

iLexIR Advanced Search Engine

This document is designed as a gentle introduction to our search technology and current prototype available for testing through our website. A more technical published paper and a video demonstrating use of the system can also be found from our website.

Search engines use software programs that analyse a set of documents (for example, webpages) to provide information for each document that the SE can then use to allow us to search for information within the documents that we want to find. For example, large search engines analyse the webpages within the www to allow users to search the www using *keywords*. A keyword can be any word, so it follows that these search engines extract and store the set of words from each webpage to allow us to search in this way.

Extracting information from the www is already a large-scale task given the number of webpages. However, if we have a smaller set of documents (like a set of PDF documents from scientific journals covering a specific field), we can provide more sophisticated search methods as it is feasible to extract more complex information required by the search engine to support more precise and advanced search.

We are actively developing generic technology within our search engine, and the corresponding software that analyses collections of PDF documents, that allows the user to build up complex queries based on linguistic and structural information (identifying for example headings, captions and reference sections of a journal paper). Our system provides sentence-by-sentence grammatical analyses of the text, resolves reference, etc, which allows us to then search using these linguistic features.

For example, if you were to search for “black dog” in a keyword-based search engine, you would find all the documents containing the words “black” and “dog” but not necessarily examples where a *dog* was referred to as *black*. However, our system can be used to find such examples by allowing the user to intuitively refine searches without having to understand the exact linguistic relationship between the words “black” and “dog” in a sentence like “The black dog chased the cat” or “The dog is black”.

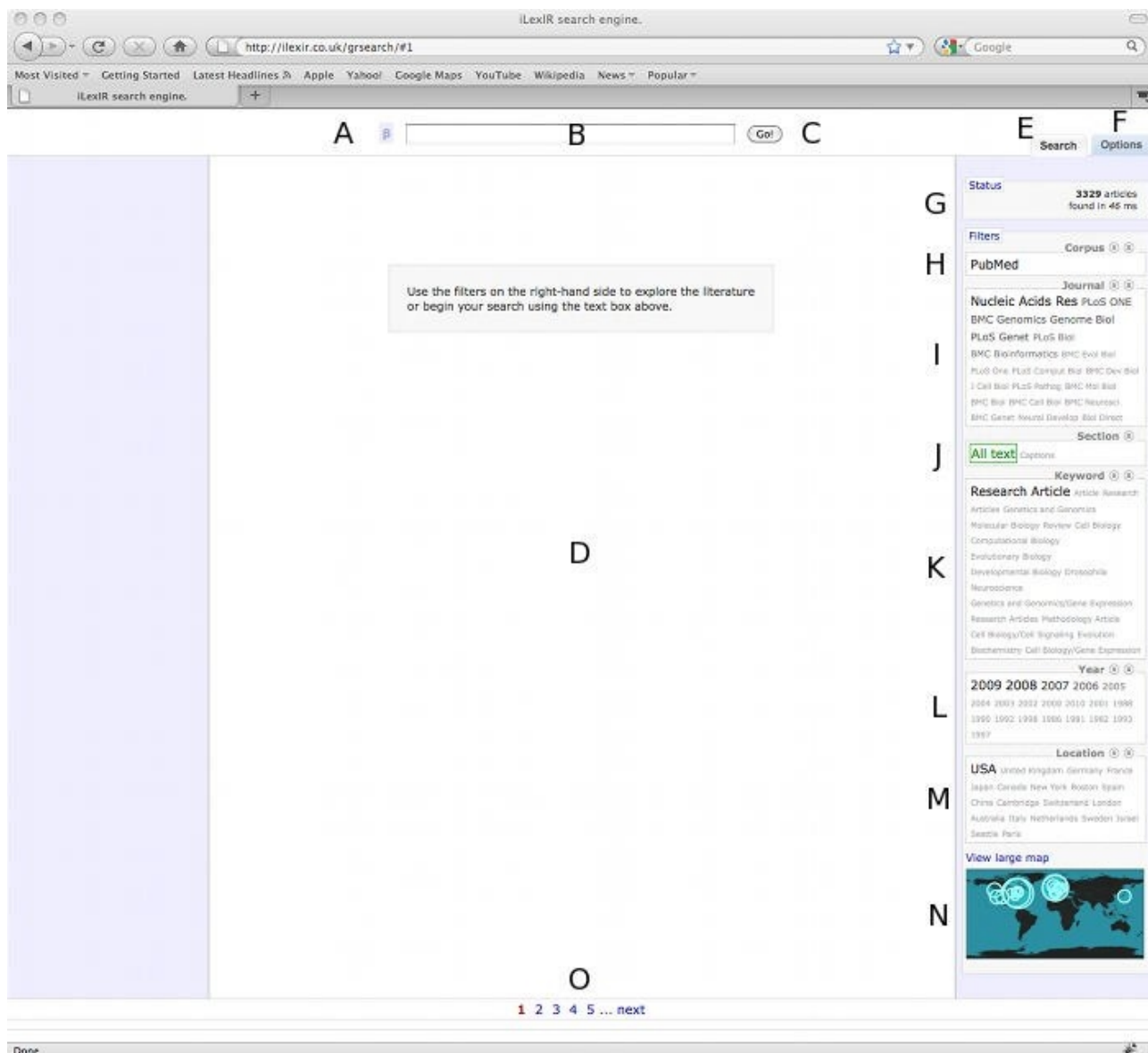
This is a simple example, but our system is capable of much more complex analysis. For example, we can easily build up a complex search to find all sentences where a black dog chased *something*, where our system might find 'cat', 'rat' and 'a boy called danny'! This kind of *information extraction or fine-grained information retrieval* can be important for complex longer documents where, for example, we might want to find or read more about the set of genes that are inhibited by a particular protein.

