# Computational Communication Science Literature - A Brief Quantitative Literature Research

Philipp Knoepfle

2022-03-15

**Summary:** This R Markdown Notebook serves as a quick presentation platform for my experiment with the LitsearchR in the Meta Rep Project. Our sample consist of 371 selected journal articles in the field of Computational Communication Science from 2010-2021. The goal of this brief analysis is to get a better understanding of potential keywords for the manual selection of articles in the second stage of our project as well as to complement our qualitative (and manual) literature research results with an automated literature analysis. As a major tool for this, we employ the LitsearchR package.

## The data

The LitsearchR package requires as a data input a Bibtex-file with author, title, journal, abstract, keyword, reference, etc. information. The more units of reference and the more comprehensive this information in general is, the better will the analysis be. To identify a first set of Computational Communication Science (CCS) journal articles, I perform an advanced reference search via the Web of Science (WOS) research interface. Our search parameters are:

- **Search term:** computational,
- **WOS Category:** Communication (Core Collection database),
- **Time period:** 2010-2021,
- **Document Type:** Journal Article.

Our first search yields 597 journal articles. After a brief manual inspection of this list I noticed that there are still some references which are classified as "journal articles" even though they are clearly contributions in a large handbook. Moreover, some journals are classified as "communication" even though they are not necessarily Communication Science contributions but can be more precisely prescribed to the field of Telecommunication or Information Theory. Even automated literature research can be really messy and always needs manual proofreading of some sort. Since these articles are not interesting for our analysis, I manually exclude them to arrive at a final sample size of 371 articles. This bibtex-file containing the data set for our analysis can be found in my Github repo for this project Link.

## LitsearchR analysis

Note: The LitsearchR package has to be downloaded via devtools and Github since it is not on CRAN yet.

```r
# Load packages, setwd and import the data set
library("igraph")
library("ggraph")
library("litsearchr")
library("dplyr")
library("wordcloud")
library("tibble")

setwd("H:\\Meta-Rep\\R Code")
naive_results <- import_results(file="savedrecs.bib")
```

First, let us see what our data set looks like. It contains 371 CCS references, i.e. journal articles, and has 42 different categories.

```
dim(naive_results)
```

```
## [1] 371  42
```

Our bib-file categories/columns include: type, title, abstract, keywords, the International Standard Serial Number (ISSN), WOS Core Collection category, cited references, and many others. A lot of very useful information for the next steps of our analysis.

```
colnames(naive_results)
```

```
##  [1] "type"                      "author"
##  [3] "title"                     "journal"
##  [5] "abstract"                  "publisher"
##  [7] "address"                   "language"
##  [9] "affiliation"               "doi"
## [11] "earlyaccessdate"           "issn"
## [13] "eissn"                     "keywords"
## [15] "keywords_plus"             "research_areas"
## [17] "web_of_science_categories" "author_email"
## [19] "researcherid_numbers"      "orcid_numbers"
## [21] "cited_references"          "number_of_cited_references"
## [23] "times_cited"               "usage_count_last_180_days"
## [25] "usage_count_since_2013"    "journal_iso"
## [27] "doc_delivery_number"       "web_of_science_index"
## [29] "unique_id"                 "da"
## [31] "article_number"            "funding_acknowledgement"
## [33] "funding_text"              "year"
## [35] "volume"                    "number"
## [37] "pages"                     "month"
## [39] "oa"                        "ci"
## [41] "note"                      "organization"
```

Now let us dive into the actual analysis and look at the keywords in our current data set. Unfortunately, 52 articles in our data set do not have keywords but that should not be a problem since we can easily infer keywords from the article abstract and title. The `method` argument specifies the type of keyword extraction method which is used to obtain the keywords. “Tagged” gives us simply the author-tagged keywords from the bib-file, whereas “Fakerake” is a quick implementation of the Rapid Automatic Keyword Extraction (RAKE) algorithm, which is a domain-independent keyword extraction algorithm, see link. The idea behind this is, that if a keyword co-occurs often, it’s information value in a literature search is high which qualifies it as an important keyword in our literature search.

