Little Scientist, Big Data Information fusion towards meeting the information needs of scholars

Nasrine Olson

Swedish School of Library and Information Science (SSLIS), University of Borås, Sweden. E-mail: nasrine.olson@hb.se

H. Joe Steinhauer

Informatics Research Center, University of Skövde, Sweden. E-mail: joe.steinhauer@his.se

Alexander Karlsson

Informatics Research Center, University of Skövde, Sweden. E-mail: alexander.karlsson@his.se

Gustaf Nelhans

Swedish School of Library and Information Science (SSLIS), University of Borås, Sweden. E-mail: gustaf.nelhans@hb.se

Göran Falkman

Informatics Research Center, University of Skövde, Sweden. E-mail: goran.falkman@his.se

Jan Nolin

Swedish School of Library and Information Science (SSLIS), University of Borås, Sweden. E-mail: jan.nolin@hb.se

Abstract

With increasing numbers of scholarly publications, and multiplicity of publication-types and outlets, overviews of research fields have become a challenge. We bring together bibliometric methods, information retrieval, information fusion, and data visualization within a new project, *INCITE - Information Fusion as an E-service in Scholarly Information Use*, with the aim to develop improved methods and tools addressing emerging user-needs. In this paper we report on ongoing research within that project. (a) We elaborate on a qualitative user-study in which the emerging needs of researchers in the age of big data are explored. The study is based on interviews and dialogue with seven scholars at different academic levels. Data analysis was informed by adaptive theory, in accordance to which iterative pre-coding, provisional codes, and memo-writing were used to reach a more abstract level of analysis. A number of challenges related to the multiplicity of information sources and extent of data were identified including difficulties in keeping track of all the relevant sources; the inability to utilize extensive sets of data being taken for granted; and using data reduction strategies that at times go against the scholar's own ideals of scholarly rigor. In analysing these difficulties, we have identified potential solutions that could facilitate the process of forming overviews of different research areas. (b) An example of such a solution is presented, which is builds on the Dempster-Shafer Theory and is designed to allow for interactive individual ranking of information sources in the process of a coordinated search across different information sources.

Keywords: e-services, information behaviour, bibliometrics, information fusion, big data, research area overviews.

Introduction and background

In this paper we report on ongoing research related to provision of e-services to scholars. Traditionally, there has been a disconnection between qualitative and quantitative approaches. We bring together both of these where project members with diverse but complementary strengths in INSU (information needs, seeking, and use), bibliometrics, information fusion and visualization join forces.

According to a report by the European Commission (2008: 51), the number of researchers in EU-27 in 2006 was listed as 1.33 million, with an annual increase of 3.1%. This was given in full time equivalent; the number of actual individuals goes beyond this. The numbers given for the US were higher and considering the countries in the rest of the world, the total number becomes rather substantial. Each of these researchers, regularly or at times, is involved in accessing scholarly communication data, making sense of and forming overviews of research fields. This is commonly a *time-consuming* and *costly* process. We plan to gain further insight in this process with the aim to facilitate and improve this scholarly practice.

Advances in digital technologies have contributed to increased production of data and new strategies for collecting and managing information. This has given rise to the advent of massive and complex data sets, which go beyond the capabilities of common software tools, and are commonly referred to as 'big data'. The definitions of this term are varied; for some the size of data (in terms of measurement units such as Exabyte) is a main issue, while for others it involves broader aspects. Researchers at the Oxford Internet Institute explain their view of 'what big data is' as follows:

> Our working definition is that they are data that are unprecedented in scale and scope *in relation* to a given phenomenon. In other words, data that represents a step change in how a field or discipline is able to address social science questions. (Meyer, Schroeder, Taylor, 2013 – emphasis added)

Here big data is defined as a relative concept where what can be seen as big (or not) depends on the context. Others have highlighted three related attributes of information assets in conjunction with other requirements by sating:

> Big data is high-*volume*, high-*velocity* and high-*variety* information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. (Gartner IT Glossary, 2013 – emphasis added)

Regardless of how it is defined, big data has become a phenomenon of our time and in relation to it, scholars have become faced with new opportunities and challenges. While processing of big data in terms of capture, storage, management, search, cross-referencing, analysis, sharing, transfer, and visualization requires technical solutions, it is also imperative to investigate the information needs and behaviour of scholars in the face of the new challenges and opportunities.

In this study, the focus is placed on scholars' endeavours in forming overviews of research fields. Although the size of data sets used by scholars in this pursuit may not yet reach millions of terabytes, we are witnessing an exponential increase in the volume of scholarly communication in different formats. The volume and variety of data that can be used to form overviews of different research fields have increasingly become of magnitudes that go beyond the scale and scope of common tools available to everyday scholars.

Whereas a literature review in a near past would have involved reading and analysing a few hundred articles, today such an endeavour becomes a challenge when the volume of relevant publications extend to thousands, or tens of thousands of items. Thus, the production of scholarly *literature reviews* or *overviews of research fields* has become a major challenge, particularly in multidisciplinary fields where publications from many different areas may be of interest.

