AI SEARCH TOOLS FOR PATENTS

How to test & compare them?

APRIL 16

IPSCREENER

Authored by: Linus Wretblad

IPSCREENER
AutoMatch technology

Intellectual Property & AI

Lately the AI buzz has also flooded the work of the information professional. This is indeed a new option to expand the toolbox to find prior art. It offers smarter automated information retrieval through text-based searches. However, the common approach from solution providers seems overall to be "Yes we got it too, AI is the future" rather than to be transparent on the actual performance.

The evaluation process of new tools has become quite a challenge. On one hand, the AI black box is exceedingly difficult to understand with all the complex algorithms behind it. On the other hand, it is a time-consuming process to validate the performance manually. Finally, it is the question of how you would make best use of this new gadget to maximize the performance as a user.

This article will address these topics, from explaining the background of evaluations to exploring hands on insights on how to practically benefit from AI search tools.

1.The background to unveil the black box

As a user you would like to know straight on if the quality of a tool is good enough. That the results contribute to a more efficient search and decision process.

In short: does it help me or not? Other questions are: what is the actual performance within my specific technical domain and how does the performance of different providers compare?

The focus of the evaluation should be on the actual performance delivered and not on technology used. Comparable to when buying a car with automatic gearing, it is not how the gearing box works but the performance of the car you are interested in as driver.

It would be great to have a universal model for evaluating such text-based search tools. To have a standard test to validate any system that from a text input, without human interaction, automatically retrieves supposedly relevant documents.

Secondly, it should be even better to have an automated procedure to run a test instead of manually defining and reviewing search queries. This is the reason why I write the article to share our knowledge from research and experiences we have done to create transparency.

Let us first back up a little to establish a ground for such an evaluation. Information retrieval, which is the general term in academia for research on searching and finding of relevant data, has been around since the early 1960s. There are two basic metrics defined for measuring the performance, called Precision and Recall [1] (read also more online at Wikipedia).

Precision defines, as it sounds, the quality ratio of the retrieved hits or documents.

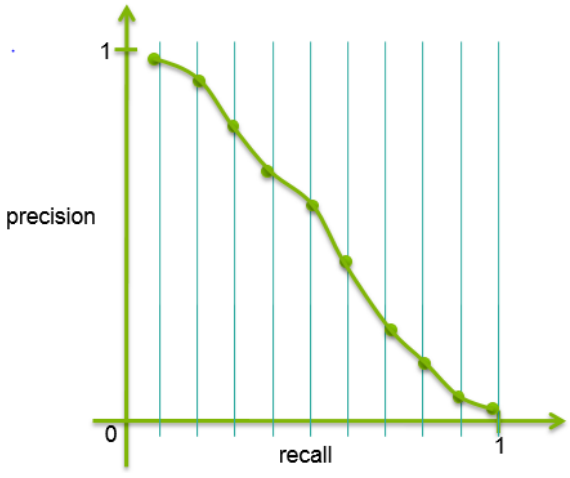
In short: how happy are we with what we have got?

This score is the relation between the number of relevant documents in a result and the total amount of documents retrieved. In a perfect world you have zero noise and all hits are relevant to the subject of the input query text.

Recall shows the proportion of correct answers found.

In common words, have we found all the documents we were looking for? This score tells us how many of the relevant documents of a known set of right answers that were found by the search tool.

The relation and tradeoff between the two scores a contradictive challenge. By looking at more documents in the result list to identify further correct answers and increase the Recall, we will also decrease the Precision accordingly. The relation is shown in the following graph.



The evaluation procedure of the two metrics demands a preparation of test queries and associated correct answers, a so-called ground truth.

Automating measurements of the Precision is complex as you need to know the relevancy of all retrieved hits in relation to a query text. Algorithms are yet not accurate enough to assess if a document is explicitly relevant or not.

Thus, it remains to do a manual review to obtain reliable score. That is, an IP specialist understanding the technical domain of the query must assess and categorize each retrieved document as relevant or noise. This is a very tedious process and, in practice, limited to random sampling with quality check procedures on a few cases.

Verifying the Recall is more straight forward. Every case must be defined with a query (text input) and a correct set of answers (documents to be found). You limit the validation procedure to identify how many of the known documents that were found.

