Text Mining i systematiske reviews

Jartoft, Vibeke; Møller, Anne Vils; Thise Pedersen, Jette

Published in:
Revy

DOI (link to publication from Publisher):
10.22439/revy.v43i2.5993

Creative Commons License
Ikke-specificeret

Publication date:
2020

Document Version
Også kaldet Forlagets PDF

Af Vibeke Jartoft, forskningsbibliotekar, vija@kb.dk
Anne Vils Møller, forskningsbibliotekar, anvm@kb.dk
Jette Thide Pedersen, forskningsbibliotekar, jtp@aub.aau.dk
DOI: 10.22439/revy.v43i2.5993

Text Mining i systematiske reviews

Artiklen er udarbejdet i forbindelse med DEFF-projektet: *"Udvikling af servicemodeller og kompetencer i forhold til bibliotekernes rolle ved udarbejdelse af systematiske oversigtsartikler – Systematic Reviews" (http://projektbank.dk/).*

Som en del af projektet har vi undersøgt, hvordan text mining kan bidrage til udarbejdelsen af systematiske forskningsoversigter, også kaldet reviews.

Det er inden for de senere år blevet mere og mere udbredt at arbejde med text mining og data mining i et forsøg på både at identificere, systematisere og lave udtræk i den enorme mængde data, der skal forceres i udarbejdelsen af systematiske reviews. Et systematisk review er en forskningsmetode, der igennem mange år er blevet anvendt især inden for natur-, sundheds- og uddannelsesvidenskaberne. Det drejer sig i korte træk om at indsamle mest mulig af den forskning, der på et givent tidspunkt er blevet udarbejdet om det forskningsspørgsmål, man ønsker at belyse, og derefter sammenfatte den indsamlede viden i en forskningsoversigt. Systematiske reviews drejer sig med andre ord om at udarbejde en forskningsoversigt og -syntese baseret på eksisterende forskning inden for det aktuelle forskningsfelt.

For at foretage systematiske litteratursøgninger i de forskellige databaser er det nødvendigt at udvælge nøgleord/søgetermer og synonymer, der skal sikre så god og præcis en søgning som muligt. Inklusions- og eksklusionskriterier som f.eks. tidsperiode, geografi, sprog mm. defineres. Det kan være svært at finde søgetermer, der kan identificere alle relevante artikler. Det betyder, at vi kan miste referencer i de tilfælde, hvor søgetermerne ikke fanger begreber i den indekserede artikel, selvom den udmærket kan omhandle det fænomen, vi søger.

Uanset hvor kvalificeret man er til at udarbejde søgestrategier, så er det desværre sådan, at sensitiviteten og nøjagtigheden i afsøgningen af databaserne er forholdsvis lav, og reviewerne må derfor ofte, i det man kalder screeningsprocessen, manuelt sortere flere tusinde titler og abstracts for at ende ud med et meget lille antal relevante titler. Hertil kommer, at mængden af forskningspublikationer, der udarbejdes og udgives på verdensplan er støt stigende.

Vi har i løbet af arbejdet med DEFF-projektet forsket i de text mining programmer, der er udviklet og ud fra vores perspektiv særligt anvendelige med henblik på at lette og evt. forbedre udarbejdelsen af systematiske reviews.

Hvad er text mining

I bogen *Systematic Searching. Practical ideas* for improving results gives et glimrende indblik i den aktuelle status for anvendelsesmulighederne af forskellige text mining programmer i forbindelse med udarbejdelsen af systematiske reviews. Julie Glanville beskriver text mining som følgende:

"Text mining er et paraplybegreb. Det dækker en stor variation af teknikker, som involverer computeranalyser af ord i dokumenter og deres relation til hinanden i dokumentet. Text mining kan strække sig fra at være simple optællinger af det antal gange, ord forekommer i dokumenter (frekvensanalyse), til at skelne relevant fra irrelevant tekst (maskinlæring). Et andet aspekt af text mining er semantisk analyse, hvor ords rolle i sætninger kan specificeres således, at ordenes kontekst kan identificeres og skabe muligheder for at gøre søgninger mere præcise og forståelige". (Oversættelse ved artiklens forfattere)

Det nationale center for text mining NaCTeM i Storbritannien (http://nactem.ac.uk/faq.php?fag=1) beskriver text mining som: (Oversættelse ved artiklens forfattere)

"... en proces, hvor man afdækker og udtrækker viden fra ustrukturerede data. Denne omfatter tre hovedaktiviteter:

- Informationsgenfinding til indsamling af relevante tekster
- Informationsudtræk til at identificere og udtrække enheder, fakta og relationer mellem dem
- Data mining til at finde forbindelser mellem de informationsdele, som er udtrukket af de forskellige tekster.

