FOI

Metodikutveckling för Scientometri

ANNA LINDBERG, JOHAN SCHUBERT, JONATAN WESTMAN, TOMMY WESTMAN

FOI-D--0716--SE Mars 2016

FOI-D--0716--SE

Anna Lindberg, Johan Schubert, Jonatan Westman, Tommy Westman

Metodikutveckling för Scientometri

Titel	Metodikutveckling för Scientometri
Title	Methodology development for Scientometrics

Rapportnr/Report no	FOI-D--0716--SE
Månad/Month	Mars/March
Utgivningsår/Year	2016
Sidor/Pages	18 p
Kund/Customer	Försvarsmakten/Swedish Armed Forces
Forskningsområde	12. Övrigt
FoT-område	Avskanning av forskningsfronten
Projektnr/Project no	E72640
Godkänd av/Approved by	Matts Gustavsson
Ansvarig avdelning	Försvars och Säkerhetssystem

Sammanfattning

I föreliggande rapport redovisas upplägg av och inledande experiment kring en scientometrisk ansats för att inhämta och analysera trender i forskning. Arbetet har utförts inom ramen för FOI:s projekt Avskanning av forskningsfronten. Två metoder har utvecklats: *Kända nyckelord* och *Revolutionary nyckelord*. Metoderna gör en skanning genom flera sökningar och identifierar forskningsområden med stigande trender. Syftet med experimenten som här redovisas är att undersöka möjligheter att tillämpa de föreslagna metodförslagen. Målet är att identifiera möjligheter och utmaningar med metoduppläggen för att under 2016 skapa en användbar scientometrisk metod för horizon scanning. En slutsats från årets arbete är att fokus under 2016 bör ligga på ett samordnat (*i*) metodutvecklande, (*ii*) implementering och (*iii*) metodtestning. Att genomföra en första horizon scanning med det verktyg som under 2016 fortfarande kommer att vara under utveckling bör vara en mindre omfattande aktivitet i projektet som kan provas mot slutet av året.

Nyckelord: Scientometri, horizon scanning.

Summary

In the present report we report on planning and initial experiments on a scientometric approach to gather and analyse trends in research. The work has been carried out in the context of FOI's project scanning of the research frontier. Two methods have been developed: *Known nyckelord* and *Revolutionary nyckelord*. The methods perform a scan through several searches and identifies areas of research with rising trends. The purpose of the experiments reported here is to examine ways to apply the proposed method proposals. The goal is to identify opportunities and challenges with the method to create a useful scientometric method for horizon scanning in 2016. A conclusion from this year's work is that the focus in 2016 should be on a coordinated (*i*) methodological development, (*ii*) implementation and (*iii*) method testing. To implement a first horizon scanning with the tool during 2016 that is still under development should be a less extensive activity in the project that can be tried towards the end of the year.

Keywords: Scientometrics, horizon scanning.

Innehållsförteckning

1 Inledning

I föreliggande rapport redovisas upplägg av och inledande experiment kring en scientometrisk ansats för att inhämta och analysera trender i forskning. Arbetet har utförts inom ramen för FOI:s projekt *Skanning av forskningsfronten*.

Två metoder har utvecklats: *Kända nyckelord* och *Revolutionary nyckelord*. Med hjälp av metoderna kan förhoppningsvis intressanta tendenser inom olika forskningsområden identifieras utifrån inhämtad data från artikeldatabaser. Centralt för de bägge metoderna är att analysera användandet av nyckelord för att identifiera forskning och vetenskapliga vinningar som snabbt utvecklas och som kan innebära disruption eller stora förändringar inom säkerhets- och försvarssektorn.

1.1 Syfte och mål

Syftet med experimenten som här redovisas är att undersöka möjligheter att tillämpa de föreslagna metodförslagen.

Målet är att identifiera möjligheter och utmaningar med metoduppläggen för att under 2016 skapa en användbar scientometrisk metod.

1.2 Bakgrund

Det metodförsök som redovisas i föreliggande rapport är del av ett större avskannande arbete inom FOI. För ytterligare teori kring upplägg och bakgrund refereras till dessa rapporter.[1,2]

De scientometriska metodernas syfte i en horizon scanning-process är att automatiskt identifiera trender inom vetenskapliga och teknologiska nyckelord och ämnesområden som underlag för fortsatt arbete. Målet är att identifiera ämnesområden som kan ha stor påverkan inom säkerhets- och försvarssektorn. Automatisk inhämtning är ett tillvägagångssätt för att undvika alltför mycket bias inledningsvis i horizon scanning-processen och förordas i viss litteratur. Annan litteratur pekar på risker vid användande av automatisk datainsamling med hjälp av data-mining, scientometrics och andra metoder. Riskerna är att man snarare hamnar i extrapolering av data, och i sökning istället för i skanning. En utmaning är därför att metodologiskt finna sätt att hitta de ämnen och strukturer som är nya alternativt inte redan identifierats, baserat på automatiska datainsamlingssätt.