The use of publication indicators and bibliometric measures as proxies for quality, and in turn as methods of assessing centrality of published literature has been shown to be marred with a number of problems (Borgman, 2007, 63ff). There are many issues associated with assessing relevant works by means of citation frequency or impact factor measures. First, using highly regarded publications based on the sources' reputation as measured by its (journal) impact factor is questionable since average performance of a publication does not indicate that an individual paper will fare well in terms of received citations (Seglen, 1997; Cronin, 2005). Furthermore, citation frequencies aggregates cumulatively, meaning that recent research always is at its disadvantage in comparison, and at the same time citation frequencies are highly skewed, e.g. in the way citedness is distributed over time (van Raan, 2006). Other issues relate to differences in publication as well as citation practices that present themselves in the problem of comparing sources to each other within and between different research areas.

Multiplicity of publication outlets, including a flourishing of open access journals and depositories not included in citation databases, complicate the situation further. It is not always possible to identify the most relevant sources of information that should be observed. The current labour intensive identification, evaluation, analysis, and mappings are no longer adequate.

To address these challenges we have witnessed the arrival of various data-mining, content analysis, and visualization tools which can be used in algorithmic analysis and visualization of bibliographic data.

Nevertheless, problems remain. Consequently, we have initiated a project, INCITE, in which we examine the current approaches and consider development of new improved methods and tools that can be of service to researchers. In the project, we address two major challenges that confront any scholar. The first challenge involves a cluster of issues including validity, quality, uncertainty, and usability of scholarly communication data. There are many problems with data integrity, duplicates, name ambiguity, and non-standardized formats. The second challenge is to put scholarly communication data to optimal use. We here see a unique opportunity in utilizing techniques and methods from the research field of information fusion (IF) (Liggins, Hall, & Llinas, 2009), where much research has been performed regarding decision support for different types of application scenarios. In particular, we anticipate that the IF methods utilized for building predictive models and handling different types of uncertainties may provide a novel and fruitful perspective on scholarly information use. One important initiative in this direction is to evaluate whether the methods for combining uncertain information, typically utilized in the IF domain, could model different types of certain and uncertain relationships between bibliographic items from various structured and unstructured sources in order to support information retrieval and use in the scholarly process.

Accordingly, the overall aim of the INCITE project is to evaluate existing procedures in data access, analysis, and visualization and to construct improved methods and tools based on a combination of information retrieval, bibliometric methods and information fusion methodologies that can be used in analysis, visualization, and interpretation of large quantities of data to support researchers in their day to day information use. The focus is placed on the production of overviews of research, especially in multidisciplinary fields in which the potential relevant items are too extensive to be managed by human reviewers.

Based on this background, in this paper we present (a) an interview-based user study, which was conducted to inform the follow up activities within the INCITE project. We then present (b) Interactive Individual Ranking as an example of the types of solutions that we are investigating within INCITE.

(a) The User Study – Introduction

Studies of researchers' information behaviour suggest that these differ widely between disciplinary categories. Some distinctions between broad meta-categories such as science, social science and the humanities tend to note that while scientists relate more to journal articles as their primary literature, humanists, on the other hand use books and archives to a higher degree, while social scientists also rely on institutional data (Case, 2007) as their primary resources. This is further emphasized by a JISC meta study of twelve user behaviour studies. Their conclusion was that disciplinary differences in search behaviour prompts libraries and other service providers to gain the ability to serve many different constituencies (Connaway & Dickey, 2010).

In terms of temporal factors, researchers generally use literature of recent age with a majority of the read literature being less than two years old (Tenopir and King (1998), cited in Borgman, 2007). The same authors also found that the time spent on searching and downloading articles roughly doubled between the years 1984–2000, a period during which the manual practices of finding literature turned into digital downloading and printing (Ibid.).

The fast web-based information searches, and the incorporation of bibliographic databases, digital archives and institutional sources within the web, suggest that differences between these sources are on the brink of levelling out. The JISC report speaks of finding ways of providing seamless access to resources, arguing that providers must be able to accommodate different needs due to changing user behaviour (Connaway & Dickey, 2010, p. 32).

An early study of user queries on the web (Jansen et al, 2000) found that in contrast to users of traditional information retrieval tools, web searches were found to include a low use of advanced search techniques (such as Boolean operators), instead having a rich variability of unstructured search terms. This implies that there was a need for new types of interfaces and methods to create term lists and indexing results (Jansen et al, 2000, p.226).

While informed by such findings, we conducted a new user study so that we are up to date with the current situation and in order to examine the potential relevance of *big data* for information behaviour of scholars. That is, we wanted to examine whether the increasing number of publications and available material (volume); diversity of publication types and outlets (variety); and the speed of production and manipulation (velocity) has had a bearing on the needs and behaviour of scholars. Furthermore, we wanted to focus on scholars' information needs and search strategies only in relation to *two specific tasks* of forming an overview of a research area and writing literature reviews.

Purpose of the study

The overall aim of the study is to gain an insight into the information needs and information strategies of scholars in the light of ever-increasing information volumes and types. The overall research question posed is:

> In what way, if any, has the availability of increased volume of information, multiplicity of sources, and emergence of new data types had a bearing for scholar's information behaviour in their processes of

forming an overview of a research field and or *writing literature reviews*?