Our current demo (available through our website) illustrates an example application of our technology applied to a small set of PubMed scientific papers within the field of genetics. However our technology is applicable, in principle, to any complex document collection such as those on a company Intranet or linked to from web pages but not necessarily effectively-indexed by a standard keyword-based search engine. The demo only illustrates a subset of the features that we have incorporated into deployed versions of the system. For example, we have also integrated image similarity search over diagrams, photographs, etc with search within captions, so that we can support searches like “find all the gene names in captions linked to photographs of fruit fly wings”.

The rest of this document will walk you through the type of searches and functionality available in our prototype. This prototype is designed to demonstrate generic technology which is not confined to use on scientific papers or PDF documents, and which could easily be extended to apply to different document types with different structure (including webpages, emails, product reviews, etc). Please contact us if you are interested in hearing about how this search engine could be applied to your documents or unstructured or semi-structured textual data, possibly also containing visual information.

Introduction:

Our demo system will initially load and you should see the following website:



We have labelled this figure to illustrate the main sections of the page, these letters are used in explanations that follow:

- (A) utf-8 symbol selection: allows users to enter in characters not available on a keyboard (e.g. β)
- (B) text search box: enter keyword-based search terms here.

- (C) text search button (labelled “Go!”): performs keyword-based search using words entered in (B).
- (D) result window
- (E) Search tab: contains information regarding the current search/result set.
- (F) Options tab: used to provide more advanced options (if any apply, please disregard in our current demo).
- (G) Status window: illustrates the status of the current search/result set (shows the number of articles found and search time).
- (H – N) Filters section: this contains a information and search filters available for the current search/result set (each will be described in more detail below).
- (O) Search pages links: links to the different search result pages.

Most people will be familiar with the search text box (B) and the search button (C) which can be used to enter in words to search for, as well as the results area (D) that displays these results and (O) that shows links to the result pages. These elements are present in some form in all the major search engines and here we assume that the reader is familiar with these elements. These elements provide *keyword-based* searches, and this simple search is also available in our demo.

Filtering Search Results:

Before we discuss how to search using our system, we will show you how to *filter* results. This is a kind of search in itself, but it effectively refines any search e.g. if we search for papers with the term “ATP”, then want to filter these results to find the papers published in 2010.

You can also perform filtering over the entire set of articles in our test set without doing any kind of search which we will discuss now e.g. all papers published in 2010.

Our demo is built over a small set of 3329 articles from PubMed (scientific articles). This number can be seen in the status window (G) when you first start the demo. (*Note: our system is scalable and can be applied to large document collections as we apply distributed search techniques*).