I föreliggande arbete har peer-reviewade tidskrifter i akademiska och vetenskapliga databaser använts för sökning. Automatiserade datainsamlingsmetoder kan också använda andra källor såsom press (magasin, tidningar, etc.) och portaler med fokus på Science & Technology (S&T), wikis, twitter, bloggar, konferensbidrag och patentdatabaser. Valet av källa beror av frågeställningar och önskade resultat. Scientometrics definieras som

> *"The scientific measurement of the work of scientists, especially by way of analysing their publications and the citations within them."* [3]

eller

> *"Scientometrics is the study of measuring and analysing science, technology and innovation. Major research issues include the measurement of impact, reference sets of articles to investigate the impact of journals and institutes, understanding of scientific citations, mapping scientific fields and the production of indicators for use in policy and management contexts."* [4]

[1] Kindvall, G., Lindberg, A., Trané, C. och Westman, J. (2016). Exploring future technology development, FOI-R--4196--SE.
[2] Lindberg, A. och Westman, J. (2015). Förslag till process för Horizon Scanning, FOI rapport (under utgivning).
[3] https://en.wiktionary.org/wiki/scientometrics (december 2015)
[4] https://en.wikipedia.org/wiki/Scientometrics (december 2015)

Scientometrics innebär studier av vetenskap, teknologi och innovation från ett kvantitativt perspektiv. Text används som empiri.

Scientometrics kan användas för olika syften. I Leydesdorff och Milojević (2015)[5] visas på fem olika användningsområden:

- The measurement of impact,
- The delineation of a reference set,
- Theories of citation,
- Mapping science, co-occurance,
- Policy and management contexts.

I det avskannande projektet vill vi framförallt fånga och använda publicerat data för att förstå dynamiken i vetenskaplig forskning. Vi vill analysera hur forskning och idéer sprids och utvecklas. Underlaget ska sedan användas för att visa på ämnen med stor framtida och möjligen disruptiv effekt. Användningsområden för de scientometriska metoder som redovisas i föreliggande rapport är därför *mapping of science*, *co-occurrence* och *policy och management contexts*.

Scientometrics och användningsområdena co-authorship analysis, co-word analysis och co-citation analysis klustrar och analyserar tids-samplat data baserat på olika faktorer. Identifierade termer länkas till varandra och strukturer över utveckling skapas. Med co-authorship analysis identifieras sociala strukturer, co-word redovisar forskning som använder samma nyckelord och co-citation identifierar strukturer för citering och kan sägas visa ett släktschema över referering mellan artiklar. För och nackdelar med metoderna beskrivs i Cozzens *et al.*[6] och Amanatidou *et al.*[7]

Flera källor uppger att de använder scientometrics som metod för att samla in data. Information kring faktiska tillvägagångssätt, vilka resultat som ges och hur resultaten väljs ut för fortsatt analys är dock mycket sparsmakad. Vi tror att scientometrics är en användbar metod, men har också fått uppfattningen att det vi försöker åstadkomma är ett område för pågående forskning inom scientometrics. Vi väljer därför att avsluta kapitlet om bakgrund med följande citat från Leydesdorff och Milojević (2015):

> *"In summary, the modeling of knowledge exchanges in scientific discourses cannot be reduced to the exchanges of information in co-authorship, co-word, or citation relations. Models as entertained in the sciences enable researchers both to provide meaning to possible future states and to specify uncertainty. The measurement of the communication/sharing of meaning among frames of reference—for example, in university-industry-government relations—is very much on the research agenda of scientometrics."* (s. 15)

> *"Models also serve to communicate possible future states. A construction of future states from a knowledge-based perspective can be modeled as hyper-incursion. The modelers/scientists thus become carriers who are differently positioned in terms of their reflexive and communicative competencies; for example, as scientific explorers and/or appliers of engineering knowledge."* (s. 16).

[5] Leydesdorff, L. and Milojević, S. (2015). Scientometrics, in: J.D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*. Elsevier, 2015, pp. 322–327. doi:10.1016/B978-0-08-097086-8.85030-8

[6] Cozzens, S., Gatchair, S., Kang, J., Kim, K.-S., Lee, H. J., Ordóñez, G., Porter, A. (2010). Emerging technologies: quantitative identification and measurement. *Technology Analysis & Strategic Management* **22**(3):361–376. doi:10.1080/09537321003647396

[7] Amanatidou, E., Butter, M., Carabias, V., Könnölä, T., Leis, M., Saritas, O., Schaper-Rinkel, P., van Rij, V. (2015). On concepts and methods in horizon scanning: Lessons from initiating policy dialogues on emerging issues. *Science and Public Policy* **39**(2):208–221. doi:10.1093/scipol/scs017

2 Metodupplägg

Detta kapitel beskriver två förslag på automatiska datainsamlings- och skanningsmetoder. Syftet med metoderna är att finna relevanta signaler på ämnesområden med uppåtgående trend. Metoderna *Kända Nyckelord* respektive *Revolutionary Nyckelord* beskrivs i kapitlet.

2.1 Sökning och skanning

Den automatiska inhämtningen syftar till att samla in data genom att källor avsöks eller skannas med olika sökord. Distinktionen mellan sökning och skanning är hårfin, men ska inte glömmas bort. En skanning syftar till att hitta något man inte vet att man söker medan en sökning normalt sett visar empiri för en ställd fråga som man direkt vill ha svar på. Nedan redovisade hypotetiska metoder använder sökord i syfte att *se längre* än vad som normalt görs vid sökningar.

2.2 Metoden ’Kända Nyckelord’

Sökmetoden *Kända nyckelord* innebär iterativa sökningar av nyckelord, dvs. sökord bestående av kända fraser och ämnesområden, i flera steg. I det första steget används nyckelord som forskare idag redan använder för att söka inom domänen försvar och säkerhet, möjligtvis kombinerat med nyckelord från specifika ämnesområden. Resultaten från den första sökningen i form av nya identifierade nyckelord utgör grunden till en andra sökning. (Sökningar kan också göras inom andra domäner och med kända civila forskningsområden.) Utifrån identifierade nyckelord (eller ett urval eller kombinationer därutav) görs nya sökningar. Urval och kombinationer av sökord kan kombineras med varandra och med andra typer av sökord för att minska mängden utdata.