To investigate this, a number of sub questions were formulated, all related to scholars' use of information in the process of conducting the mentioned two tasks:

i. Which information sources are used/ prioritized?
ii. How are new upcoming sources and publications identified and to what extent are these accessed and used?
iii. What amounts of documents and bibliographic data are typically accessed and or reviewed?
iv. What are the methods and tools used when faced with large amounts of data/ publications?
v. How is prioritization done, if at all?

Methods, materials, procedures

This paper reports on on-going research. For the purpose of gaining a better understanding of scholars' information behaviour, the use of interview technique was deemed to be appropriate (see e.g. Case, 2007; Denzin & Lincoln, 2005; Silverman, 2005; Yin, 2003). So far, seven interviews have been conducted with scholars from seven different countries, two of whom had English as their native language. The participants comprise of three professors with extensive knowledge of their fields and numerous well received publications. Of these, two hold the position of scholarly journal editor. Two participants are seasoned researchers with several years of postdoctoral research activities and numerous publications. One participant is a PhD student at an early to mid-stage of completing the PhD programme. The final participant is a researcher / expert bibliometrician whose current role involves supporting other researchers with various bibliometric studies.

All these study participants have had a background in, or are currently closely associated with, the field of Library and Information Science (LIS). The assumption behind this choice was twofold. First, LIS is a multidisciplinary field; hence an overview of a topic of interest may involve knowledge of, and access to, publications from different fields. Second, it was hoped that by being from the field of LIS, the participants would be very familiar with a variety of relevant information sources and have a solid knowledge about different ways of accessing and making use of potential sources. Further interviews with scholars from other fields are planned.

Prior to the interviews, the objectives of the INCITE project were discussed with the participants. The semi-structured interviews each took from 45 minutes to over one and a half hours. The interviews were recorded and transcribed. In several of the instances, the interviews were followed by looking at actual examples from participants' related experiences in forming an overview of a new field, as well as studies and visualizations conducted by the interviewer that involved larger data sets. In looking at these examples, a number of problems related to search methods, data access and visualization were discussed. These were used to prompt comments from the participants. If new information beyond what was said in the interviews came forward, notes were made and included in the data analysis.

Data analysis was informed by adaptive theory (Layder, 1998). A brief provisional coding was conducted at the time of transcribing the interviews. The recordings were listened to and transcriptions were read on multiple occasions and at each time the allocated codes and memos were revised and cross-referenced. Iterative coding and memo-writing were used to reach a more abstract level of analysis. The qualitative data analysis software AtlasTi was also used to facilitate the coding and analysis process.

Findings

A basic assumption underpinning the INCITE project is that *accessing and reviewing publications and forming overviews of different research areas are integral parts of the scholarly practice*. In the dialogue with the study participants, not only did we find grounds for this, but a nuanced variation of the goals for such efforts emerged. As part of the objective with a literature review, one participant, for example explained "I think it's necessary in order to realise that we are not reinventing the wheel all the time" and so that one does not address "something that has already been studied tonnes". Another comment was "I need to start with a little bit of sense of overview at least of what's going on there". But there was a variation in reasons expressed, such as "well I do the literature reviews because it is expected", or "I want to position myself within a scholarly discourse", and "we know that there are some rules and these are probably silly, maybe they are, but this is the only way for us in order to be accepted". The different reasons that were imbedded in the discussions could be summarized as follows. (a) Environmental scanning: with aims such as getting informed about a research field and new emerging topics; keeping abreast of fast evolving areas; and getting a sense of what one's research community regards as valuable or of importance. (b) Intellectual and creative work: with aims such as identifying interesting gaps in earlier research; positioning one's work within a field of study; to find supportive or contradicting evidence for one's own research findings ideas, or writing; and to avoid reinventing the wheel. (c) Meeting the norms: the aims here include getting to know who one should cite in order to get accepted by the research community; to avoid being seen as a newbie; so that one's writing gets accepted for publication; and to meet the expectations and play the academic game.

The dialogues with participants, therefore, lent support to our first basic assumption. We also examined a second conjecture. In the recent times we have witnessed major advancements in the digital communication technologies

which in turn have led to the advent of new data types, upcoming information sources, as well as emergence of novel research areas. Therefore, a second assumption associated with the idea of the INCITE project is that *the combination of emerging new phenomena, and diversity of information sources, as well as the sheer size of the available information may lead to difficulties in locating, accessing and processing the data required for forming overviews of different research fields.* We also found support for this assumptions in the interview data as a nuanced picture of potential challenges materialized as described below.

All the study participants showed awareness of a multiplicity of information sources and channels present on the information landscape. While the preferences of information sources varied among the participants, that is, although a source that was valued highly by one participant could well be described as less suitable or relevant by another, all of the participants to a lesser or greater extent had accessed and used multiple data types and sources, (some to a very advanced and extensive level). Depending on how one chooses to count, around 40 different information sources/resources were named by the participants. These could be categorized as: freely available search engines, databases, open access repositories, social networking sites, other web-based resources, printed sources, and human recommendations. In the dialogues that took place, one could find a pattern emerging where the expressed or implied challenges would be associated to two different types of situations.