I det følgende vil vi redegøre for, hvordan vi har arbejdet med at finde frem til et bredere udvalg af tilgængelige text mining programmer. Vi har udvalgt programmer, som vi har testet og afprøvet i egen praksis og i samarbejde med forskere og studerende.

Projektets forløb

Som udgangspunkt foretog vi en omfattende litteratursøgning i databaserne Web of Science, Scopus, PubMed, LISTA (Library, Information Science and Technology Abstracts), ACM DL (Association for Computing Machinery Digital Library) i foråret 2018 og fandt 322

artikler, hvilke vi screenede og fandt 68 relevante for det videre arbejde. Heraf har vi fordybet os i et udvalg af forskningsartikler og samtidig ladet os inspirere til det videre praksisarbejde.

Derudover har to af os deltaget i et kursus om systematiske reviews ved York Health Economics Consortium (YHEC). Her blev en vifte af forskellige text mining programmer præsenteret. Chris Marshall, som forestod en del af kurset, er hovedkraften bag hjemmesiden SR Toolbox: http://systematicreviewtools.com/, som er en af de mest omfattende ressourcer, hvis man vil orientere sig i feltet og finde text mining programmer til forskellige faser i arbejdet med systematiske review.

Netop for at finde rundt i junglen af programmer kan det give et overblik at opdele disse i forhold til, hvilken del af processen i et systematisk review, de kan anvendes i:

Grov inddeling af programmer

1. Inspirationsfasen/eksplorativfasen til at få et overblik over emnet, identificere nøgleord, synonymer, begreber og fraser eller finde umiddelbart relevante artikler.

2. Screeningsfasen, hvor programmerne kan assistere, øge hastigheden samt evt. tilbyde prioritering af resterende referencer ved hjælp af machine learning.

3. Data extraction/quality appraisal (kan endnu ikke klares uden menneskelig indblanding, derfor ikke medtaget her).

Text mining programmerne findes tilligе i forskellige former, f.eks. som et sidesystem (interface) til en eksisterende database. F.eks. PubReMiner for PubMed. De findes også som et standalone program, hvortil man uploader data. F.eks. VOSviewer eller Voyant. Desuden findes de som en større softwarepakke. f.eks. EPPI-reviewer.

Præsentation af programmer

På baggrund af litteraturstudiet, anbefalinger fra York samt egne kriterier såsom anvendelighed –tidsforbrug – sværhedsgrad - hvilken fase i processen – vedligehold– validitet samt målgruppen har vi testet nedenstående 6 programmer.

1. **Carrot2** er et document clustering tool, der ud fra en hurtig søgning i enten PubMed eller i en samling søgemaskiner (herunder Google, Bing, Yahoo o.l.) identificerer termer, begreber og relevante artikler og organiserer dem i tematiske samlinger. I eksemplet herunder er der søgt på "Filter bubble" og vi får vist temaer, andre mulige søgeord som eksempelvis "Personalized filtering" samt artikler, der kan give en bedre forståelse af emnet. Ved klik på et cluster fremkommer relevante links i højre side af billedet.
https://search.carrot2.org/#/search/web/filter%20bubble/treemap

2. **PubMed PubReMiner** laver frekvensanalyse ud fra en indtastet søgestreng og viser f.eks. de mest anvendte MESH- termer (kontrollerede emneord), hvor de fleste artikler er udgivet, og hvilke forfattere der har skrevet dem.
https://hgserver2.amc.nl/cgi-bin/miner/miner2.cgi

3. **HelioBLAST** finder ud fra en kendt tekst lignende tekster fra PubMed og giver dem en sammenlignelighеds score. Man kan desuden se foreslåede eksperter inden for emnet samt keywords. Det er muligt mod betaling at få Helioblast til at søge i andre databaser.
https://helioblast.heliotext.com/