Metodsteg:

Identifiering av första ordningens sökord

1. Identifiera första stegets nyckelord.

Sökning 1 (relaterade nyckelord)

2. Sök efter nyckelord i valda källor (inom abstract, nyckelord, topics).
    - Sök individuellt för första ledets alla nyckelord i valda källor (abstract, nyckelord, topics), för en utvald period (t.ex. 2015, 2014–2015, …).

   Resultat: en mängd artiklar som ger andra ordningens nyckelord, den kompletta listan med dessa nyckelord används som input till nyckelordsanalys.

Sökning 2 (frekvens) + Nyckelordsanalys 1 (urval)

3. Urval baserat på nyckelordsanalys. För alla andra ordningens identifierade i sökning 1 görs en årsvis sökning för de senaste fem åren, dvs. en sökning för 2011, en för 2012, osv. till 2015, för att göra en frekvensanalys på andra ordningens nyckelord. Denna analys utförs enligt följande:
    - Urval på andra ordningens nyckelord:
        i. om robust trend i alla kategorier (femårsperioden, större än y%)
        ii. om robust trend i en enskild kategori (femårsperioden, större än z% (> y))
        iii. om robust trend i minst två kategorier (större än x% (< y))

iv. om helt nya nyckelord, som består kommande år. Ett nytt nyckelord får inte existera år 2011, men måste existera år 2015.
v. Möjlighet till fine-tuning av procentsatser för att få hanterbart antal träffar

Resultat: ett filtrerat urval av andra ordningens nyckelord. Nota bene: Om dessa resultat är tillfredsställande så att Nyckelordsanalys 2 kan hoppas över, så söker man på artiklar och går till steg 6 i arbetsgången.

Sökning 3 (relaterade nyckelord) + Sökning 4 (frekvens) + Nyckelordanalys 2 (urval)

4. Med resultatet från steg 3 som indata så upprepas steg 2 och 3. NB: Om resultaten från nyckelordsanalys 1 bedöms som tillfredställande kan denna punkt hoppas över.
   Resultat: ett filtrerat urval av andra ordningens nyckelord, kallad HS-lista.
5. Automatisk analys av HS-lista, flera olika alternativ att filtrera HS-listan enligt nedan a–d. Ett syfte är att reducera HS-listans storlek till något som är hanterbart av människor.
   a. Co-author analysis,
   b. Co-word analysis,
   c. Grafiska kluster (om möjligt),
   d. Impact factor.
6. Resultatet lämnas som input till nästa steg i processen, för analys av människor. Eventuellt sker insamling av relaterade artiklar.

2.3 Metoden ’Revolutionary Nyckelord’

Sökmetoden *Revolutionary nyckelord* innebär iterativa sökningar av nyckelord, i två steg. I det första steget används nyckelord som indirekt kan indikera identifierade framsteg, t.ex. *order of magnitude*, *revolutionary*, *disruptive* – det har föreslagits i viss litteratur att detta skall kunna vara en bra markör på disruptiv utveckling.

Tillvägagångssättet kan ge en förutsättningslös skanning gällande S&T-områden men förutsätter att forskaren eller annan författare identifierat ett forskningsområde eller upptäckt som mycket viktig. För att metoden ska fungera måste någon således ha kopplat upptäckten till dess möjliga implikation och dessutom ha publicerat kring reflektionen. Exempel på skanningsord från litteraturen är: *“innovation, emerging, issues, impact, change, future, emerging, promising, threatening, solutions, discoveries, problems, crisis, tensions, growth, breakthroughs, breakdowns, or new insights in combination with the domain demarcating keywords”*[8].

Till skillnad från metoden *Kända Nyckelord* görs bara en första ordningens sökning på nyckelord, vilket sedan genererar slutresultatet. Urval av och kombinationer av sökord kan kombineras med varandra och med andra typer av sökord för att minska mängden utdata. Metodsteg:

[8] Amanatidou, *et al*. (2012), p. 210.

Identifiering av första ordningens sökord

1. Identifiera första stegets nyckelord

Sökning 1 (relaterade nyckelord)

2. Sök efter nyckelord i valda källor (inom abstract, nyckelord, topics).
    - Sök individuellt för första ledets alla nyckelord i valda källor (abstract, nyckelord, topics), för en utvald period (t.ex. 2015, 2014-2015, …).

   Resultat: en mängd artiklar som ger andra ordningens nyckelord, den kompletta listan med dessa nyckelord används som input till nyckelordsanalys

Sökning 2 (frekvens) + Nyckelordsanalys 1 (urval)

3. Urval baserat på nyckelordsanalys. För alla andra ordningens nyckelord identifierade i sökning 1 görs en årsvis sökning för de senaste fem åren, dvs. en sökning för 2011, en för 2012, osv. till 2015, för att göra en frekvensanalys på andra ordningens nyckelord. Denna analys utförs enligt följande:
    - Urval på andra ordningens nyckelord:
        i. om robust trend i alla kategorier (femårsperioden, större än y%)
        ii. om robust trend i en enskild kategori (femårsperioden, större än z% (> y))
        iii. om robust trend i minst två kategorier (större än x% (< y))
        iv. om helt nya nyckelord, som består kommande år. Ett nytt nyckelord får inte existera år 2011, men måste existera år 2015.
        v. Möjlighet till fine-tuning av procentsatser för att få hanterbart antal träffar

   Resultat: ett filtrerat urval av andra ordningens nyckelord, kallad HS-lista
4. Automatisk analys av HS-lista, flera olika alternativ att filtrera HS-listan. Ett syfte är att reducera HS-listans storlek till något som är hanterbart av människor.
    a. Co-author analysis,
    b. Co-word analysis,
    c. Grafiska kluster (om möjligt),
    d. Impact factor.
5. Resultatet lämnas som input till nästa steg i processen, för analys av människor.