In the filter section (H-N), there are a number of windows, each of which is a *tag cloud* illustrating the content of the set of results (initially all results) by showing values that occur more frequently in larger font. You can see the most common journal type in the journal filter window (I) is the “Nucleic Acid Res” journal:

If you click on “Nucleic Acids Res” in the journal filter window (I), then we'll search to find the articles that occur for this journal type.

You will see the number of results shown in the status window (G) will be 682 and the results window (D) shows the articles for this journal type. You can also see that in the journal filter window (I) the journal value is highlighted in green, illustrating that this filter value is being applied.

The other filter windows will also have updated, showing the counts for the current result set i.e. the values for these other filters in the 682 articles from the “Nucleic Acids Res” journal.

If you want to remove a filter, then simply click on filter you are applying (those

highlighted in green) and it will be removed and your search results will update. You can also press the small “x” button in the top right corner of any filter window to remove any filters applied in that filter window.

Every time you perform a search (e.g. a keyword-based search) the filter windows will update showing the counts for the new result set.

Note: the filters were selected to correspond to the information present in the test set which are scientific articles (journal type, mesh terms, year of publication etc) but we can easily extend or modify our system to store and apply any number of filters as we have developed generic software that requires very little customisation to port to new datasets. For example, as we perform structural analysis of our documents, this structure can be used to refine searches. In the demo we provide the choice of searching within “All text” and “Captions” as seen in the section filter window (J) but this could be extended to any and all document structures.

Map filter:

The map filter window (N) illustrates the locations of the authors of the papers, where the size of the circle represents the frequency of these location values within the result set.

If you wish to apply this filter, you need to click on the “View large map” button above the map filter window (N), this will bring up a large version of the map within the results window (D). You can then roll-over the map to view the different locations, and click on the circles to apply the location as a filter.

The same locations can be applied within the location filter window (M).

To hide the map again, click on the “Hide large map” button.

Multiple filter values:

Previously, we only selected one filter value for journal type, which removed the other journal values from the journal filter window (I), but we can also select multiple values from the same filter window. This will find all results that are from *any* of the journal types selected.

First click on the small “x” button (if you haven't already) in the journal filter window (I) to clear the previous filter applied (and show all filter values available). Now click on the small “s” button in the top right corner of the filter, which makes the filter window “stick” so you can apply more than one filter value. Now if you click on multiple journal values you'll see that each time the results and other filter windows update as they did previously. You can see the total number of results in the status window (G) along with the time it took for our search server to find these results.

Every time you click on a filter, a new search will be performed with the new filter set and you can apply any number of the different filters, from any of the different filter windows, at anytime while you are searching.

Article-level information:

In the results, you can see a small “+” sign under each article's title, if you click on this then the article abstract, figures, tables and other related documents can be viewed.

You can click on an article's images (not the pointing hand symbol which we will discuss later) or any other document type to view the image or open the document.

Keyword-based Search:

For example, if we perform a keyword-based search by entering “mRNA express” in the search text box (B) then pressing the search button (C) you can see 2413 sentences in 771 articles are found.

Note that while previously we found articles, we now are finding the number of sentences that contain these terms. That is, our searches are generally performed on a sentence-by-sentence level. The number of articles reported is the total number of unique articles across all the sentences that are results.

Each result in the results window (D) is a sentence and are ranked in descending order of how well our system thinks the sentence matches a search. We also group sentences if they appear in the same article, and show the article-level information for each group of sentences. If a sentence was part of a caption, then we will show the corresponding image thumbnail or document symbol which can be viewed by clicking on it.

The results show the keyword-based terms highlighted in either blue, green or orange. We highlight a keyword in blue if the word has no relationship with any of the other keywords. If a keyword is highlighted green then this illustrates that a grammatical relationship exists *between* two of the keywords searched while orange illustrates a grammatical relationship *between* three or more keywords. This helps the user to try and identify the kinds of results they are looking for, and is used to refine the searches as discussed below.

Wildcard searches:

As well as simple keyword-based searches, we can use wildcard characters “*” (any number of characters) and “?” (a single character) to search for a range of different keywords that match the regular expression. This kind of search is referred to as a 'wildcard keyword search'.

For example, searching for “*RNA express” will find sentences that have “mRNA”, “miRNA” etc. That is, the pattern “*RNA” finds all words where any number of characters are followed by “RNA”. You can see this matches a lot more results - 7424 sentences in 1261 articles.