Complexity related to variety and velocity – First, *related to recent research or new research topics*, a lack of relevant publications in scholarly journals and indexed databases was highlighted. While much research may be conducted, the publication of results in scholarly journals and indexed databases lags behind. Therefore it has become important to access alternative sources of data that might allow access to new discussions or findings. One participant explained about a topic of interest by saying, "the research on this new phenomenon is very limited and much about it is written in press articles, promotional material, articles by practitioners, and blogs" going on to indicate a need to access these and other alternative sources. This need could also be exemplified in the following comment by another participant; "if you try to develop a research project it should be an area where there is not much earlier research, and then it's necessary to look at blog writings and stuff people have in Facebook and things like that, which are not sort of reviewed; or alternative journal sources, but which are as up to date as possible. And which might have a very new insight, or finding of a viewpoint, because I think academic research lags behind a lot; if we try to keep up to date with development of networked environments via scientific literature it's not going to work".

It became evident that identifying and accessing the many different alternative sources of information, which might prove to be relevant, is however, a challenge. One participant, for example, while talking about research on a new phenomenon, said "[it] is so recent that it's transforming really, as we speak. That most of the documents that we have are press articles and blog entries. But simply as researchers, don't manage to keep up with the pace of transformation". Another participant who had mentioned that access to researchers' websites and blogs would be useful did not access these with the comment, "I seldom do that, but maybe that's a good idea. But I wouldn't be sure how to find them. Yeah, yea; those researchers that I know by name within my own field, those websites I find easily of course. But otherwise, there are.., in another instance I wouldn't know who to look for." Yet another participant, referred to the need to learn webometrics for analysing information on different websites, blogs, etc. explaining, "because it's..; it's also a reasonable delay before something gets indexed in subject databases or multidisciplinary citation databases. So time is a disadvantage. So I would [if the participant could] probably look at the web somehow and collect data from the web." Similarly other participants also discussed the need to access alternative sources and data types indicating a difficulty in knowing where to look or how to get access to these.

Individual strategies – To access the up-to-date information and alternative sources, participants had, therefore, formed personal strategies; for example one of the participants, would keep the calls for papers and or conference participations to follow up after the calls' deadline. This participant would access the conference programme and after the conference date would look for the potentially interesting presentations on researchers' websites and blogs as well as their academic social media accounts, or conference websites. Other strategies included accessing practice papers; using various forms of news alerts for capturing the reports by research centres and market surveys by big survey agencies; periodic searches on key-researchers or project websites; joining mailing lists; periodic random searches on different search engines; following related debates in mass-media; attending seminars of potential relevance; building network of contacts and receiving tips to then be followed up by snowballing and so on.

Still, much of the findings seemed to be seen as random and often serendipitous. A respondent described that an open depository related to a topic of interest was found randomly; or that the respondent had by chance got to know about "some empirical work going on in Europe" which was then followed up. This respondent also provided other examples of random discoveries which had proved to be of much interest and relevance for the respondent. Related comments could be exemplified by: "it's very serendipitous also, it's not just a linear process.

It's just – ok, I start from here and I jump there – a bit like that", or "sometimes I also find the things completely unexpectedly" or "this is another one, this is from a journal and this is probably, I also found a little bit unexpectedly" or "I don't know, I just came across this one. And that was an excellent, excellent, excellent discovery".

Challenges remain – What was indicated in several different forms was *the inadequacy of the traditional methods and tools to help find and access the relevant sources and or help form an overview of the topic in hand*. In relation to the inability to use traditional bibliometric methods in forming an overview of new topics, one comment was, "bibliometrics also require certain period of time to accumulate citations or papers or whatever". Another similar comment was, "new emerging topics, interdisciplinary research, new ideas that are not really communicated through standard modes of communication; like journal articles or books, where you have open repositories for input and output and so; this new media is not covered by bibliometric. […] So you probably have to use web resources in another way; blogs, links, between web pages etc".

Although the participants were aware of different potentially relevant sources, *the task of identifying and including them in a systematic search seemed to be a challenge*. For example, although a participant discussed the relevance and importance of a number of sources, when asked if those sources are included in the information collection strategy, the response was, "not always, I must admit, not always. There are a lot of sources that are not… I forget about them, I don't think about them. It's not that I don't want to include them, simply I don't think about them".

Accordingly, the first main problem identified in association to emerging new research areas was the lack of tools and services that would facilitate a systematic and coordinated effort in identifying and accessing the relevant sources.

As in these examples, the participants most frequently related the use of alternative newer sources to recent research or to studies of new phenomena. When traditional scholarly publications are available, those are preferred. This could be seen in an example of a participant who had earlier indicated the centrality of blogs and other new media in relation to research in new areas. When asked whether the participant had a good way of bringing together those types of data, the response was, "definitely not; and honestly I think we must be very, very, very careful. When you want to publish in selective journals, you know, they are traditional; they are conventional. When they see that you are referring to things in blogs or ..,'ooh!!'[gesturing a negative response presumably by the reviewers], there is a kind of status; okay? All the scholarly works carry a status and are worth mentioning, the rest, mm, don't look very nice". This brings us to the challenges experienced in relation to topics and research fields that are of enough age to have been addressed in scholarly publications in the traditional sense.

