query ensures that DorothyAI’s NLP has the context it needs to find the best results. NLP is much better at searching things in context than Boolean based search engines. Indexing also allows NLP based engines to understand the words “kitchen” and “knife” are often used together to describe a cutting utensil. Indeed, 100% results returned by Activ8 Novelty using the search query “kitchen knife that cuts food using a laser” are directed to cutting utensils and apparatuses. The top references describe inventions that cut food: broadly (#1), eggs (#2), rolls (#3), or tuna (#4), and all of these references describe a laser cutting means somewhere in the specification. Notably absent from our results are the laser cooking apparatuses that make up the majority of the results returned by Google Patent.

The results returned by Activ8 Novelty are clearly more relevant than the best results that were obtained using Google Patents in this simple search. Moreover, DorothyAI’s results both solve the problem posed by the query, and it’s obvious why they were returned.

Tactics for Legacy Searching

IAs we saw above, the key to understanding relevancy is understanding **context**. Humans are extremely adept at understanding context. Because the Boolean based search engines have no means for extracting context from a keyword query, the terms “kitchen” and “knife,” in the above example, are analyzed in a vacuum, completely without context. The lack of context makes keyword based search engines unreliable and over inclusive, increasing the amount of time necessary to carry out a search.

Nonetheless, even the most intelligent humans can have trouble extracting context from patent disclosures. Boilerplate definitions, lists of potential components and ingredients, inventors and practitioners being their own lexicographer, inconsistencies in the disclosure, and hypothetical examples make understanding context tough. Moreover, the terminology used in patent applications can have different meanings in different contexts and, in some cases, similar terms can have different meanings depending on the technology it is describing.

Searchers have attempted to solve the context problem in three main ways: (i) by adding additional terms to the query in an attempt to limit the search to a particular type of technology, (ii) limiting the search to particular fields, and (iii) by limiting their searches to specific classifications.

For many searchers, adding terms to the query is the first tactic used to eliminate the spurious results from their searches. The logic is sound: Require the search engine to find more terms and fewer results will be returned. This is the same tactic used with searching consumer products. For example, an upcoming birthday may prompt you to search for “kitchen knife” in Google or Amazon. This search will likely yield too many results, so you may limit your search to “paring knife,” “chef’s knife,” “Japanese kitchen knife,” or “Shun kitchen knife.” In each case, fewer more targeted results are returned, but at what cost? By limiting your search to “Shun kitchen knife,” for example, you might miss a more highly rated, Mitsomoto knife, or a sale on Wusthof knives.

This is what happens when you add elements to a patent search query: your search returns fewer results. However, potentially important references that discuss key concepts in the detailed description will undoubtedly be left out.

The second tactic often employed by searchers is to limit the search to particular fields, for example, title, abstract, and claims. Like adding terms to a keyword query, limiting the search to fields that contain less information, like the title, abstract, and claims, will undoubtedly miss relevant documents.

Tactics for Legacy Searching cont.

IThis is particularly dangerous for novelty and invalidity searches. Hypothetical descriptions and examples that support, but are not integral to a patented invention and are not necessarily into the title, abstract or claims, are often the most relevant disclosures in a novelty or invalidity search. The heroes of the patent world find these disclosures, and put them to use for their clients.

Patent Examiners mitigate the impact of limiting search fields and adding terms by performing numerous searches with various combinations of key terms. The USPTO's primary search engine, EAST, is set up to allow patent examiners to combine search terms and results quickly, producing a large number of searches in a relatively short amount of time. For example, on average 75 individual searches for each application were carried out by a group of 4 U.S. patent examiners in the medical device art unit that we studied. Nonetheless, one Examiner we interviewed commented that this average is low.

As the prevalence of keyword searching has increased, the third tactic, limiting the search to particular classifications, has become less prevalent. From the Dewey decimal system to the West Key Number System to the Coordinated Patent Classification (CPC), humans have been classifying information into codes for centuries to help other humans find the information. Classification is a powerful tool, particularly when the information is held in a card catalog or patent office "shoes." Unfortunately, choosing the correct classification from the 250,000 CPC categories can be tricky. Searchers often find it more intuitive to add terms to the query or limiting search fields rather than limiting the search to particular classifications even though classifications more directly provide context for the keywords.

The DorothyAI Answer

Embeddings are great. However, producing embeddings that accurately untangle patent disclosures is impossible or, at least, improbable given the current state of technology. This explains the inconsistencies searchers experience when using most AI patent search tools.