An IP specialist could manually assess this number and calculate a performance ratio. It still requires quite some manual work, both to prepare the queries and to assess the score. However, for this task automation is possible, as the analysis is limited to identify known answers.

For performing these kinds of evaluations, spinning a handful of queries is far from enough. You would need thousands of test queries to get statistically reliable quality

measurements, which is not feasible to do manually. There is some research done on defining models for automatically evaluating Precision and Recall based on log data or search reports.

For an information professional, the performance question is quite simple. You like to see the most relevant document in the top of a hit list (high Recall) to save time during the reading process. At the same time, you only want to see relevant documents in the result list (high Precision) to trust the results.

Here we have the dilemma with Recall and Precision in its very essence, improving one will lower the other.

The balance between the two is the key question to create trust in black box solutions for automated searches.

To have a result that shows both high Precision and Recall.

Furthermore, how many documents is reasonable to review in a screening process, given a certain Recall and Precision score? From these thoughts, we should define an evaluation model. The next step is to create such a test platform for automatic evaluation and creating a baseline structure.

[1] Information Retrieval Evaluation: Harman, Donna; Morgan & Claypool Publishers, 2011

2. Defining a Baseline and a Ground Truth

I described in Chapter I the pains of measuring the performance of a black box search solution within the industry and how precision & recall are key indicators to assure the quality of such performance. This section covers how a platform for automated measurements of the performance could look like.

For the task we did a research study together with the TU Vienna (special thanks to Mihai Lupu and Linda Andersson) to review and refine existing procedures. Much of the research was based on evaluation models presented in the TREC research initiative [1] (Read more online at NIST).

The research ended up in a platform for assessing the quality of text-based searches on a large scale. Our key questions were:

- What input delivers good results and where are potential pitfalls?
- How can we reliably measure the performance to guarantee continuous improvements?

The goal is to exemplify quality measurement in an information retrieval procedure and to track such scores across different industries. As finding good documents is a primary objective of a search, the first evaluation model traces recall scores and the associated ranking (at which position in the result list answers are found).

The challenge is to define a methodology and standard set of queries together with the known correct answers (documents), i.e. a ground truth.

Luckily, the patent domain has indeed such a "truth" if we presume that the examination reports from the Patent Offices are the correct answers to be found. Of course, it might be possible to find other prior art documents being as relevant (or even more relevant) than the documents cited in the examination report.

However, this definition of ground truth presents a transparent evaluation and permits us to automate the procedure on large set of queries.

That fact is essential for obtaining a reliable performance score in an efficient way.

Thus, the ground truth consists of queries (the patent applications) and the known answers against which results are compared (the relevant documents cited by the examiner). We measure the recall by calculating the ratio of how many of the relevant citations mentioned in the patent office examination report did the automated tool find.

The approach makes it possible to create an automated evaluation procedure and is a first step to open the black box to comparison.

From the information professional's perspective, we should focus the evaluation model on recall scores within limited number of presented result hits.

We want to understand the ranking of correct answers retrieved; how many of them are found and where in the result list?

Optionally, you could consider measuring at what position in the result set did the tool find the first correct answer (e.g. the first knockout citation document was found as document number 187 in result set). However, we chose to focus on the interval approach for clarity reasons and better overview.

To achieve this, we could define ranges corresponding to intervals normally accepted or used for screening purposes. E.g. how many of the predefined answer documents that "should" be found in a search were retrieved among the first top 10, 25, 50 and 100 founds hits, respectively.

The second challenge is to trace the performance across technologies. Again, the patent collections are very structured by the IPC/CPC classification system, categorized in main technical sections from A to H and an associated hierarchy (explained more at WIPO, reference cited).

This permits us to measure and compare the search performance for different technologies, as the query sets (documents) are possible to group according to the classification. It also creates transparency around the quality, respectively. The second challenge is to

trace the performance across technologies. Again, the patent collections are very structured by the IPC/CPC classification system, categorized in main technical sections from A to H and an associated hierarchy (explained more at WIPO, reference cited).
This permits us to measure and compare the search performance for different technologies, as the query sets (documents) are possible to group according to the classification. It also creates transparency around the quality of automatically generated results.

