02_11_2022

Initial search results

Using the query `"topic label*" OR ("topic model*" AND "label*")` in the time period *2017-2022* yielded 388 conference papers and 549 journal papers, for a total of **937** papers.
The same query in the time period **2020-2022** yielded 178 conference papers and 332 journal papers, for a total of **510** papers

Further filtering using proximity operators (2017-2022)

Given the large amount of papers related to the query spanning the time period 2017-2022, a further filtering step is proposed on the initially gathered papers by imposing a proximity constraint between the root terms and `label*` and `topic*`.

This is possible because imposing the proximity constraint `"label*" NEAR "topic*"` on the papers gathered from the query
`"topic label*" OR ("topic model*" AND "label*")`
will lead to the same set of results that would be obtained by directly executing the query
`"topic label*" OR ("topic model*" AND ("label*" NEAR "topic*"))`
on the selected repositories.
In fact, the query containing the proximity operator is simply a stricter version of the one used to gather the initial set of papers.

Since proximity operators (e.g. `NEAR`) are not supported by most the chosen repositories (with the exception of SpringerLink), it was decided to first gather the initial set of papers using the more relaxed query `"topic label*" OR ("topic model*" AND "label*")` and then to further filter them locally by imposing the proposed proximity constraint .
The filtering on local files has been performed on a tool called “FoxTrot Professional Search” using, in the indexed folders, the following query:
`[{20} "topic*" "label*"]`
Note that `{20}` indicates the value of the proximity constraint (i.e. The maximum distance between the two root terms).
For readability purposes, for the rest of the chapter the syntax `"label*" NEAR "topic*"` will be used instead in order to indicate the query used within the FoxTrot search tool.

In this context, all the papers gathered from the first part of the query (`”topic label*”`) are kept, since the two root terms always appear in direct proximity to one another.
On the other hand, for the second part of the original query (`"topic model*" AND`

`"label*"`) only those papers meeting the newly imposed proximity constraint are kept. In other words, the initially gathered papers mentioning topic model(ing) and label(s) are kept only if the proximity constraint between the two root terms (label and topic) is met.

Formally, if we define:

- P as the set of papers contained in the selected venues for the chosen time period

And:

- Q_0 as the query `"topic*" NEAR "label*"`
- Q_1 as the query `"topic label*" OR ("topic model*" AND "label*")`
- Q_2 as the query `"topic label*" OR ("topic model*" AND ("label*" NEAR "topic*"))`

Then:

- $Q_1(P)$ represents the set of results obtained by executing Q_1 on P.
- $Q_2(P)$ represents the set of results obtained by executing Q_2 on P.

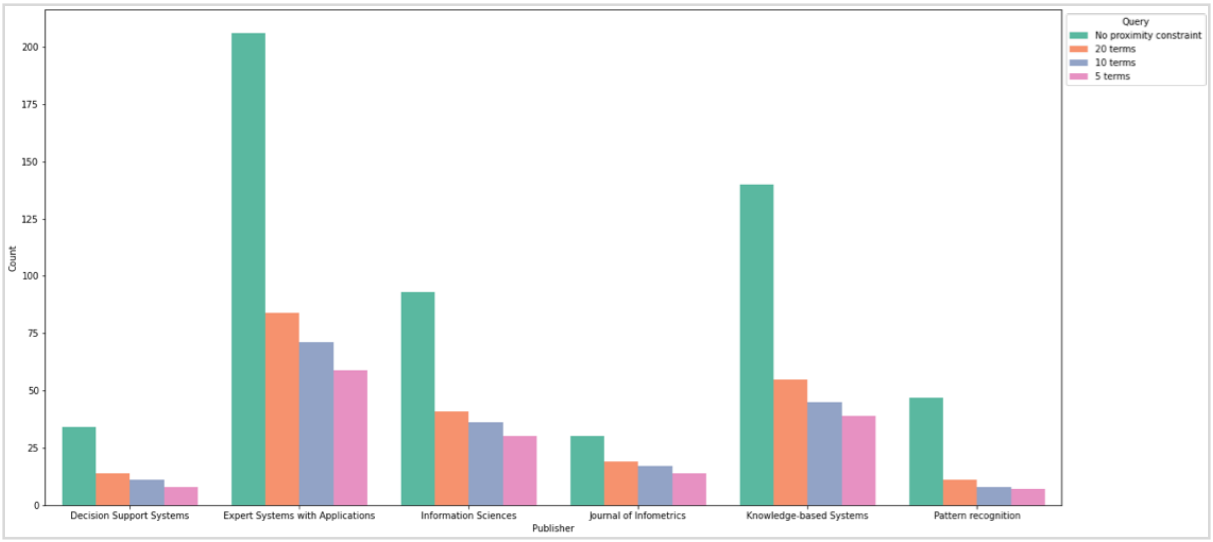
In this context, we always have that: $Q_2(P) \subseteq Q_1(P)$