2.4 Risker

Tanken är att steget Nyckelordsanalys ska minska mängden träffar till ett hanterbart urval med hög grad av relevans. En stor risk är dock att de bägge metoderna genererar för mycket data. Vi riskerar vidare att missa nyckelord med låg frekvens. Dessutom riskeras med metoden *kända nyckelord* (baserat på de sökord som används i första steget) att missa områden där ingen säkerhets- eller försvarskoppling gjorts, medan vi med metoden *revolutionary nyckelord* riskerar dels att nätet kastas så brett och vitt att vi drunknar i felaktiga träffar, dels att vi riskerar att missa områden där författarna själva inte gjort en koppling till disruptivitet.

I vilken utsträckning dessa problem uppstår och hur de påverkar resultatet, kan bara testas genom provkörningar på relevanta datamängder.

3 Experiment

I två experiment har de praktiska förutsättningarna för att kunna använda de föreslagna metoderna *Kända nyckelord* och *Revolutionary nyckelord* testats. Upplägg och resultat redovisas i detta kapitel. Om metoderna ger relevanta resultat har därför ännu inte testats.

3.1 Bakgrund

Syftet med de experiment som beskrivs är att hitta möjliga problem kring dessa och att få uppslag kring hur man kan tänkas gå vidare med att lösa eventuella problem som uppstår. Att det kommer att finnas problem med att gruppera ihop nyckelord går att veta på förväg. Detta då nyckelord som fritt fått definieras av artikelförfattare kommer att användas och det finns inte några etablerade standarder för hur det ska göras. Men exakt vilka problem och i vilken omfattning de kommer att dyka upp är inte i förväg känt, inte heller hur pass långt man kan komma i just denna datamängd med väldigt förenklade och snabbt konstruerade metoder för en första hantering.

Metoderna föreslår en inhämtning av data i flera steg. Det är för de nedan beskrivna experimenten som syftar till att skapa fungerande metoder dock inte nödvändigt att genomföra samtliga steg i processen. Flera stegs sökningar påverkar framförallt bara hur stor den datamängd vi måste hantera blir. Nödvändigt i ett första test-skede är endast att ha en tillräckligt stor datamängd för att de första och mest grundläggande problemen att hantera ska uppstå. Det kan visa sig att när datamängden blir avsevärt större så dyker fler utmaningar relaterat till detta upp, men det är då ett senare problem att hantera.

3.2 Experiment med metoden ’Kända nyckelord’

3.2.1 Datamängd

Den datamängd experimenten utförts på är 5453 manuellt nedladdade artiklar publicerade under åren 2010 till 2015 från två olika databaser ur Web of Science baserat på ett fåtal riktade sökningar (vilka är dock inte relevant då syftet primärt var att hantera en godtycklig testdatamängd att utföra grundläggande experiment på). Dessa artiklar gav upphov till 13 962 nyckelord (vilket varierar beroende på hur mycket vi försöker ”städa” bland dem).

Ur denna testdatamängd användes inte all data tillgängliga i posterna utan bara det data som vi antas kunna ladda ner från web-services från samma källa[9]. Att använda dessa web-services kommer med stor sannolikhet att vara nödvändigt för att kunna inhämta och bearbeta större mängder data. Vid experimentets utförande så har vi inte haft någon åtkomst till dessa, men det måste vara målet på sikt att få (antingen för Web of Science, eller en annan jämförbar datakälla).

3.2.2 Problem vid extraktion och gruppering av nyckelord

De nyckelord som finns i datamängden är ofta fritt definierade av artikelförfattarna. Problemen med dessa nyckelord är även just detta; att de fritt definierats av artikelförfattarna, vilket innebär att de kan se ut hur som helst.

Det finns ingen standard vad gäller att ange ord i exempelvis singularform, pluralform, verbform, fullt utskrivet eller i form av en förkortning. Utöver det så innehåller orden gott om *skräptecken* såsom parenteser av olika typ, semikolon och annan kommatering. De kan

[9] Titel, författare, vilken källa artikeln publicerats i, författarens nyckelord, länk till artikelnummer i WOS. Notera att information såsom vilken vetenskaplig kategori artikeln tillhör eller annan bibliometrisk information såsom hur ofta den citerats och i vilka artiklar, räknar vi inte med att kunna få i dessa sökningar.

även fritt kombineras med andra ord sammanskrivet på olika vis vilket försvårar att gruppera ihop dem på lämpligt vis (det vill säga, ett nyckelord är inte på något vis nödvändigtvis ett enstaka ord). Det förekommer givetvis även felstavningar och olika sätt att stava samma term. Som exempel på detta kan vi tänka oss att vi är intresserade av att se trender inom 3D-animation.

Nyckelord inom området kan förekomma som:

> 3D Anim., 3D Animation, 3-D Animation, 3 D Animation, 3Danimation, Animation in 3D, Animations in 3-D, Free form animations in 3-D, Animations of 3D objects, 3-dimensional animation, Three-dimensional animation, m.fl.

Den term man letar efter kan som synes stå lite vart som helst i ett nyckelord. Det skulle även kunna vara så att det är registrerat som två separata nyckelord *3D* och *Animation*.