For statistical relevance, we required to use at least 100 queries per defined technical area.

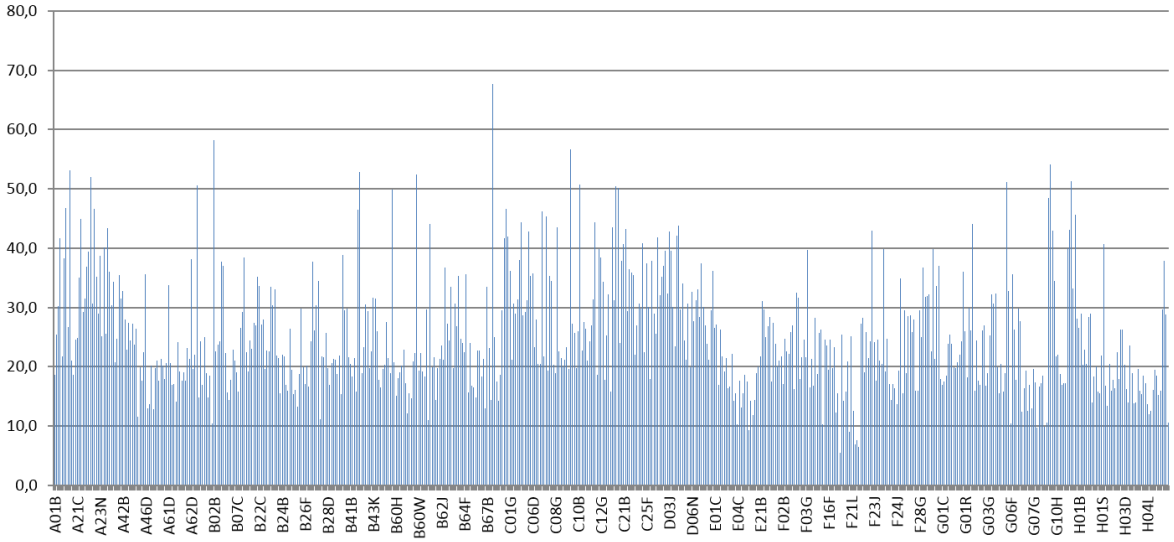
To cover the overall populated technical domains, the test collection added up to a total of 175000 queries. They are composed of randomly distributed patent applications with examination reports, spread over classes and applicants from the latest 20 years. It constitutes then the baseline query set to be used for measurements.

Based on the assumption that the text input affects the results the most, we elaborated on different formats and associated output performance. The old rule of "Garbage in, Garbage out" applies even more for algorithms.

Thus, we chose to have toughest input possible for a "worst case" evaluation: only using title and abstract as input, with grouping of test queries on Subclass level (e.g. G06F) and Group level (e.g. G06F-003). This represents a good tradeoff between granularity of query sets, harmonization of the length of the query and the technology focus.

As a measurement reference, and to establish a graphic presentation of the results, we run the baseline queries through a text based prior art screening algorithm.

The resulting diagram below shows the recall scores as a percentage. It shows the average number of answers (documents) found within the first 100 hits shown in view of the total set of citations reported by the examiner.



The scores in our example are grouped on a Subclass level (G06F). opportunities considerably. You really need to make conscious decisions. There are horrible mistakes out there that hopefully will work as reminders.

The performance is showing a major variation depending on the technical classes in question. The fluctuation does not come as a surprise and applies typically to any automated tool out there on the market. However, even though this conclusion is obvious, it is normally not communicated to me as a user.

Furthermore, as the input text has the highest impact on the quality of the output, it is important to understand when I need to redesign an input for better performance.

Using a baseline is one way to unveil and identify such domains.

More important, by including multiple result sets in one view you may also compare the performance for different providers of automated tools. It would support to understand where one is better than the other.

However, the diagram divided on individual classes still becomes complex. The end user would certainly wish for more transparency. It is not common knowledge what each class relates to, and even experts do not know them all by heart. It would be great with an analysis view in a more comprehensive presentation format.

[1] *TREC chemical information retrieval – An initial evaluation effort for chemical IR systems*: Lupu, Huang, Zhu, Tait; World Patent Information, 2011 advantageous than to reveal the content through a patent. Drawback is that it is difficult to keep the secret.