Note: However, it is important to note, that keywords which do not co-occur often are crucial to identify meaningful studies in our analysis and still have to be taken into consideration. Even studies which are not well connected to other literature can be meaningful and relevant to our analysis.

```
sum(is.na(naive_results[, "keywords"]))
```

```
## [1] 52
```

```
keywords <- extract_terms(keywords=naive_results[, "keywords"], method="tagged")
```

```
## Lade nötigen Namensraum: stopwords
```

```
Keywords_abstr <- extract_terms(text=naive_results[, "abstract"], method="fakerake", min_freq=3, min_n=

Keywords_title <- extract_terms(text=naive_results[, "title"], method="fakerake", min_freq=3, min_n=2)
```

A first look at the author-tagged keywords reveals a lot of expected CCS keywords. More interestingly though, is the sheer quantity and variety of keywords. Of course, there is a large variety in the type of study and methodology, yet the inconsistency in keywords is surprising.

```
extract_terms(keywords=naive_results[, "keywords"], method="tagged")
```

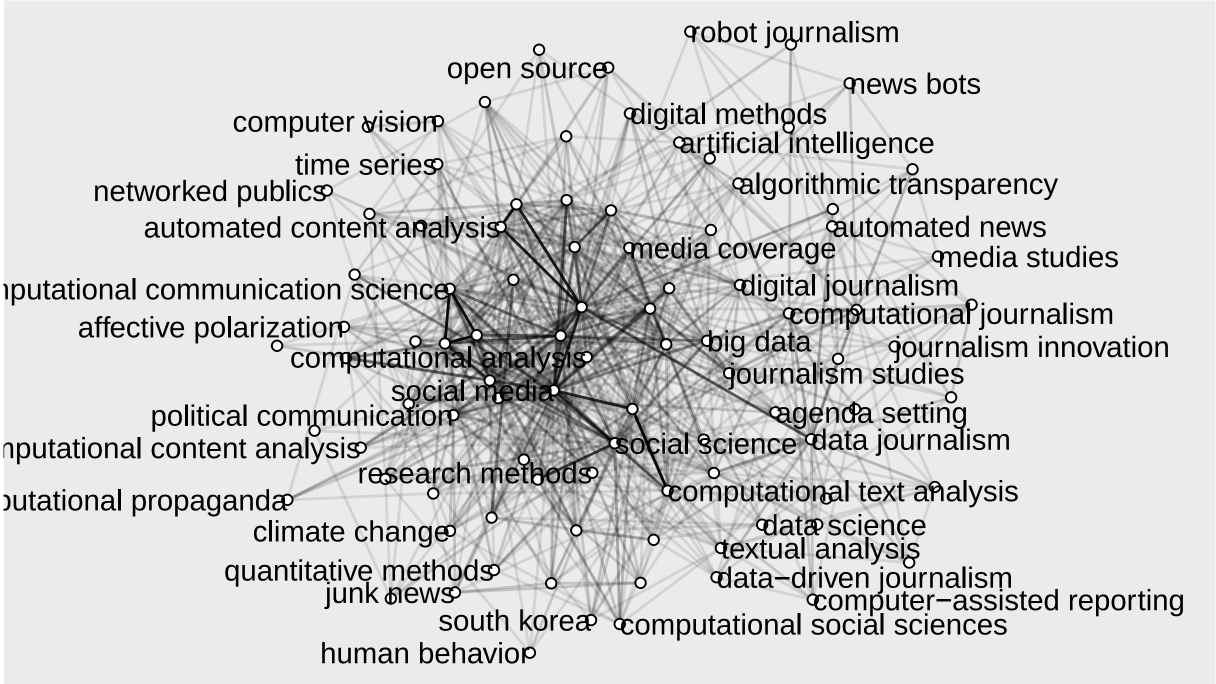
```
##   [1] "actor-network theory"
##   [2] "affective polarization"
##   [3] "agenda setting"
##   [4] "algorithmic culture"
##   [5] "algorithmic journalism"
##   [6] "algorithmic transparency"
##   [7] "artificial intelligence"
##   [8] "automated content analysis"
##   [9] "automated journalism"
##  [10] "automated news"
##  [11] "big data"
##  [12] "climate change"
##  [13] "communication theory"
##  [14] "computational analysis"
##  [15] "computational communication research"
##  [16] "computational communication science"
##  [17] "computational content analysis"
##  [18] "computational journalism"
##  [19] "computational linguistics"
##  [20] "computational methods"
##  [21] "computational narrative"
##  [22] "computational propaganda"
##  [23] "computational social science"
##  [24] "computational social sciences"
##  [25] "computational text analysis"
##  [26] "computational thinking"
##  [27] "computer-assisted reporting"
##  [28] "computer vision"
##  [29] "connective action"
##  [30] "content analysis"
##  [31] "data-driven journalism"
##  [32] "data collection"
##  [33] "data journalism"
## [34] "data journalism awards"
## [35] "data science"
## [36] "data visualisation"
## [37] "data visualization"
## [38] "digital cartography"
## [39] "digital communication"
## [40] "digital ethnography"
## [41] "digital humanities"
## [42] "digital journalism"
## [43] "digital media"
## [44] "digital methods"