Complexity due to volume – Second, therefore, in relation to established and especially multidisciplinary topics, as expressed by a participant "the main problem is not that I don't have many references, or literature", in these instances the problem is rather "that I have too much". *Accordingly, the main problem identified in relation to established fields of study was the challenges brought forth by the huge size of relevant publications.*

One challenge relates to the difficulty in identifying all the fields and disciplines (and related journals) in which a topic of interest may have been explored. As one participant who had found it necessary to identify and use publications from different fields explained, "because there are contributions from people from different disciplines so I..; and also I'm a little bit at the interface myself, I don't consider myself very neatly positioned in any particular boxes so I use contributions from [several different fields]". The searches for this participant would start with a common search engine, known journals, and then snowballing, describing "then from there I see that probably there are other journals, other resources that I didn't know of that they can have some interesting stuff for me" indicating that at times this leads to discoveries in other fields than originally were imagined.

When it came to the volume of the data, one participant, the bibliometrician, regularly accessed and processed huge sets of data. The other participants, however, seemed to take it for granted that access and processing of big data sets were not possible for them. They indicated that a comprehensive coverage of the relevant material is not possible with comments such as "that's impossible, I try to get the central information" or "when you are a beginner researcher you don't know where to stop. People don't teach you. Because you have this idea that you should cover as much as you can, this is completely impossible." One participant said, "I've never been concerned with preparing something which is comprehensive, exhaustive – this is something that I cannot do." In another comment a participant said, "I am very pragmatic, you see. I have a limited amount of time, right, I have a limited amount of time, I want to quickly discover things of interest, I don't want to discover *every*thing of interest."

While the typical magnitude of the data successfully accessed and analysed by the bibliometrician was large (e.g. in one instance 19 million references and around one and a half million documents), a typical number at each instance by the other participants was a lot more modest (ranging from around 20 articles to several thousand publications).

The search and the required reading were often described as a very time-consuming process. For example,

a professor who through participation in a collaboration project had been introduced to a new research area had identified only some of the key books in the new area for reading in order to get an overview of the key ideas. This professor indicated that the reading of those books took a whole lot of time, including a whole summer vacation, without being able to cover it all. Another researcher discussed the way some researchers account for a rigorous systematic search in their publications and added "this is not what I can do, because it takes a long time and it takes more than one person." This participant elaborated further, "when you don't have much time, what I do instead and what I think a lot of other researchers do, we are not that systematic, we identify a little set of literature which we consider relevant for our research […] and then you use those." The limitations in time and financial resources for systematic searches, therefore, were highlighted time and time again with comments such as "there isn't that time anymore, unless you have money". When a systematic extended review is not seen as feasible, this is dealt with in different ways. One participant explained, "so I try to find an elegant and very nice and acceptable formulation when I write my papers to make people understand that yes, I did some literature review, it's not comprehensive. So I try to see what people usually write in papers when it comes to this and I found that a lot of people, much more authoritative than I, think like: this is a very short literature review so no ambition of being comprehensive, it just covers the most recent literature you can identify in a kind of a period of time. This is what I do usually. So I perform selective, short literature reviews. This is what I do right now in my research. I was a bit more comprehensive when I was a doctoral student, but I had more time at that point and it was probably more expected."

Individual strategies – Several participants were in agreement that more systematic and rigorous searches are only feasible during one's PhD studies, when one has the time and when this is expected. At other times, in response to common restrictions, reduction in the volume was seen as necessary. Often when a search would return voluminous results, only the first few items or pages of results were considered. At times, a search would be concluded just as soon as a small number of relevant items were found. For some participants, relying on human recommendations was a core strategy and was identified as a preferred trusted means of finding documents of interest. Strategies for reduction of large volumes of data to manageable sizes included modifying the search terms; delimitations by date of publication, publication source, and document type. The refinements were done in order to include mainly items that are "considered within the scholarly community" as important pieces of literature, items that are "somehow recognized as really belonging to the field", papers produced by organizations that "are recognized as well-reputed and well-established and authoritative", or documents by authors who are regarded highly. The idea often was to choose those items that could be recognized as significant and could show the selection to be justified. For this, often the items selected would comprise of those top ranked by the search engines or most highly cited items as identified by a citation database or items highly recommended by an expert, and so on.

Being cited highly was a recurring response in different forms. Some of the participants who indicated the high level of citations as a quality measure for selection in parts of the interview were somewhat in contradiction to what they said in other parts of the interview. For example, in case of one participant, when asked whether in selection any attention is paid to the number of citations, the response was, "not at all", and elsewhere in the interview this participant commented "sometimes people do a lot of honorary citing that I don't like, because not all articles written by respected or well acknowledged scholars on a specific topic are the most important articles on that topic", still in another part of the interview this participant indicated high level of citation to be a criteria for selection at times. This could be interpreted to exemplify the scholarly ideals that are not easy to fulfil given the limitation of current praxis and norms. This could be better illustrated in another example where a participant demonstrated informed awareness of the limitations of the citation practices. Still this participant would base relevance ranking on the assumption that these measures "are of a good standard", indicating that given the current situation and available information this is "the best assumption" that one could have.