3. AI performance understandable by everyone

I elaborated in part 2 around how-to setup an evaluation platform for text-based search tools. I defined a measurement methodology using the patent examination citations as ground truth and showed what performance in such a graph could look like.

In this section I will take it one step further to convert the results from a technical classification view to a more transparent and understandable representation.

The classification distribution is difficult to read for other than information professionals. It would be more convenient to have the query sets explained by industry domains and expressed in common words. In the ambition to show the quality score of a text based search tool in a more user-friendly way, we ended up exploring the WIPO definition of technical areas [1] (Read more online at WIPO).

The work clusters patent subclasses and groups the data into 35 generic technical domains.

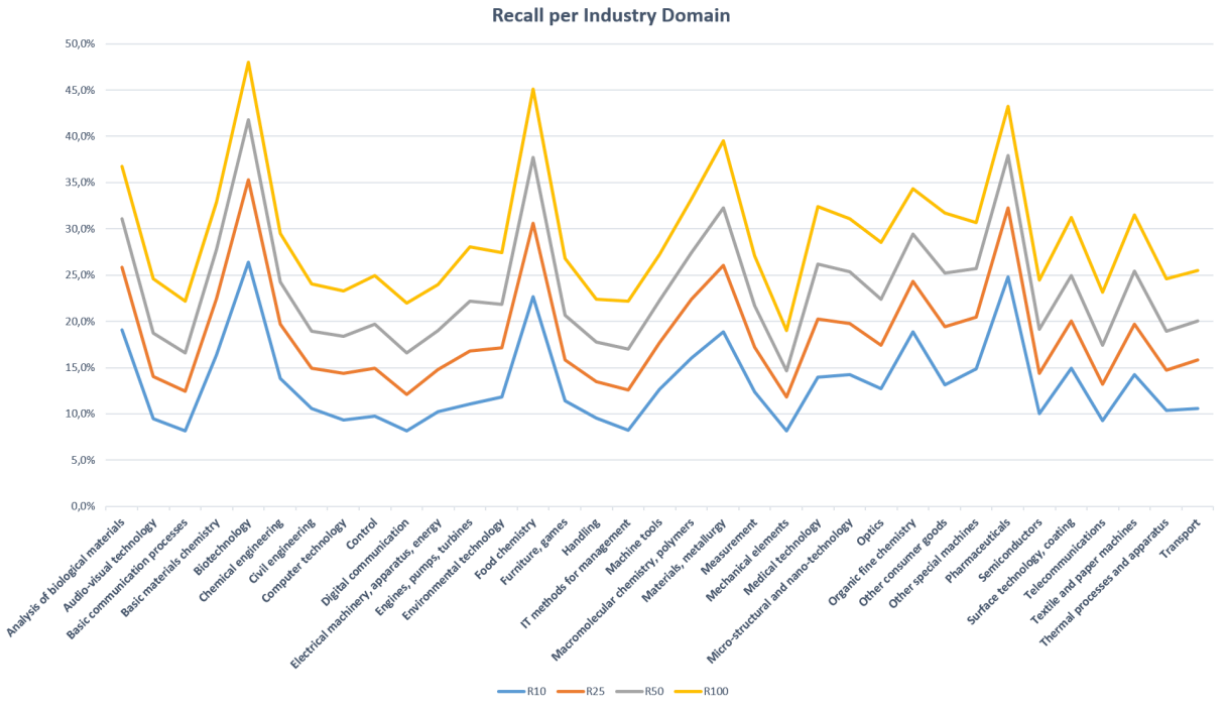
The result covers the head topics for human research and development.

Transforming the baseline score for the class distribution presented earlier into an equivalent graph across the WIPO domains shows another perspective and overview. As common headlines describe each technical area, we get at a more straight forward and transparent representation.

In addition, the aggregated view also enables us to include the Recall for several ranges, where we choose to show top 10, 25, 50 and 100 simultaneously. The expanded presentation shown below yields indeed a much better overview.

However, we did some general observations during the analysis. For technical areas where the invention is presented by non-text descriptors such as drawings, e.g. fine mechanics or process control, the recall is much lower. On the other hand, for technical fields such as biotechnology and food industry where the invention idea is better explained (and understood) by a regular text, the recall is impressively high.