4. **VOSviewer** viser temaer, ord frekvenser og sammenhænge inden for et sæt af referencer.
https://www.vosviewer.com/

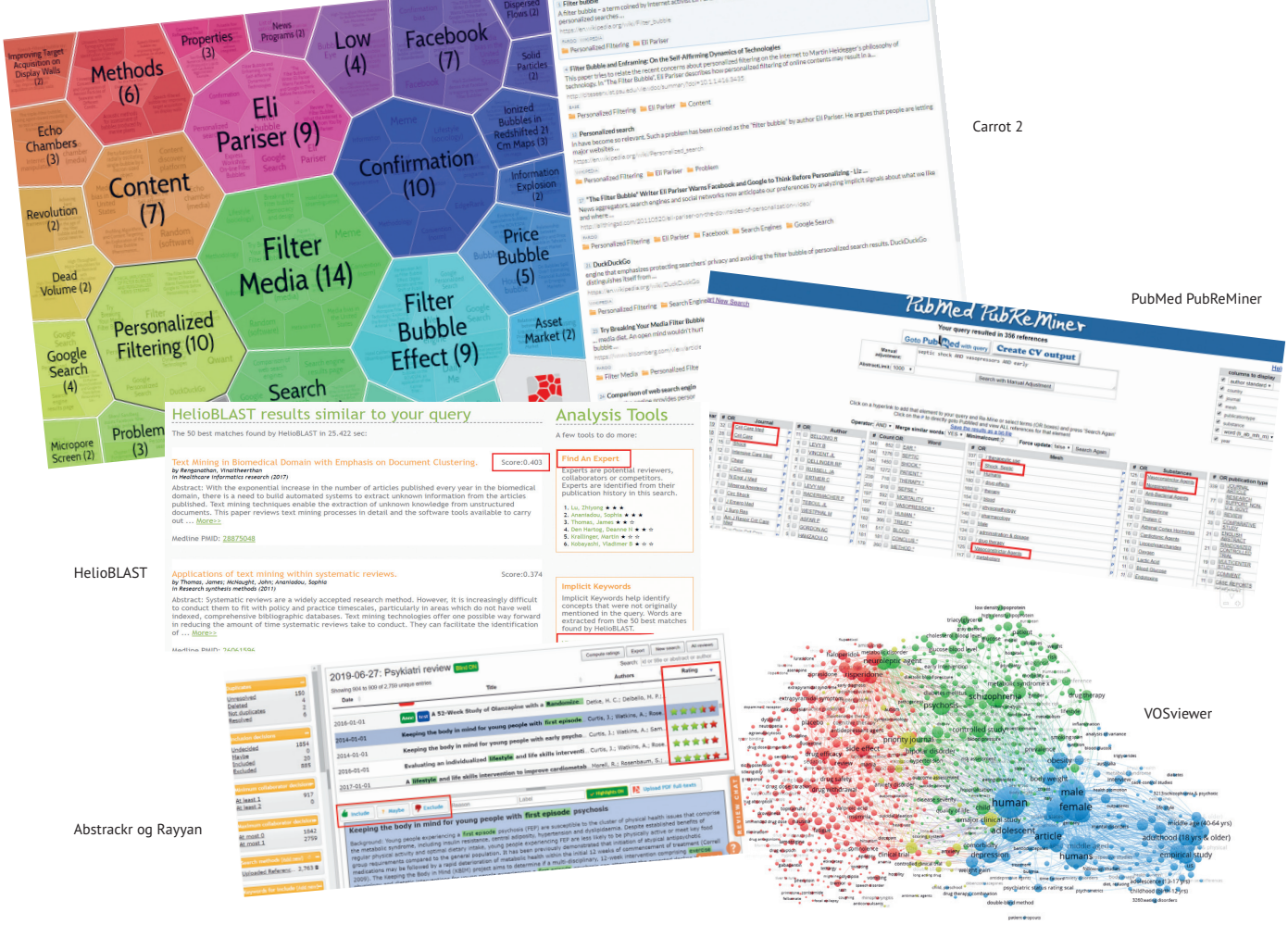
5/6. **Abstrackr og Rayyan** er begge screeningsredskaber, hvortil man kan uploade referencer og screene (dvs. vælge om den enkelte reference er relevant eller irrelevant for forskningsspørgsmålet) ud fra titel og abstract. I Rayyan kan man efter screening af 50 referencer udregne en rating, og Rayyan giver de resterende referencer stjerner ud fra, hvor relevante den regner med, de er. Abstrackr har en lignende funktion med en prioritering.