## [45] "digital news"
## [46] "digital platforms"
## [47] "digital sociology"
## [48] "digital trace data"
## [49] "election prediction"
## [50] "expectancy violation theory"
## [51] "experimental methods"
## [52] "facial expressions"
## [53] "fake news"
## [54] "field theory"
## [55] "filter bubble"
## [56] "gulf crisis"
## [57] "health communication"
## [58] "information warfare"
## [59] "intermedia agenda setting"
## [60] "international broadcasting"
## [61] "internet studies"
## [62] "journalism education"
## [63] "journalism ethics"
## [64] "journalism innovation"
## [65] "journalism practice"
## [66] "journalism studies"
## [67] "journalistic field"
## [68] "junk news"
## [69] "machine learning"
## [70] "main model"
## [71] "media bias"
## [72] "media economics"
## [73] "media effects"
## [74] "media studies"
## [75] "mixed methods"
## [76] "natural language generation"
## [77] "network agenda setting"
## [78] "network analysis"
## [79] "networked publics"
## [80] "news automation"
## [81] "news bots"
## [82] "news consumption"
## [83] "news diffusion"
## [84] "news evaluation"
## [85] "news flows"
## [86] "news media"
## [87] "news production"
##  [88] "news sources"
##  [89] "news values"
##  [90] "newsroom management"
##  [91] "online experiments"
##  [92] "online journalism"
##  [93] "online news"
##  [94] "open data"
##  [95] "open science"
##  [96] "open source"
##  [97] "participatory media"
##  [98] "platform governance"
##  [99] "political campaigns"
## [100] "political communication"
## [101] "political economy"
## [102] "political journalism"
## [103] "political news"
## [104] "political polarization"
## [105] "practice theory"
## [106] "public diplomacy"
## [107] "public opinion"
## [108] "public service media"
## [109] "public sphere"
## [110] "qualitative interviews"
## [111] "quantitative analysis"
## [112] "quantitative methods"
## [113] "research methods"
## [114] "robot journalism"
## [115] "selective exposure"
## [116] "sentiment analysis"
## [117] "social bots"
## [118] "social media"
## [119] "social media analytics"
## [120] "social movements"
## [121] "social network analysis"
## [122] "social science"
## [123] "sociology of news"
## [124] "software development"
## [125] "south korea"
## [126] "structured events"
## [127] "structured journalism"
## [128] "structured narratives"
## [129] "supervised machine learning"
## [130] "technology studies"
## [131] "text analysis"
## [132] "text mining"
## [133] "time series"
## [134] "topic modeling"
## [135] "user-generated content"
```