The insight, on how the recall performance for text-based tools are varying and depending on the technical field you are active within, gives us a better understanding on what to expect and how to handle it. Let us say that in the innovation review process we require 20% recall on automated searches for improved efficiency.



As you probably noted, regardless of the representation, by class or by technical domain, the performance is showing variations depending on the technical field. There are of course many possible sources for this, which are yet not fully explored.

Then, for "Biotechnology" cases we only need to review the Top 10 documents to ensure that effective performance. However, we would need to analyze the Top 100 documents to reach the same Recall score for "Mechanical elements" inventions. It stresses the importance of being able to verify the performance for

your specific target R&D domain and adapt the search analysis process accordingly.

With such a baseline analysis run on several providers, it is possible to compare AI based search tools and to identify the specific performance within a certain technical field as well. The baseline measurement is the basis to be able to improve a tools performance as well and to keep track of developments.

Therefore, quality metrics are important. It enables the client to choose the best option for that industry and supports the providers to optimize the text matching procedure to perform better.

Our vision is that the community will have a Standard Baseline Toolkit with queries and appropriate settings.

It could assist users in performing comparative measurements and support them to take informed decisions. Such a toolkit could also inspire the industry to be more transparent and hopefully open for quality comparison. Preferably, an initiative like that should be governed by organizations such as WIPO, PIUG, CEPIUG, PDG and major patent offices.

We want to contribute to the evaluation and the better use of new smart text-based search tools by sharing our data and knowledge. Consider this as a starting point for further discussions and collaboration. To explore this further would need input from the different stakeholders, where some important topics are.

- New skill sets for the information professional using more text-based searches.
- How to collaborate on a Standard Baseline Toolkit and evaluations.
- Share experiences of pains and challenges within the IP community.

This may give future thoughts on best practice also serves to evolve the role of the information professional. By understanding the boundaries of a tool, what to enter and when, we will make better use of these new search possibilities.

How good must the recall be to be acceptable for me in my daily work as an information specialist? What could I do as a user to improve the performance? How may new tools improve my searching processes?

[1] Concept of a Technology Classification for Country Comparisons: Schmoch; WIPO, 2008

4. Best practice for AI enhanced searchers

In previous sections I elaborated on how to measure the quality of search tools which are using just a normal text as input. This included thoughts on how to define an evaluation platform to measure performance of AI search tools. A test methodology was defined using patent examination citations as Ground Truth. I also showed some graphs illustrating what performance results across different technology domains could look like. This part will, based on feedback and further analysis, review how to use AI search tools more efficiently. The aim is to identify some possible best practice.

We should simply accept that being able to use text-based queries adds a brand-new gadget to the searcher's toolbox and that we need more knowledge on how to

use them properly. This is well described by Aalt van de Kuilen article (read more here), addressing how to embrace new tools and possibilities. However, this probably also demand a somewhat different mindset for the search approach.

Being an information professional or a "searcher" for prior art, the work traditionally involves understanding a technical concept, performing a search, composing search strategies, assessing the relevance of documents and finally communicating the findings on the state of the art in a proper report format. The part affected by the new options is the converting a technical description into a representative search strategy, which today is an essential part of an information professional's skill set.

Thus, I am trained to extract the core of the idea and tailor my own search strategies from scratch by applying boolean and proximity operator s and combining essential keywords with synonyms, applicant names, classification codes, meta-tagged data etc. There are numerous good articles on this topic, especially by Evert Nijhof [1] (online version), [2] (online version), [3] (online version).

I demonstrate this procedure of creating a search profile by a schematic example applied on a simplified text:

> *We are proposing a new trapping device for catching unwanted animals (e.g. mice) in home or office environments. The cage construction has a door that is activated when a mouse is touching a substance (e.g. cheese) on a sensor provided inside the cage.*

In an associated query strategy, you would try to identify relevant words, synonyms, and related distances in between key words. The searcher then run the queries, reviews found documents and in an iterative manner refines the query to generate more relevant results. For the text above an associated very simplified figurative query could look like:

> *((mouse or mice or rodent+) proximity (cage+ or trap+)) proximity (sensor+) and (group) in cpc/ipc-class*

By adjusting the number of words of "proximity" value as well as the granularity of the class description the search string will be either broader or narrower. That is, a longer distance and higher-level class will catch more documents (with higher recall and lower precision), whereas a shorter distance and precise class will catch fewer documents (accordingly with lower recall and higher precision).