Additionally, we also have that: $Q_0(Q_1(P)) \equiv Q_2(P)$

Filtering the initial set of papers by setting the proximity constraint at 20, 10 and 5 terms leads to the following results:

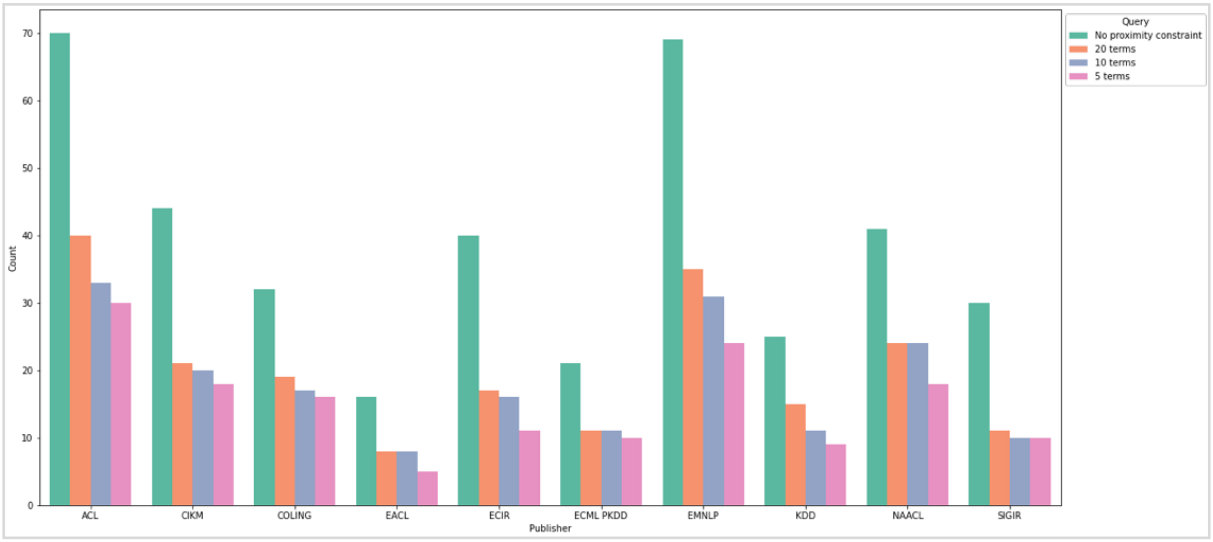
Journals

- Decision Support Systems: 34, 14, 11, 8
- Expert Systems with Applications: 206, 84, 71, 59
- Information Sciences: 93, 41, 36, 30
- Journal of Infometrics: 30, 19, 17, 14
- Knowledge-based Systems: 140, 54, 44, 38
- Pattern recognition: 47, 11, 8, 7