Aside from this, we can combine the author-tagged keywords and the “Fakerake”-derived keywords to get a comprehensive list of keywords and create a simple co-occurrence network in the form of a weighted graph based on the co-occurrence frequency of our comprehensive keyword list and the abstract and title of each paper. In simple terms, we look how often a keyword co-occurs with other keywords in the abstract and title. If they co-occur frequently, they are rated closer and vice-versa.

```r
terms <- unique(c(keywords, Keywords_title))

docs <- paste(naive_results[, "title"], naive_results[, "abstract"])

dfm <- create_dfm(elements=docs, features=terms)

g <- create_network(dfm, min_studies=3)
```

We can plot our results in a nice graph.

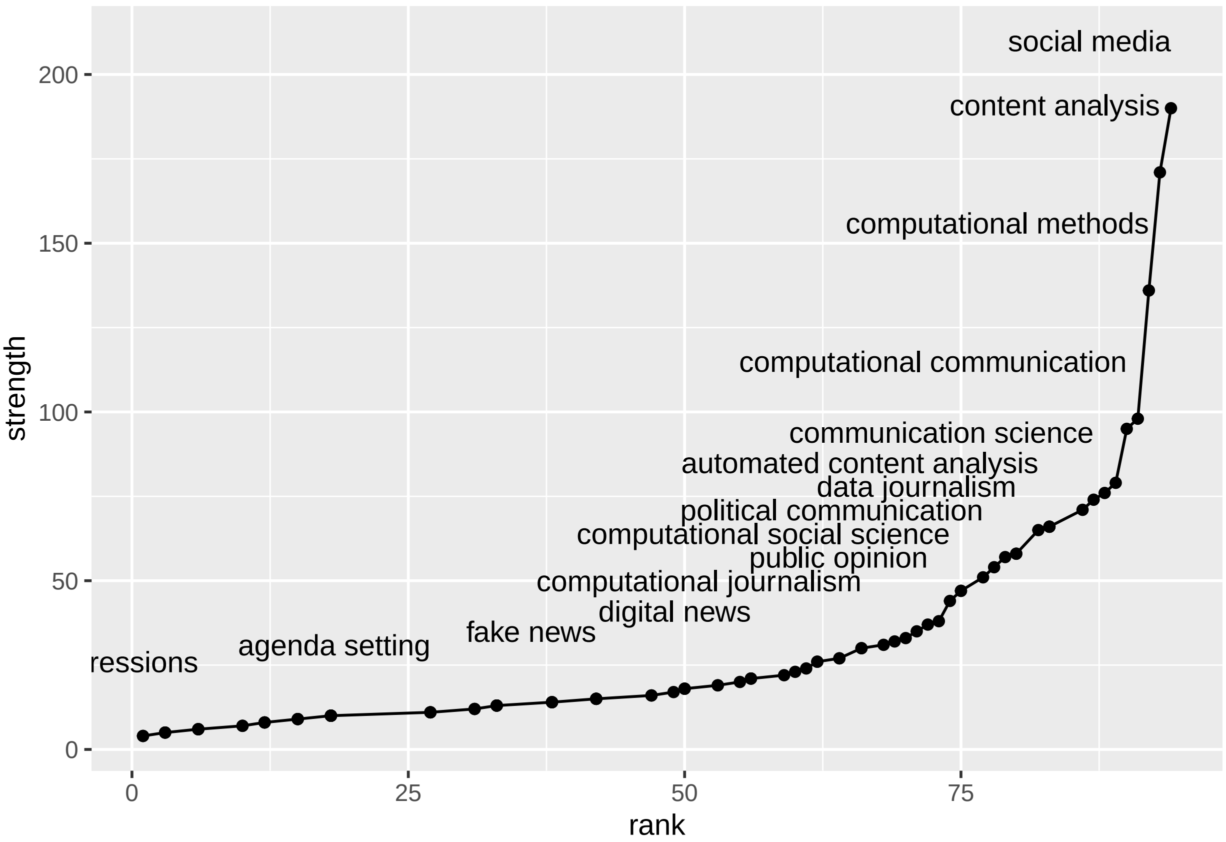
```r
ggraph(g, layout="stress") +
  coord_fixed() +
  expand_limits(x=c(-3, 3)) +
  geom_edge_link(aes(alpha=weight)) +
  geom_node_point(shape="circle filled", fill="white") +
  geom_node_text(aes(label=name), hjust="outward", check_overlap=TRUE) +
  guides(edge_alpha="none")
```



Judging by a first look, there seems to be a certain set of keywords which form a tight core in the middle of the graph. This core looks almost like a polyhedral cube. We can keep this in mind for a later analysis. The outer aura of the graph is dominated by political CommSci terms, such as political communication, affective polarization, computational propaganda, etc. This could merely be a product of the visualization. Methods are also relatively prevalent in our graph, e.g. automated content analysis, computational analysis, computational text analysis, computational content analysis, computer vision, etc. All in all, it's safe to say that the graph can give us a lot of impulses to think about when it comes to the identification and selection of keywords.