Fordele

Ved gennemgang af litteraturen og gennem egne erfaringer kan vi se flere fordele ved disse text mining programmer. Programmer som Carrot2, PubReMiner, HelioBLAST, Voyant, VOSviewer og andre lignende programmer er gode til identifikation af keywords og kontrollerede emneord og kan evt. forkorte tiden ved udvikling af søgestrenge samt gøre søgestrengen mere præcis og sensitiv. Det gælder dog hovedsagelig ved komplekse emner. De kan derudover synliggøre trends, sammenhænge, subtile emner eller irrelevante temaer. Screeningsprogrammerne Rayyan og Abstrackr kan i sig selv forkorte screeningstiden, og de indbyggede prioriteringsmekanismer (algoritmer) kan med en god portion kritisk tænkning evt. bruges som 2. screener eller til sortering af screeningsrækkefølgen. Alle ovennævnte programmer er gratis at anvende og flere af dem er hurtige og lette at gå til.

Ulemper

Der ses dog også ulemper eller betænkeligheder ved programmerne. Text mining programmerne kan være effektive at anvende, men det kan være svært at dokumentere, hvad man har gjort, og man kommer nok ikke udenom at skulle supplere med sædvanlige metoder. Selvom programmerne kan synes at være objektive, har de kun det subjektive input at arbejde med, og man kan altid være kritisk overfor, om træningssættet til machine learning funktionerne var det rette. Endvidere skal man være opmærksom på, at man med de programmer, som kan benyttes gratis på internettet, vanskeligt kan gennemskue, hvilke



prioriteringsmekanismer (algoritmer), der ligger til grund for programmerne, ligesom det kan være vanskeligt at gennemskue, hvad eget input og data kan blive brugt til af udbyderne. Derudover er flere programmer stadig betaversioner, som ikke nødvendigvis bliver opdateret. Sidst men ikke mindst findes der inden for feltet mere sofistikerede programmer, hvilke kræver ekspertise og tid at sætte sig ind i.

Afslutning

Som vi har vist i artiklen, er der mange raffinerede måder text mining kan bidrage til udarbejdelsen af et systematisk review. Der findes også mange andre text mining programmer ud over de ovennævnte, og der er bestemt potentiale for, at disse kan både lette og supplere arbejdet med systematiske reviews.

Men uanset, hvor veldefinerede, præcise og tilrettede søgninger man har fået udarbejdet, kan man heller ikke med disse metoder sikre, at der ikke er væsentlige artikler eller anden vidensformidling, der er faldet uden for søgeresultaterne, så det er til stadighed relevant at benytte sig af de traditionelle afsøgninger i databaser, faglige netværk, 'peer to peer' udvekslinger, kædesøgninger osv. Det er med andre ord stadig nødvendigt at have faglig udveksling i forskningsnetværk og andre relevante sammenhænge for at nå så langt omkring belysningen af det aktuelle forskningsspørgsmål som muligt.

Brugen af text mining programmer er i biblioteksverdenen et relativt nyt fænomen. Det betyder, at det ikke er en integreret del af informationssøgningsprocessen, og derfor først skal opdages og prøves af i det danske biblioteksvæsen. Vores oplevelse viste en erfaring hen imod, at det er vigtigt, at man tager programmerne i brug over flere omgange, før man reelt kan vurdere, hvor i informationssøgningsprocessen programmerne kan udgøre et værdifuldt bidrag. Text mining er et område i vækst i biblioteksvæsenet, derfor håber vi at have skabt interesse for feltet med denne artikel.

Links til programmer:

Systematic Review Toolbox - http://systematicreviewtools.com/

Carrot2 - http://search.carrot2.org/stable/search

PubReMiner - https://hgserver2.amc.nl/cgi-bin/miner/miner2.cgi

HelioBLAST - https://helioblast.heliotext.com/

Voyant - https://voyant-tools.org/

VOSviewer - http://www.vosviewer.com/

Rayyan - https://rayyan.qcri.org/welcome

Abstrackr - http://abstrackr.cebm.brown.edu/account/login