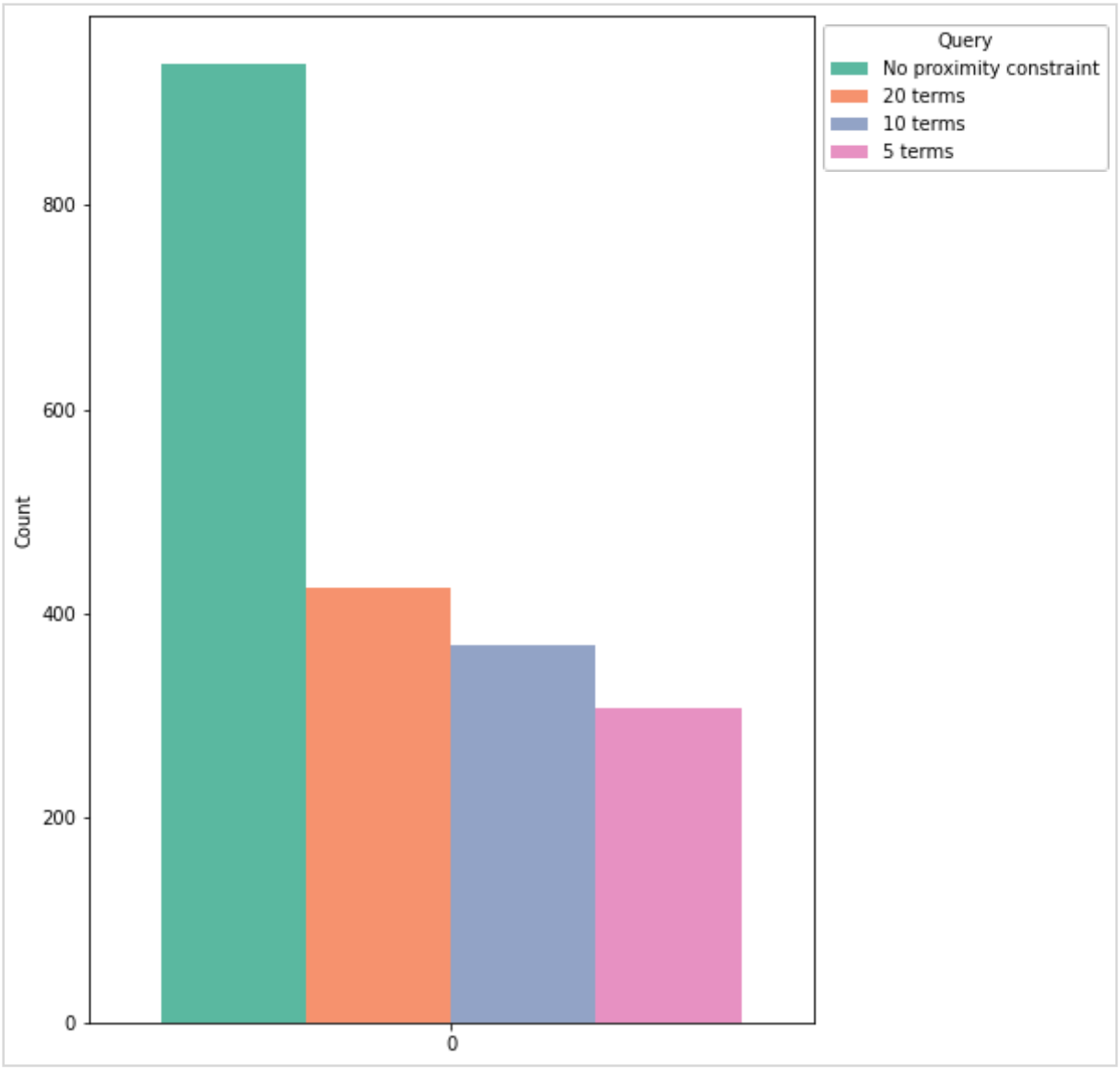


Conferences

- ACL: 70, 40, 33, 30
- CIKM: 44, 21, 20, 18
- COLING: 32, 19, 17, 16
- EACL: 16, 8, 8, 5
- ECIR: 40, 17, 16, 11
- ECML PKDD: 21, 11, 11, 10
- KDD: 25, 15, 11, 9
- EMNLP: 69, 35, 31, 24
- NAACL: 41, 24, 24, 18
- SIGIR: 30, 11, 10, 10



New Totals: 938, 424, 368, 307



Selection process (Inclusion/Exclusion criteria) (Cont.)

In order to be selected for this review, a paper should propose or actively apply, either with a primary or secondary focus a topic labeling technique.

Papers appearing in the selected research do not necessarily need to describe the implementation of a novel labeling approach, but it is important that they do not meet any of the following **exclusion criteria**:

- The paper does not actively apply any topic labeling techniques
- The choice of labels and the label assignment procedure are not clearly described and justified
  - (e.g. labels are assigned manually and the details of the assignment process are not described)
- All the described labeling approaches are taken from existing work and re-proposed as-is (on the same corpus and set of topics)
- The paper and/or the analysed corpus do not match the imposed language restrictions

- The paper is a systematic review (secondary/tertiary study)

Reference extraction

Extracting references from the final selection of papers is useful for both snowballing and for creating paper networks with tools such as Pajek.
In order to automate the reference extraction process the anystyle CLI tool is utilised.

```
$ anystyle -f bib find main.pdf
# returns BibTeX formatted list of all the references in the paper
```

CLI command details

#Thesis/Temporary notes#