Antag att det inte ens var en så komplex term, utan att det bara var ordet *3-D*. Inte ens att lyckas slå samman detta enkla begrepp till en term är trivialt. I stort sett varje ingrepp när datamängden ska städas kan ställa till det på annat håll. Att exempelvis plocka bort detta bindestreck skulle kanske förbättra situationen just i detta fall, men då i gengäld ställa till det i ett annat sammanhang (exempelvis 14-3-3 protein). Motsvarande kan gälla för de flesta andra liknande operationer.

Nästa typ av problem är användandet av synonymer, där det givetvis är önskvärt att gruppera ihop synonyma begrepp. Ordlistor och annat kan hjälpa till viss del, men ibland är termerna så extremt specialiserade att en domänexpert med djup kunskap inom området krävs för att veta att begrepp är synonyma. Man kan givetvis även vara intresserad av att gruppera ihop begrepp som är väldigt närliggande även om de inte är äkta synonymer och det beror lite på hur man vill arbeta med datamängden hur pass närliggande begrepp man önskar gruppera ihop.

Det omvända, homonymer, är givetvis också ett problem. Inom det data vi tittat på har exempelvis ordet *landing* förekommit. Det kan betyda landning, landsättning, trappavsats, landstigning eller landningsplats, men stavas helt identiskt och kan förekomma som nyckelord. Där måste vi använda annan metadata för att kunna särskilja dem från varandra.

Det finns även ett nära besläktat problem vilket är att orden kanske inte i sig är äkta homonymer, men kanske används begreppsmässigt olika i olika ämnesområden eller som liknelser eller metaforer. Ett exempel är ordet *cloud* som både inom datornätverk och meteorologi avser moln, inom datornätverk dock bara som en liknelse. Däremot så är det inte relevant att gruppera ihop artiklar inom dessa områden.

Likaså finns problem att särskilja sådana fall som baseras på tämligen generiska termer, exempelvis att skilja på *growth* i kontexten ekonomi från *growth* inom agrikultur eller medicin. Men troligt är i alla fall att hopgruppering av artiklar relaterade med detta nyckelord med varandra inte är av intresse. Däremot så kan man mycket väl vilja göra det med artiklar inom olika ämnesområden om termen är klart mer specialiserad för att hitta innovativa tvärvetenskapliga tillämpningar eller likande.

Ytterligare ett problemområde är till vilken grad man vill gruppera ihop nyckelord eller inte. Det styrs i allra högsta grad av vilken finkornighet man önskar i sin sökning och vilket resultat man önskar uppnå. Vill man se generella och övergripande tendenser så vill man i hög grad gruppera ihop artiklar baserat på nyckelord. Är å andra sidan tendenser inom ett lite smalare område av intresse så vill man gruppera ihop dem desto mindre.

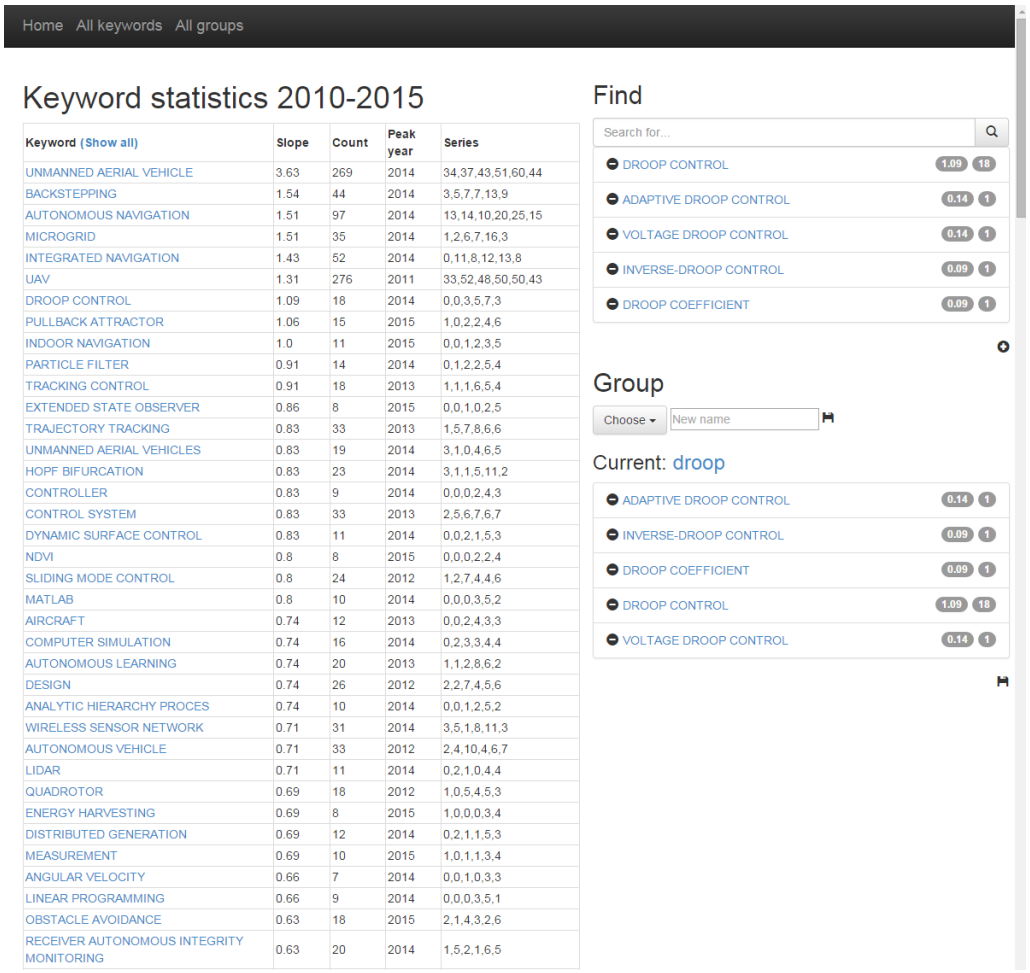
Det finns gott om olika automatiserade metoder för att angripa problemet inom natural language processing (NLP), maskininlärning och andra relaterade discipliner. Däremot finns inte någon möjlighet att inom detta års projektram göra ett sådant försök. Problemet är inte enkelt, så det kan krävas en insats i större omfattning för att hitta robusta metoder för att hantera det.