Advanced GR-based Search:

Once we have performed keyword-based search, we can intuitively refine the results to the kinds of sentences we are looking for. This process always starts with a keyword-based search. From the keyword-based search results the user can find an example sentence where keywords relate to each other the way they are trying to find. Our advanced search then effectively searches for *similar* sentences to this example sentence. This requires the user to simply use their underlying linguistic understanding

of their own language and doesn't require any theoretical linguistic knowledge.

For example, we previously searched for “mRNA express”, where for arguments sake we really wanted to find instances like that in the second sentence of the keyword search results: “Figure 3 Correlation between mRNA expression and decay rate...”. The search terms are highlighted in green, confirming that the system has found this relationship correctly, so we'll now use our system to find more examples of this relationship.

To do this, we will use the “Advanced search” box that appears in the bottom right of the demo once you have performed a keyword-based search.

First perform the keyword-based search “mRNA express” which has 2413 results found. From the result window (D), we see that the second result has the relationship we're looking for.

We will first clear the “Advanced search” area by clicking on the “X” button in the top-right corner of the window.

Now, if you click once on the words “mRNA” and “expression” in the second result and you will see that “mRNA” and “expression” will appear in the advanced search box with a purple link illustrating that a grammatical relationship occurs between them.

Click on the pointing hand symbol in the top-right hand corner of the advanced search window in order to search for this relationship which will show 362 sentences (vs. 2413).

This simple example illustrates that the system can be used to quickly and intuitively refine keyword-based search results to perform complex linguistic searches without the user understanding the actual grammatical relationship between terms. Rather than looking through 2413 results for these 362 results, we have quickly and easily found the set of sentences which contain the information we really wanted. We can continue to refine the search using filters or adding further keywords and relationships to the advanced search window.

Wildcard searches:

We can perform advanced wildcard searches as well. Continuing the previous example, we will find instances where the verb express is applied to words ending with “RNA”.

Within our advanced search window roll-over the word “mRNA” and click on the small “+” sign in the top left-hand corner. In the pop-up window type “*RNA” in the “Search for” text box and click “OK” (or press return key). You will now see the advanced search window shows a relationship between “*RNA” and “expression”.

Click on the pointing hand symbol in the top-right hand corner of the advanced search window in order to search for this relationship which will show 1023 sentences (vs. 7424).

Advanced search window:

For each word shown in this window if you roll-over the term four options arise (one in each corner of the square surrounding the word).

- “x” (top right) – removes the word
- “-” (top left) - lets you edit the word (to add wildcard characters “*” or “?” or replace words with another one)

- “+” (bottom right) – a short-cut to change the word to “*”
- “>” (bottom left) – a short-cut to change the word to it's original value.

For each relationship, a purple link is shown between words in the advanced search window. This link illustrates a particular kind of grammatical relation, but if you want to create a more generic search, you can indicate that **any** kind of relationship can exist between the words by clicking on the link (turning it pink). (You can toggle between exact relationship (purple) and any relationship (pink) by clicking on the link.)

For example you may want to search for information regarding *what* is expressed by *RNA using the advanced search: “*RNA -- expression -- of -- *” (where – indicates a link in the advanced search window).

[You can try and do this yourself by searching for “mRNA express of” in the keyword search then selecting “mRNA expression of samples” in the third result. Then refine the keywords the advanced search window as required. You should find 3 sentences (as this is only a small test set).]

Similar Image Search:

We have integrated similar image search into our system, which uses a mixture of content-based image similarity scoring techniques and our advanced search similarity scoring (similarity based on the caption text and relationships between the words in the caption) to determine a set of similar images (and captions) to one the user is interested in.

This works independently of any prior search you may have performed as our system analyses the image selected and it's corresponding caption text in order to perform the search. This is a new feature which we plan to develop further in the future so that similar image search is integrated within our search refinement process. That is, similar image search will be added to our search criteria (i.e. to those defined in the advanced search window).

This similar image search is easily accessed by simply clicking on the small pointing hand symbol in the top right corner of an image thumbnail. Note: This symbol only appears on an image thumbnail if similar image search is available for the image (and will not appear for example on the .xls or .doc type files for a article).