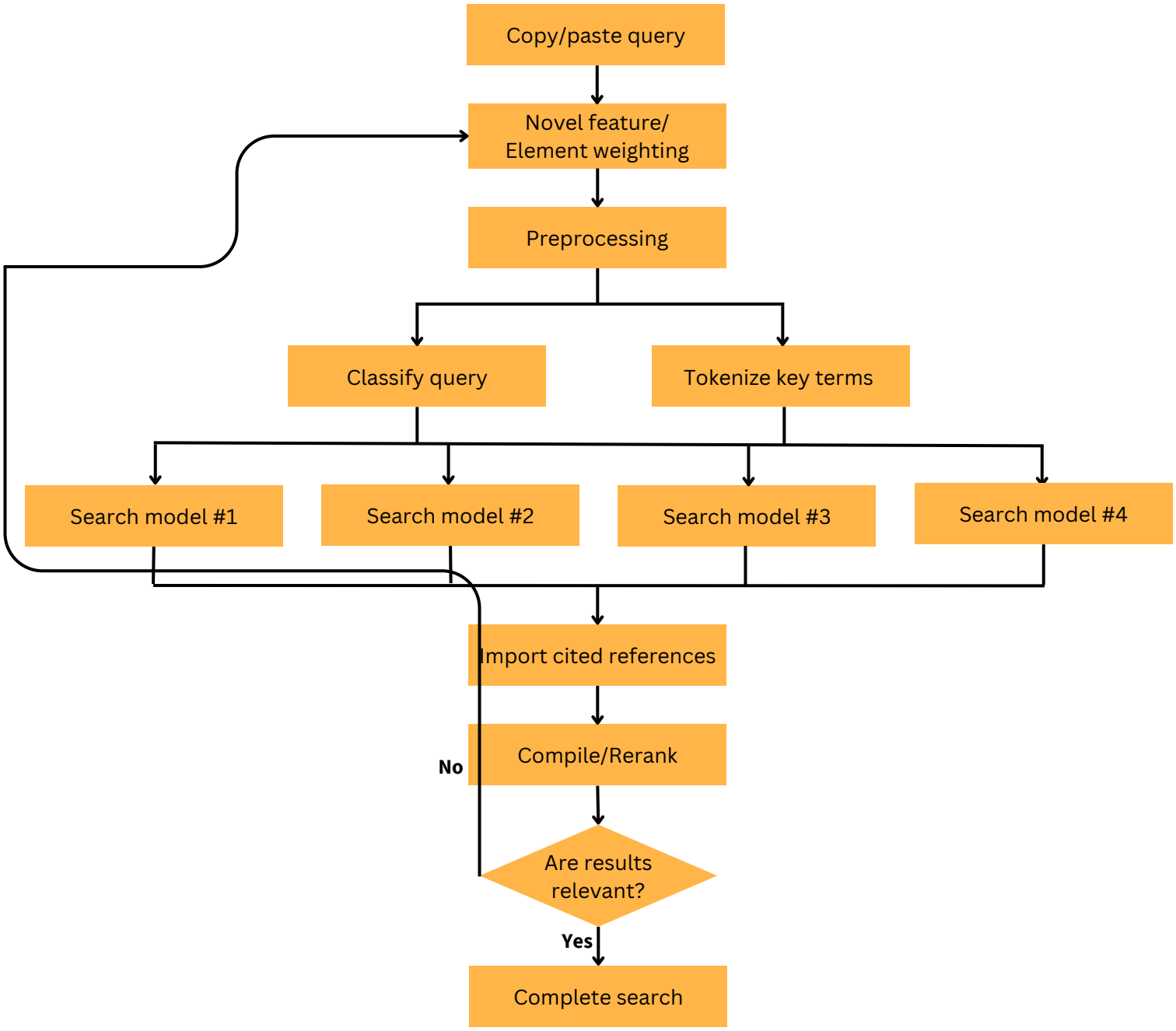

FIG. 2 DorothyAI workflow: NLP models and automated keyword search tactics

The DorothyAI Answer

Like humans, some search algorithms are better at searching certain technologies like software or electronics and not so good at searching other technologies such as chemistry. AI search tools are generally very accurate if the disclosures on which the model is trained is similar to the technology being searched. The results are inaccurate, however, if the searched technology is less prevalent in the training documents. To date, no generally applicable AI model has been created. Developers and data scientists continue to train and tune AI models with varying results.

DorothyAI exploits these characteristics of search models by incorporating several search models into the platform, called an "ensemble." The platform provides better results than each of these search models individually, making Dorothy an expert at searching all technologies.

In addition, DorothyAI has automated many of the "tactics" human searchers use to remove spurious results and focus the search. Like human searchers, DorothyAI uses classification, synonymization, and collocations along with embeddings models and Boolean search in the ensemble. A specially trained neural net that scores the results returned by combinations of the search models for relevance.

A proprietary framework underlies DorothyAI's search platform. This framework allows the various components of the platform and search models to work together seamlessly and simultaneously. The input query is processed, keywords are extracted and replaced with synonymization tokens, and necessary information is passed to an ensemble of search models. This process allows the user to simultaneously carry out multiple searches simultaneously while eliminating the need for iterative searches that are required to manually create a query incorporating a sufficient number of synonyms. This portion of the DorothyAI workflow alone saves users hours of time spent carrying out iterative searches.