3.2.3 Verktygsprototyp

Det som gick att uppnå inom ramen för dessa försök var ett snabbt ihopsatt verktyg som kan hjälpa en användare att manuellt gruppera ihop nyckelord till intressanta grupper att studera. Detta är nödvändigt för att få någotsånär rättvisande resultat. Notera att nedanstående bilder är baserade på en liten exempeldatamängd och säger endast något om just den mängden.

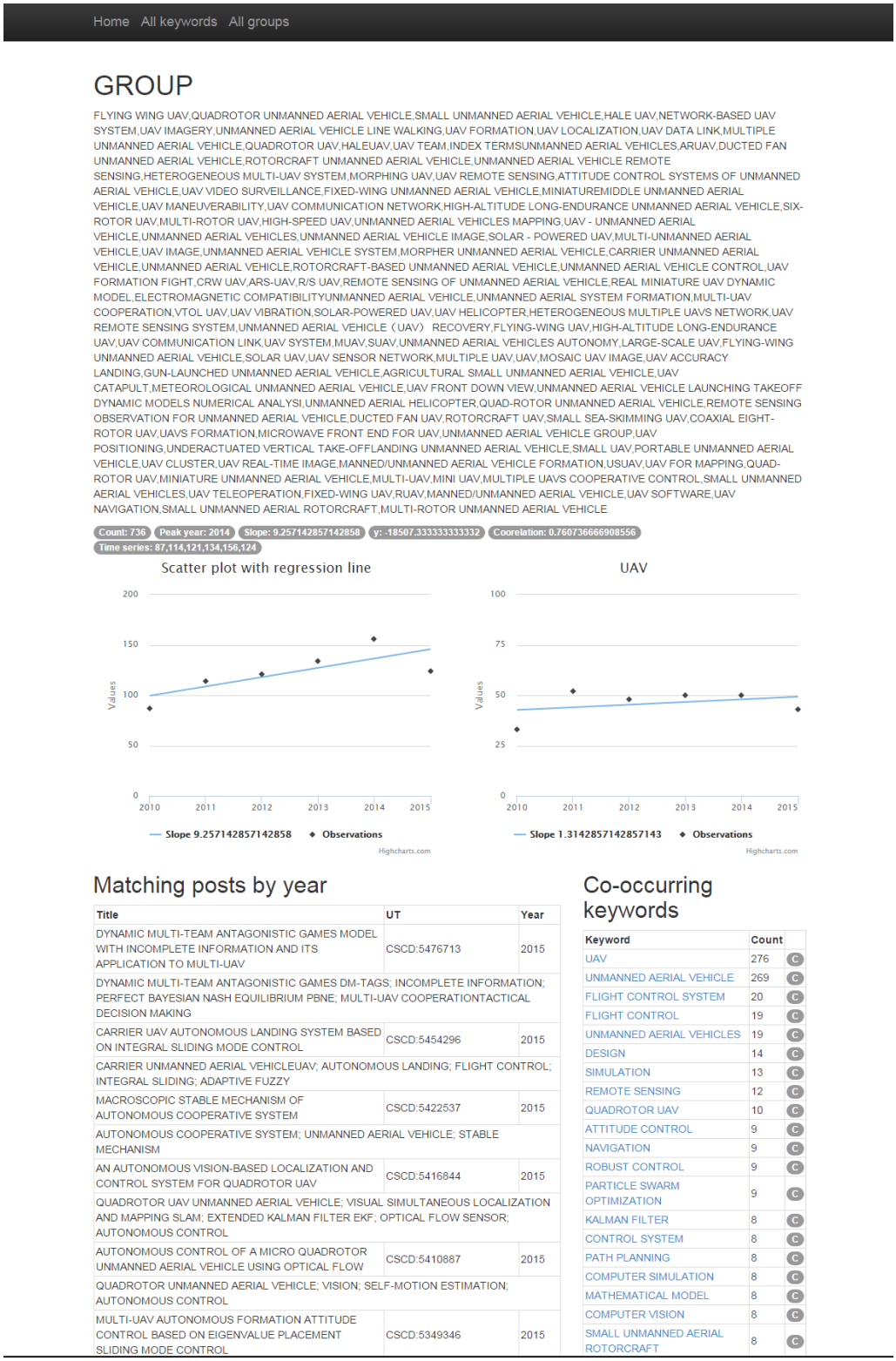
Det finns några operationer av enklare typ som har applicerats på nyckelord såsom att transformera alla ord till kapitäler, rensa bort skräptecken och kommatering, slå ihop multipla blanksteg i rad samt i vissa fall dela upp nyckelord när kommateringen antyder att det kan röra sig om felaktigheter i inmatningen och att det egentligen rör sig om multipla nyckelord inkorrekt separerade. Tid har dock inte funnits för att specifikt mäta effekten av varje enskild operation.

Nyckelord presenteras i en lista, Figur 1. Men hjälp av sökfunktionen på sidan kan termer enkelt läggas till i en grupp som användaren själv skapat och namngett.



Figur 1. Nyckelordstatistik.