So, we are as searchers accustomed to manually convert a text into a corresponding query representation, then successively adjusting the search strategy.

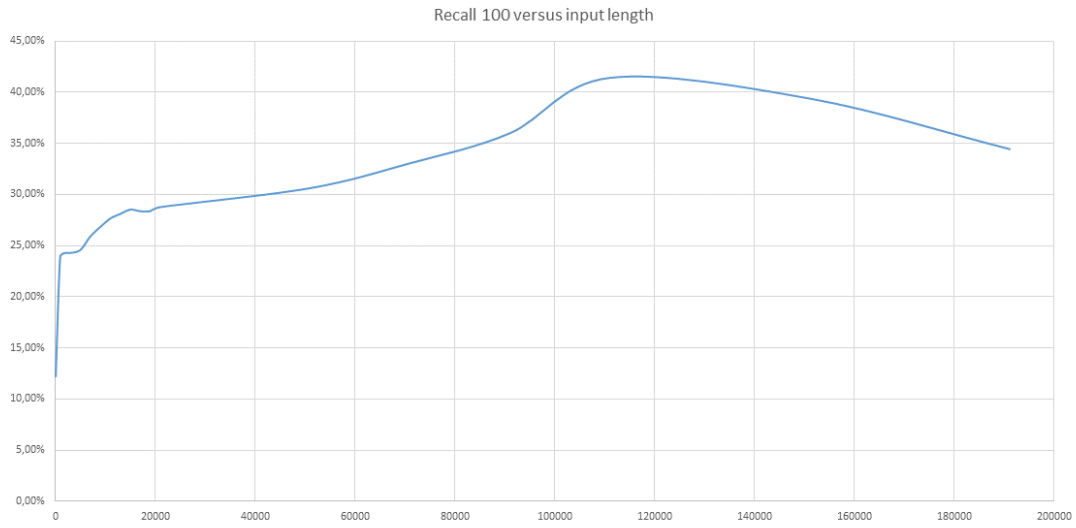
However, when using AI-based tools we shall feed in a text instead. Thus, in the example you would simply start-off by using the whole paragraph containing both problem, solution and maybe background as well. Even though obviously compared to manual approach this input includes irrelevant information such a prepositions and other noise terms.

However, for the algorithmic analysis this semantic data provides additional teaching and ground for training.

Consequently, a first challenge for the information professional is then to accept to use a "raw" text (such as an invention disclosure or application) and avoid transforming the text into one or more manual queries.

Another insight is the possible need for adaptation of a search text depending on the technical domain.

We saw in part 3 that you would have to review quite different numbers of hits depending on the technical field to get a similar recall score (the example showed



that for an area less suited for text-based searches such as "Mechanical elements" you could need to review up to Top 100 documents to have the same Recall score given by the Top 10 documents for a high performing area such as "Biotechnology").

These results were given using the same format of the query texts (in that example the title and the abstract only).

In view of those insights some questions are raised; what text parts should be entered and how could I as a searcher change the input to get better performance? Also, what tuning towards a technology domain is needed, and how long should an input text then be?

This is especially interesting for areas with performance challenges, and where a possibility to boost the score is wanted.

Simply put, could an information professional give the AI a manual "push" for better results by adapting the text?

Consequently, the question is to understand what amount of input data to use for best performance. To achieve a better understanding of this, we did a study on how different text length inputs affect the quality of the output. The first test collection is a technology mix of 50000 patent applications with more than 200 000 characters of text length. The analysis was based on baseline runs varying the input from using just the title to using the full document. Non-text parts such as charts, chemical formulas etc. were excluded.

In general, we see that adding more text to the input did yield considerable

performance boosts to begin with, only then with much larger texts slowly decreasing. This is shown in the graph above, where recall is plotted against the amount of text (number of characters) entered. The score shows the recall within top 100 hits retrieved.

This behavior is probably logical; too little data might be too vague and lead to a misunderstanding the actual focus, where too much data adds noise and you are suddenly unable to identify a distinct target description.