In a next step, we can look at the interior of the core and see which connections between keywords are the strongest. I use the `strength`-function of the `igraph`-package which sums up the edge weights for each vertex and plot our results in a ranked chart.

```
strengths <- strength(g)

# create ordered data frame for the plot
data.frame(term=names(strengths), strength=strengths, row.names=NULL) %>%
  mutate(rank=rank(strength, ties.method="min")) %>%
  arrange(strength) ->
  term_strengths

ggplot(term_strengths, aes(x=rank, y=strength, label=term)) +
  geom_line() +
  geom_point() +
  geom_text(data=filter(term_strengths, rank>5), hjust="right", nudge_y=20, check_overlap=TRUE)
```



Next is the chart above in a ranked order as a list. A high strength value indicates a strong connection of a keyword to others. The first and second page of keywords exhibit high strength values after which the keyword co-occurrence drops relatively sharp Interestingly enough, page one and two contain a variety of different keywords. Naturally, social media is on top of the list. Methods descriptions such as, automated/computational text analysis, automated content, supervised/unsupervised machine learning, network analysis, etc. are relatively frequent. Yet, terms such as “news media”, “online news”, “data journalism”, “big data”, “political communication”, etc. seem really important as well. This list should definitely spark an interesting discussion about keywords in our project.

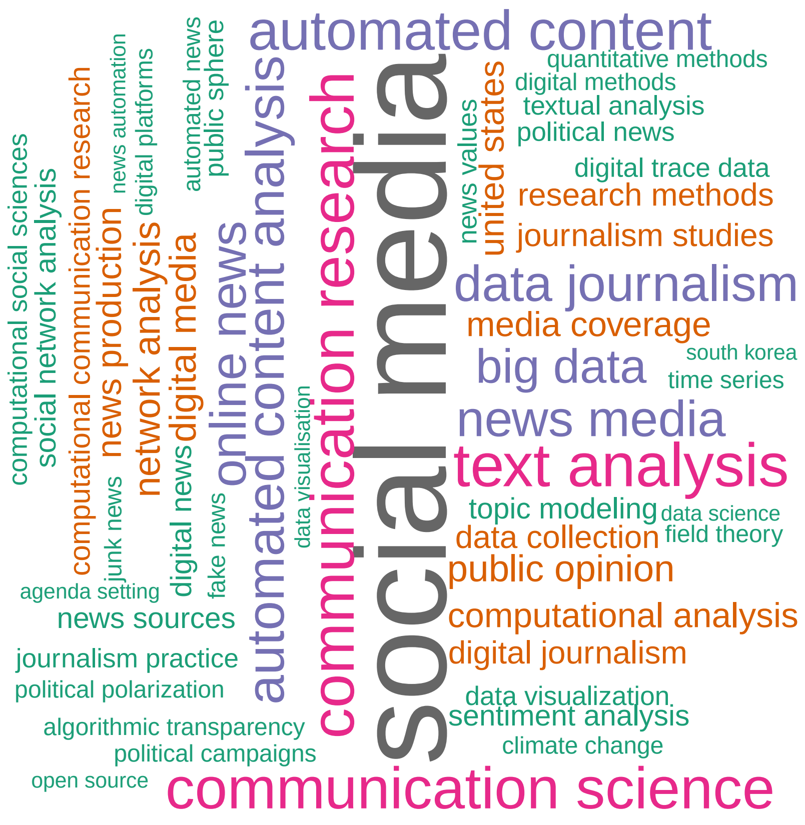
```
# change to decreasing order for better interpretation
rev_df <- term_strengths[rev(rownames(df)),]
rev_df
```

```
## [1] term     strength rank
## <0 Zeilen> (oder row.names mit Länge 0)
```

## Wordcloud

To round up our analysis in a pleasing visual manner, I created a word cloud with the most important topics based once again on the strength, i.e. co-occurrence.

```
strengths <- strength(g)
df <- as.data.frame(strengths)

df <- strengths %>% as_tibble()
df <- add_column(df,names(strengths))

wordcloud(words = df$`names(strengths)`, freq = df$value, min.freq = 10,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```



## Conclusion

Both the graph and ordered list of keyword importance are a really interesting contribution to our keyword discussion.

I also performed this analysis with the comprehensive keyword list comprising of author-tagged keywords and “Fakerake” keywords derived from both the abstract and title. The results are relatively similar, which is why I spared you with them in this Markdown-file. In case you are interested in the latter analysis, I can gladly share them of course.

End of this document.