Det går att klicka på antingen ett enstaka nyckelord eller på en grupp för att erhålla en sammanställning, Figur 2.



Figur 2. Trendanalys för nyckelord.

I gränssnittet är det möjligt att välja att jämföra gruppen eller det valda nyckelordet med ett annat nyckelord. I detta fall har användaren valt att jämföra en grupp sammansatt av termer som innehåller *unmanned aerial vehicle* eller *UAV* med endast nyckelordet *UAV* (som en

ensam term). Det visar ganska väl hur missvisande resultat det är möjligt att erhålla om nyckelorden inte grupperas samman på ett ändamålsenligt vis.

3.2.4 Annat

Experimenten har utförts sent under året, så vi har inte i hög grad märkt av vad som annars kan vara ett problem, att senaste årets resultat kan vara missvisande (då bara en del av årets publikation registrerats). Men detta ses inte som någon större utmaning, så det kommer bara bli att se till att innevarande års antal justeras till någon lämplig prognos.

Alla nedladdade poster sparas även i en grafdatabas. I sökgränssnittet mot grafdatabasen är det även möjligt att traversera den skapade grafen och utforska relationerna mellan posterna, Figur 3. Frånsett att det är ett trevligt visuellt sätt att interagera med datamängden så finns intressanta möjligheter som denna typ av teknologi medger för att eventuellt lösa vissa av problemen associerade med nyckelord.

När posterna sparas i denna typ av datamängd (när den blir tillräckligt stor) kommer relaterade poster att gruppera sig tillsammans i olika delar av grafen. Detta kommer naturligt av att poster inom samma ämnesområden tenderar att vara skrivna av samma författarmängd (inte nödvändigtvis som huvudförfattare), publicerade i samma tidskrifter och använder samma *ytor* av nyckelord. Detta är möjligt att göra även i en annan typ av databas, dock är just grafdatabaser väldigt anpassade för denna typ av sökningar (att gå ett steg i någon riktning från en specifik nod är tidsmässigt väldigt snabbt jämfört med i en relationsdatabas). Dessa möjligheter har dock inte hunnit undersökas i någon större omfattning inom ramen för detta års projekt.

3.3 Experiment med upptaggning av ’revolutionära termer’

En hypotes var att en ordlista med olika begrepp av typen *breakthrough*, *significant*, *revolutionary*, m.fl. skulle kunna användas för att eventuellt hitta särskilt intressanta artiklar genom att matcha dessa mot använda nyckelord och artiklars titlar, och sedan sålla den erhållna mängden.

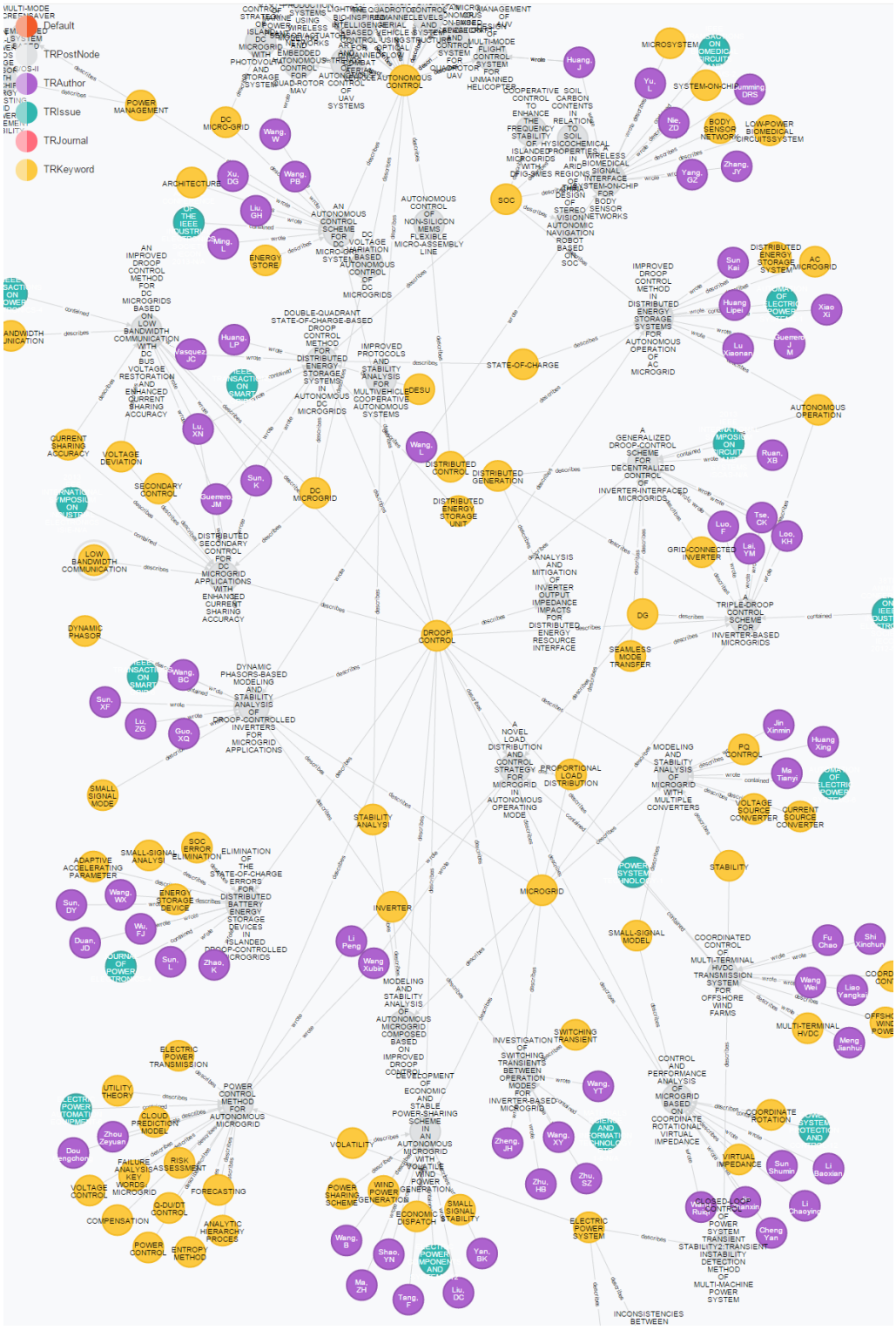
Det finns dock ingenting i det experiment som genomförts som tyder på att detta leder till några användbara resultat. Termerna i fråga används, dock inte i den kontext som det var tänkt och ibland är de dessutom i sig termer som används som facktermer inom olika ämnesområden. Möjligen skulle dessa termer om de i stället användes som ordkombinationer som kan tyda på något av intresse fungera bättre. Det vill säga; istället för *breakthrough* kanske ordkombinationerna *recent breaktrough* eller *breakthrough in*. Några av de termer som lagts i ordlistan var även alltför generiska, till exempel *major*, *growth*, *future*, m.fl.