Our indicative first analysis has two main takeaways:

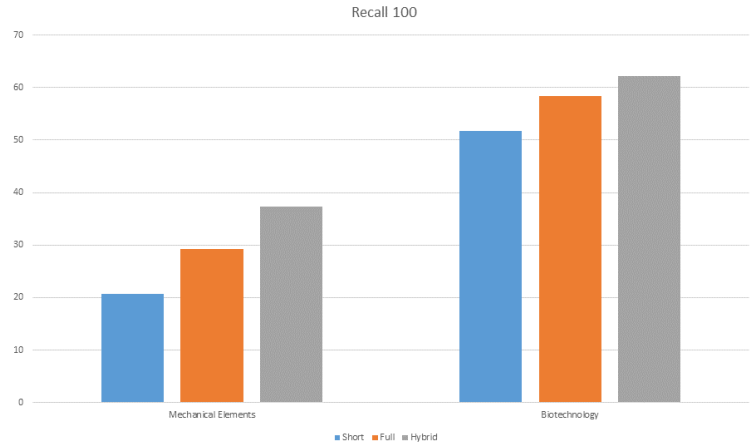
Firstly, it was indeed surprising to identify the amount of data to be "too much" as input, and hence reach the peak where the recall starts to decrease.

We see the performance increase slowly all the way to finally peak off at about 100 000 characters.

Simply put, this represents about 10-15 pages of technical text until we see a performance dip. "This will introduce too much noise", would be the normal reaction where it is indeed shown to be quite the opposite. As reference, our baseline example in Chapter II of 175 000 patent applications sampled over all technology domains has an average length of 89 000 characters per patent document.

Secondly, the analysis shows that the performance increases dramatically when using about a page (around 3000 characters) as input, prior to which the quality is quite poor. Seemingly, shorter texts are not providing enough data for the text analyzing algorithms to perform on.

As a follow up, we analyzed the statistics of query lengths used in our pilot programs. The interesting finding was the average input was only 500 characters long, even though the users were asked to use at least one page. Maybe this was just because they wanted to test how good the system is by throwing in a short phrase. Or maybe searchers are so used to tailor slimmer search strategies that by habit they reduce the text volume. Then using larger text input, it feels uncomfortable from a user perspective. Thoughts on



this are welcome!

In a worst case we have a "lose-lose" scenario when exploring text-based search engines. The searcher is spending time on analyzing and reducing a text to what is thought to be a better and more query like input. But in the end the result might be too short for the algorithm to deliver on and the effort degrades the performance.

This is rather easily remedied by thinking in terms of larger paragraphs instead traditional keyword analysis. The challenge is not to mix up the manual search approach with the algorithmic

one. For the latter, as indicated by the graph, you should as a starting point just use whatever descriptive text you have as input, even though long and seemingly too broad. The results are then used to boost and support analysis and further manual search strategies.

In an additional follow up to verify the findings, we compared the performance of the lesser performing technical domain “Mechanical elements” with the high performing “Biotechnology” with respect to the input. Thus, looking at a title/abstract type query compared to a full text based one. Looking at the baseline measurements (i.e. how many of the relevant citations found by the patent office that were found by the tool) the recall ratio when looking at top 100 hits given by the following figure.

In short, using a full text document as input improves recall for the mechanical domain considerably, in average boosting performance with around 50%.

However, for biotechnology we got only a boost of about 10% and *only* when also capping the length of the input text to include a maximum of 100000 characters (the peak value for best recall value shown in the previous graph).

The optimal input length is also somewhat dependent on the technical domain, if the technical field is better or less suited for using text-based queries.

This suggest a need to run a baseline on your specific patent portfolio or at least within the technical domain reflecting your research. This helps to have a better understanding of best use and to be able to optimize the workflow and need of information for improving the performance.

So then back to the initial question; may I do something as user to improve the results? The answer is yes. It is still valid that a better input always yields better output, if the text used is large enough for the algorithm to analyze properly. The grey bar "hybrid" in the diagram shows indeed a further boost of the recall when a user manually selects the most appropriate text sections and use this assembly of text as input (about 1-5 pages together).