DorothyAI saves additional searcher time by automatically compiling cited references from each of the results returned from the ensemble of search models, checking yet another box in the searcher workflow described in Fig. 1.

The DorothyAI reranker is another marvel of technology. The compiled results from the ensemble and their associated cited references are passed to the reranker, where duplicates are removed and the results are scored for relevancy, using statistical criteria, for example, how many models returned the document and a neural net trained using data curated by professional searchers and patent practitioners. The reranker also identifies key phrases and/or claims that match the input query (incorporating synonyms of course).

The DorothyAI Answer

The entire process is carried out in about 2 minutes producing a list of results sorted by relevance, including text snippets to help our users quickly confirm that the results answers the question or solves the problem posed by the search query and to allow the user to easily understand why the search engine retrieved the returned document.

By combining NLP and AI techniques and automating more traditional search tactics, DorothyAI solves the biggest problem with AI-based search engines, producing excellent results regardless of the technology. What's more, the framework is flexible, so as new search models are trained and new search techniques are developed, DorothyAI's results will continue to improve.

Searcher-Platform Synergy

Our goal is to produce a tool that creates synergy between the searcher and technology. DorothyAI allows the searcher to control nearly every aspect of the search by targeting specific elements, adjusting various weighting factors , and searching across different databases. The platform becomes an extension of the searcher. A partner, using the user's expertise to produce better results.

At the same time, giving more control of the search to users reduces bias. Bias is a concern for all searchers and patent practitioners. Bias is often introduced into AI systems through training data. An AI system will understand that the information is correct if it consistently “sees” that information in training data. Bias can also be introduced when the models are tuned. Information that a developer perceives to be correct may be weighted more highly than other information, producing “weighting bias.”

It is very difficult to remove bias introduced by training data, since the disclosures of patent publications cannot be changed. Weighting bias, on the other hand, can be reduced. Weighting factors used in most search platforms are static. For example, Google Patent applies the same search model with identically set weighting factors to every search. Weighting bias is introduced by the developers perception of what references answer the question posed by the query. In contrast, DorothyAI exposes certain weighting factors to the user, making them adjustable.

For example, Activ8 Novelty allows the user to identify a “Novel Feature.” The Activ8 Novelty query is a block of text describing the invention, the “Invention Disclosure” and a novel feature or an element of interest described in the Invention Disclosure. The Invention Disclosure is meant to encompass an inventor's description of an invention in an invention disclosure document form produced by many IP departments or a brief description of an invention in a patent application. The user is free to copy such disclosures and paste them directly into the query box. The Invention Disclosure gives the search context, providing a means for focusing the search by classifying the Invention Disclosure and upweighting similar technologies.

The Novel Feature acts as a “north star” around which returned results can be assembled, putting control of the search in the hands of the user rather than a developer, who may or may not understand the technology being searched and providing yet another means for eliminating spurious results. Users can adjust the weight of the Novel Feature. A broad search can be carried out by downweighting the Novel Feature, or to make the search more targeted toward the Novel Feature, the Novel Feature can be upweighting.

Searcher-Platform Synergy

Activ8 Freedom was also created to put control of the search in the hands of the user. An Activ8 Freedom query is composed of individual text boxes for each component or “Element” of the product being searched. The individual elements reflect the claim elements of claims describing the product, providing another opportunity to simply copy and paste. Breaking the query into individual components allows Activ8 Freedom to identify references that describe the individual Elements alone or in combination with other query Elements.

Activ8 Freedom allows the user to weight the elements of the query individually. Weighting in an Activ8 Freedom search is carried out in two steps: (i) identifying “Required Elements” and (ii) weighting required elements against each other.

The first step, identifying the Required Elements, allows the user to downweight elements by clicking a checkbox. Required elements are the elements that the user is concerned may present a Freedom-to-operate risk. Unchecked elements are typically those elements that are known to be in the public domain and do not present a Freedom-to-operate risk. For example, a user searching a pharmaceutical composition would check the active agent, but might leave an Element describing a diluent or excipient unchecked. Similarly, a user searching a device would check the Required Element checkboxes for a combination of components that perform a necessary function for the device and not check Required element boxes describing power supply elements that feed those components.

Downweighting elements in the public domain eliminates a vast number of results with claims describing products that include commonly used elements or components. Unchecked components remain part of the search, but they are down weighted, allowing the elements of most interest to the user to move to the top of the list of results even if that element or combination of elements is rare. Weighting required elements against each other allows the user to focus the search even more specifically on the components of the product that present a risk to the client. Here again, Activ8 Freedom allows the user to control the search, ensuring that the elements that are most important to the user to move to the top of the list of search results.

More Information

www.dorothyai.com

Contact

Sharon Shofner-Meyer, J.D., MBA
President
sharon@dorothyai.com
651-271-7169