När sökordet väl används i den kontext vi önskade så refererade det till tidigare historiska genombrott snarare än att artikelförfattaren påstår att den egna artikeln på något vis skulle vara det. Ett exempel är ur inledningen på en artikels sammanfattning som refererar till tidigare upptäckter:

> *”A single lineage of Nicotiana benthamiana is widely used as a model plant (1) and has been instrumental in making revolutionary discoveries about RNA interference (RNAi), viral defence and vaccine production.”*

Å andra sidan, om det hade fungerat, så skulle inte metoden att tagga poster baserat på detta varit relevant, då det helt enkelt räckt med att göra en enkel sökning på termerna direkt från början.

Generellt kan det dock inte uteslutas att ordlistor kan vara väldigt användbara, men för att nå önskad effekt behöver de användas tillsammans med metadata.



Figur 3. Gruppering av artiklar och författare.

4 Sammanfattning

Projektet har ännu inte haft tillgång till en datakälla som medger en automatiserad inhämtning av större mängder information. Därför har vi hittills bara kunnat undersöka datamängder mycket mindre än de vi skulle vilja analysera. Den datamängd vi har tillgång till har dock varit tillräcklig för att se ett antal tydliga utmaningar vad gäller att hantera nyckelord.

Att gruppera ihop poster baserat på vilka nyckelord de använder kommer att kräva en betydande insats om detta ska kunna fungera robust. Till detta ska dock läggas att de flesta problemen är tämligen välkända och att det finns gott om redan existerande metoder för att hantera dem, så en del av detta arbete kommer att bestå i att välja ut och tillämpa lämpliga existerande lösningar. Det tillkommer dock viss problematik då nyckelord kan vara extremt specialiserade och det måste undersökas hur väl lösningarna fungerar på just denna typ av ord och det är sannolikt att en del modifikationer måste till. Därtill så måste även lösningarna fås att fungera väl ihop och tillämpas i rätt ordning.

Till dess att automatiserade metoder kan användas så har vi byggt en prototyp där en mänsklig användare kan gruppera ihop nyckelord för att se aggregerade resultat kring hur ofta nyckelordet använts och tillsammans med vilka andra nyckelord och i vilken utsträckning under den aktuella perioden.

Nedladdade poster har även lagrats i en grafdatabas för vidare experiment kring hur denna typ av teknologi kan användas, både för att visualisera datat i en explorativ analys och för att se om det kan vara till hjälp för att lösa någon del av problematiken kring gruppering och kategorisering av poster.

En hypotes att det skulle gå att flagga för särskilt intressanta artiklar med hjälp av en specifik ordlista har testats men funnits fungera väldigt dåligt i nuvarande form. Det innebär inte alls att det nödvändigtvis är en dålig metod att använda ordlistor för kategorisering. Men just denna ordlista har inte visat sig vara användbar för önskad effekt.

4.1 Förslag på arbete 2016

Att utveckla en fungerande metodik för horizon scanning och att implementera den i ett datorsystem är ett arbete som kommer att löpa över flera år. Båda dessa uppgifter är var för sig omfattande.

Metodutvecklingen som beskrivs i denna rapport är endast ett första steg i riktning mot en metod som när den är färdigutvecklad och implementerad kan leverera de önskade resultaten. Implementering av metoderna är som beskrivits i kapitel 3 ett omfattande arbete som kräver lösning på många olika problem.

För att komma framåt är det helt nödvändigt att gå fram iterativt där de tre arbetena med (*i*) utvidgad metodutveckling, (*ii*) implementering av metoderna i ett datorsystem och (*iii*) metodtestning med detta datorsystem, varvas i en evolutionär utvecklingsprocess med metod- och systemutveckling i flera varv (En *Utveckling–Implementering–Test*-loop i flera varv). Att i en traditionell approach tro att man först kan utveckla en metod och därefter implementera och testa denna kommer inte att nå framgång.

En slutsats från årets arbete är att fokus under 2016 bör ligga på en samordnad verksamhet bestående av: (*i*) metodutveckling, (*ii*) implementering och (*iii*) metodtestning. Att genomföra en första horizon scanning med det verktyg som under 2016 fortfarande kommer att vara under utveckling bör vara en mindre omfattande aktivitet i projektet som kan provas mot slutet av året.