This "hybrid" study consists of spot-checks on a limited number of about 100 patent documents. It indicates that the more challenging technical fields are still improved more than the already good one.

This was, as you might guess, especially true when the text was noisy, had too much background focus or where it contained inconsistencies or dealing with numerous completely different technical ideas.

Proving the query text with appropriate patent IPC/CPC classes to steer the analysis against, yield even a further performance boost. This indicates that a professional user indeed may make use of traditional manual search skills also in an AI text-based search procedure.

Then it is all about setting up a proper process for when and how to perform a prior art search and with what tools. We have one trade off identified between i) spending more time initially and improving the query by selecting text paragraphs to get better accuracy of the first output and ii) use the text as is with potentially lower accuracy and spend more time on manual search afterwards based on the results retrieved.

One consideration could then be to use it more as a rough pre-screening to get a first indication of the prior art or as a

more integrated search tool within the manual search process. I believe, this is not "one" or "the other", rather "both" and probably more depending on where in the innovation workflow it is being used and for what purpose.

Summarizing, based on our extensive analysis on input format versus output performance we discovered a few hands-on insights which should be part of best practice for text-based searches:

Avoid using too short query texts.

We discovered that test users used an average length input text of 500 characters, while the threshold for performance boost starts at around 3000 characters. This could relate to the fact that query-based searching searchers are used to create shorter search strings, which indeed contain limited text data.

The impact of small text input sets is that the algorithms have little information to "work on" with a poorer result than it potentially could have achieved. This is an important insight and therefore a good aim is to enter much more text than compared to a search query; a recommended minimum is at least one page.

For exceptionally long texts, and especially in high performing domains, you could consider cutting out data, e.g. use only the first 1-5 pages or selected paragraphs.

Adapt text length and strategy to the technical field.

There is a correlation between the technical field and the need for longer descriptive texts. The harder it is to describe the subject area in a text, the more text you should add.

Thus, areas like the "Mechanical elements "could increase the recall considerably by feeding the whole document with both descriptive passages and even background to support better search results. However, do avoid too large text documents as well. Keep in mind that it may make sense to also manually select and combine appropriate best sections to use as input, which especially applies to exhaustive and inconsistent documents.

You may also consider adding patent classes for fields with lower recall if the tool offers that option.

Explain short topics with more details.

As in real life, the more comprehensive and concise you are, the better you will be understood by the other. Algorithms cannot guess your underlying intentions, thus always avoid general concepts, especially when working on short query texts of a few sentences.

If you want to use a tool for ideation and inspiration to find related prior art, do try to almost over explain the topic. By adding details about the problem /solution when describing an idea, it will improve the performance. One comment I got from a user was a very to the point instruction; "Explain the new case as detailed as if you talked to your teenager, then it works best".

Always run a baseline on your technical domain.

You need hundreds of queries per class for statistic relevance of the performance as discussed in part 1. A manual verification of a few cases is merely

indicative. Thus, suggest using your own (or combined with a competitor) patent portfolio as basis for running a performance test showing the recall scores.

By knowing how a text-based search tool is performing in your technical domain, you get information for optimizing your prior art searching as well. The analysis may also help to improve designing invention disclosures (e.g. amount of text, format for summary, problem and/or solution, potential key terms etc.) to boost the search performance further.

The bullet points above are only an indicative summary and we are running further studies to explore how a user may get better performance. Also, as the findings disclosed in this article are based on IPscreener only. Of course, the optimal scenario would be to compare several tools run on the same data to understand differences and similarities.

Optimal text input versus quality of output will certainly vary among the providers. I look forward to a common standard soon! I also reckon there are more tips & tricks on best practice from other studies as well. You are very welcome to share comments to gather further experience on this topic.

In the next chapter I will share my thoughts on when to use AI tools and the future of searching; look out for "The future role of AI enhanced searchers".

[1] Subject analysis and search strategies – Has the searcher become the bottleneck in the search process? Nijhof; World Patent Information, Volume 29, Issue 1, 2007

[2] Searching? Or actually trying to find something? – The comforts of searching versus the challenges of finding: Nijhof; World Patent Information, Volume 33, Issue 4, 2011

[3] Want to find? Break the rules! Nijhof; World Patent Information, Volume 52, 2018