System shortcomings – Even those who had the know-how and resources to invest in this task did not find the processing of large sets of data an easy task. One participant, while talking about an established database, mentioned, "they have other problems. They index a lot of rubbish, for instance. I looked at some of my papers that are not really very interesting, not peer-reviewed or nothing, but they still index it. So there are a lot of garbage too". Beyond the quality of the data, several participants described the limitations of current tools in helping them beyond the selection. Although the participants were aware, and to various degrees took advantage of the search refinement possibilities offered by different databases and bibliographic services, when it came to huge data sets their needs for refinement went beyond the options offered. A participant, for whom theoretical discussions were the key interest, gave a reason for seldom using databases as, "you can't really search for theories". Another was interested in studies that had adopted a particular perspective in investigations of a particular phenomenon while using a particular method. This participant had also found it difficult to find relevant

items. Although one could refine the search (by different criteria such as date, language, research area, journal, etc.), and could use a sophisticated combination of terms to get close to what one wanted, this did not seem enough when huge sets of data were retrieved.

In the interview sessions, participants were shown examples of analysis and visualization tools which most of the participants found of interest. When for example one participant was asked if such a system would be of use, the response was "Yes! It would be. Definitely, absolutely. This is, I mean, you see, how can a person manually manage something like identify what's relevant when you have 11,000 hits or something." Beyond the facilities that are commonly offered in databases, there were other wishes for features that did not seem to be available in current systems. For example, one could not easily identify only those documents in which *definitions* of a term are provided. At other times one was interested in publications about terms that could have different meanings or be treated differently as a "sociological phenomenon" or "a technical subject matter". Just selecting the research area or journal did not help to refine the retrieved items to a satisfactory level. It was a wish that search facilities in databases could help with some sort of content management, for example for a system to "collect me everything that has been written and then put it into categories like..; well I would see if it's about fans, or if it's about tagging, or is it about something commercial or cultural criticism […] So if it had that kind of presentation or had some keywords that tells what their angle roughly is, then that would be very useful for me."

Some sort of content analysis was sensed to be done in freely available search engines, but the algorithms and reasoning behind the selection were opaque to participants. It was common knowledge that these search engines often return huge numbers of results in magnitudes of tens of thousands or millions. However, dealing with the full extent of the results produced by search engines was not of interest to any of the participants, as much of it was found to be of no relevance, with comments such as "I notice that a lot of garbage comes up that is absolutely not relevant for me". In these types of services, the participants typically viewed only a handful of the first pages of results with comments such as "so maybe in the first page, or in the first two pages I tend to find resources that look more interesting to me and then I give up, because I realise that then it's all completely irrelevant". Here the non-transparent, non-interactive algorithmic selections were seen as a shortcoming. It was not so that in every case the most relevant would end up on top, as one participant mentioned, "if you just have the patience to skim then you can find some really, really, relevant pearls", but as mentioned earlier, in many cases the time and resource restrictions defined the boundaries of what was included in the collection of items retrieved and reviewed.

Challenges remain – Therefore the problem identified is not a lack of seemingly relevant material, the problem is rather identifying the items actually relevant among the results found by the search engines and the lack of facilities to help analyse the large sets of findings in a meaningful fashion.

Accordingly (a) the sheer size of available material, (b) limitations in the available tools and methods for locating and accessing the material, (c) limitations in the available tools and methods for meaningful content analysis of the material, (d) limited time and resources and (e) scholars' knowhow are identified in this study as factors affecting the level of information accessed and used.

Visualization – Returning back to the content analysis and visualization tools, the level of familiarity varied, but most of the participants did not use these on regular bases. Except for the bibliometrician, who had qualified knowledge of some such tools and regularly used them, only a couple of the participants had, on some occasion, used simpler such systems. Two participants had commissioned production of bibliometric analyses, and or visualized maps of the research areas of their interest. All the participants except for one, however, found such tools and services of interest. The one participant who did not, elaborated, "I think you have to take into account that this notion of learning styles, visualizers verses verbalizers. You know if you are a strong verbalizer, you will never find that kind of picture of any value at all". However, this participant was one of those who on earlier occasions had used a simple visualization tool that had helped categorized the results into clusters of closely related items. Therefore, even for this participant analysis and visualization tools fulfilled a purpose and were seen as useful as long as the outcome was simplistic and self-explanatory. This participant continued, "when I want something I want specifics, I don't want totalities". Similarly, in several instances the example visualizations were found to be interesting and useful but they did not go far enough in their analysis or in meeting the wishes of the users.

Another stumbling block was indicated to be the learning curve required for their use and for interpreting the outcomes. Even a participant who had attended workshops with the system designers could not yet use the tools. This was despite the fact that this participant had found visualization to be particularly useful and relevant. Regarding the outcomes, the participant who did not appreciate visualizations as much as the others discussed the work required for deciphering the outcomes by saying, "does it tell me something intelligible right now or am I gonna have to work at it in order to discover what I want to know [laugh]? If I have to work at it I don't want to do it because I'm lazy, right. Which is the other factor that one has to take into account, how much effort is somebody going to have to put in to learn how to use these tools". For most participants, user-friendliness was

seen as a must especially in areas where the researcher users are not technically oriented. User-friendliness was not an issue for the bibliometrician though; there, other qualities were of more importance. As expressed by that participant, "the most important thing is that I understand mathematics and statistics on which they are based. Otherwise, the other stuff like which button to push, and what happens then, and what is practical, you learn by time, so to speak. And because I have used them so much, thousands and hundreds of times, so it's never an issue any longer. So, and I never bother about if they could be more practical or more user-friendly". For this participant the algorithmic transparency was very important in order to understand and ensure that the analysis is correctly done. This participant at times had excluded the use of some tools due to the black-boxed nature of the tools.

Wish list – Accordingly, a number of wishes were identified in the dialogues with the participants. These included a tool or service that could help users to identify potentially interesting sources of information. A tool or service that would facilitate coordinated searches in these different sources. As user preferences and information needs and circumstances vary from one instance to the next, such a system should be flexible enough to allow for inclusion or exclusion of the sources. It should also allow for the allocation of the level of importance to each source based on individual preferences. Furthermore it should be transparent and interactive to allow the user to modify the findings to fit individual needs. There is a need for ability to combine common selection criteria with meaningful content analysis options. That is, tools and services are required both for reducing the huge numbers in more meaningful ways in some instances, and / or for assisting in the analysis of the contents of huge sets in a more meaningful fashion in other cases. In relation to the latter, this is a final excerpt from the interviews: "Yes, with references of wishing something, technique, or method or theory; when you're doing bibliometrics, you're always looking at the tip of the iceberg. You look at the most frequent, the most central authors, papers, or journal of a field, but that doesn't tell you.., that doesn't really give you a measure or an understanding of what the whole field looks like. So if you look at co-citation analysis, most of the papers.., you look at one percent of everything in a subject area, 2%, or 5%. Of course you can download all the.., the whole subject field, but just even then only use 1% or 2 %. Because you're then.., in my.., in one sense it is reasonable to do that, because.., and you regard the rest as noise, right. But I would really be interested in a method or theory that could sort of visualize the whole field; "what is the total content and the total context of this field?". So, but the problem is you can't cluster the whole field because then you get associations, they are so weak that they are meaningless. So that would be really interesting. A brand new method of mapping the field without losing 90% of all the items. Thank you! (laugh)"

Discussion

Although no general trends can be identified based on this limited number of interviews, we can still discuss how the findings so far relate to the research questions posed.

In response to the first question (i) we found that a large number of information sources and search tools were used by the participants. These could be categorized as freely available search engines, databases, open access repositories, social networking sites, other web-based resources, printed sources, and human recommendations. The priority given to each of these varied from one person to the next, and based on the situation. The way new upcoming sources were identified and included in searches (question ii) varied as to the level of their sophistication. Some of the participants combined extensive mixes of strategies to keep updated with new relevant material. The amount of information collected and used (question iii) also varied considerably among the participants from just tens to millions of items. To deal with the large amounts of data, reduction (e.g. by date, number of citations, recommendations and other strategies) was common (question iv). The use of more specialist semi-automated tools in data analysis and visualization was not very common. In relation to the last research question (question v), most participants used various features offered by search engines, databases, and journals for some level of analysis and ranking, although problems with these tools and their function were identified. These problems ranged over, lack of algorithmic transparency, limited interactivity in the selection process, inability to indicate individual preferences, absence of a coordinative function, limited flexibility in automated analysis features, and inability to form a more comprehensive view instead of adopting a reductionist approach among others.

As shown above, previous research has identified a number of problems with the use of publication indicators and bibliometric measures as proxies for quality (Borgman, 2007). Some of the user-study participants showed informed awareness of such problems. Even so, in the absence of other means of adequately dealing with the sizable data, bibliometric measures were still commonly used in identifying the key resources and for the reduction of data to manageable sizes.

Similarly problems of data integrity were acknowledged by some of the study participants. Such problems were dealt with at times (mainly by the bibliometrician) by employing laborious time consuming manual manipulations, while at other times they were accepted as a fact of life and not dealt with.

Previous research (Tenopir and King (1998), cited in Borgman, 2007) indicated that researchers generally use literature of recent age. This was the case for several of

the participants. In our study we found the 3V-attributes of information assets (volume, variety and velocity), to be of relevance here. In response to challenges of volume, several of the study participants indicated that they use the age of publications as a way of reducing the number of documents that they use. When it came to newer phenomena and new research areas, it was mainly the recent scholarly publications (if in existence at all) that became of interest. When it came to the velocity aspect, the challenge was in the efforts to keep up. This was also associated with challenges related to the third V, i.e. the variety of different sources and data types. While sophisticated strategies were put in place to access multiple sources, one could not be assured that all items of interest are found, as new sources and types would emerge.

The study so far has been limited in its scope in two respects. First, it has only been based on interviews. Use of other methods such as observations, screen dumps, and journal writing may prove to be of value. Second, the number of participants and their field of study have been limited. Studies comprising participants from other fields of studies may shed light on new insights.

We intend to address some of these short comings in the continuation of the INCITE project. This study has, however, provided us with some indication of a number of solutions that would facilitate scholars' information behaviour in the face of big data. We have already presented the application of information fusion to the problem of author name disambiguation elsewhere (upcoming). In what follows we present an example of how we intend to address another problematic area as identified by this user-study.

(b) Example: Interactive Individual Ranking

As presented above, most of the participants in the user study expressed that a somewhat comprehensive search was not feasible given the time and resource restrictions. In the INCITE project, we are investigating ways of facilitating extensive searches for more meaningful and relevant results given the known restrictions, by taking advantage of improved automated techniques in content analysis and visualization. Meanwhile, we also investigate other solutions that would improve and facilitate the current routine practices of the scholars. The example described below is one such solution.

As mentioned, it was a common practice for participants to conduct searches across multiple sources. In some instances it was a common or desired practice to include a selection of the top items as ranked by different systems in the pool of their selected items. However, the preferences for sources and the values attached to each of those would vary from person to person, and also for the same individual in different circumstances. That is, while one person might prefer information sources *A*, *B*, and *C*, a second person might value sources *B*, *D* & *E*. Meanwhile, the preferred sources and value judgments attached to each of the sources may change for the same person given different circumstances (e.g., blogs are desirable and valued highly when searching for discussions of new phenomena, but are not seen as trust-worthy and are valued low when searching for established research topics). Another problem was to keep track of the different sources and to remember (or find time) to include them in the suit of sources to be accessed. Accordingly, the interviews revealed several difficulties experienced by the scholars. The two that we will address further in this paper are: (1) that the internal ranking procedures differ from source to source, which make it necessary to be interpreted in different ways, and (2) that the scholar has certain individual judgments towards the different sources regarding, e.g., trustworthiness, comprehensibility, perceived impact of the source. Depending on the purpose of the information search that the scholar is performing, each of these attributes might be regarded more or less positive or negative.

In this section of the paper, we provide an example of how information fusion can be used to automatize this process. This will benefit the scholar in three ways. (1) He or she will get a single ranked list of all the papers found by the different sources, taken his or her personal judgment of each source into account. (2) The automatized process is able to include far more items of the ranked lists than a scholar would be willing or able to do by oneself. (3) The search can be extended to more information sources than a scholar would be willing to search in manually. Thereby this approach will improve the scholar's ability to search within the rising amount of information available.

Background – Information Fusion

Information fusion (IF) (Steinberg & Bowman, 2009) is a research field where the aim is to combine information from different sources for the purpose of achieving an effective decision support for the task at hand. The research field can be roughly divided into two subfields: (1) *low-level IF* and (2) *high-level IF*, where the former typically focuses on data pre-processing and estimation of a singleton unknown state, whilst in (2) one is interested in combining all the estimations of these singletons to determine multi-dimensional, most often also more abstract, states, for the purpose of obtaining an understanding of the current situation.

One common theme to all fusion processes is that they rely on some framework to model, combine, and perform reasoning under *uncertainty*. In fact, reducing uncertainty by using multiple sources of information can be seen as one of the main goals of an IF-system (Bossé et al, 2006). Within these frameworks, e.g., Karlsson et al. (2011), the main mechanism to model uncertainty is to encode information as *pieces of evidence* (including *counter evidence*) with respect to the unknown state of interest.

Evidential Frameworks – *An evidential framework* (Karlsson, 2010) consist of (1) a mathematical structure that models uncertainty, denoted *evidence structure* and (2) a way to *combine* evidence structures to *a joint (fused) evidence*. There are many different theories such a framework could be based on, however, one can categorize these theories into two main groups, namely (1) *precise probability* (Bernardo & Smith, 2000) and (2) *imprecise probability* (Walley, 2000), where the distinction between these two groups lies in the evidence structure and in the combination schema. In the former group, one only allows for probabilities in a precise form, e.g., as in ordinary probability theory, whilst the latter one allows for imprecision probabilities, e.g., by specifying probability intervals. The idea behind imprecision is that by using a more general structure one can obtain a better model of the different uncertainties involved in the fusion process.

Dempster-Shafer Theory – In this section, we present one of the imprecise probability theories, namely *Dempster-Shafer theory* (Shafer, 1976), which we later will use for demonstrating our approach for individual ranking.

In Dempster-Shafer theory, also known as *evidence theory*, one models pieces of evidence by so called *mass functions*:

$$m(A) \geq 0$$

$$\sum_{A \subseteq \Omega} m(A) = 1,$$

where Ω denotes the set of possibilities for the unknown state of interest. Two different pieces of uncertain information, modelled in terms of mass functions m_1 and m_2, can be combined by using *Dempster's combination operator* (Dempster, 1969), defined as:

$$m_{12}(C) \stackrel{\text{def}}{=} \frac{\sum_{A \cap B = C} m_1(A) m_2(B)}{1 - \sum_{A \cap B = \emptyset} m_1(A) m_2(B)}$$

where $A, B, C \subseteq \Omega$. This *joint evidence* can then be used to calculate *lower and upper bounds on probabilities*, i.e., the *imprecision*, for a set A by:

$$\underline{p}(A) \stackrel{\text{def}}{=} \sum_{B \subseteq A} m(B)$$

$$\overline{p}(A) \stackrel{\text{def}}{=} \sum_{A \cap B \neq \emptyset} m(B)$$

which is the reason that the theory can be regarded as belonging to imprecise probability. Furthermore, one can obtain a single precise probability based on what is known as the *pignistic transformation* (Smets & Kennes, 1994):

$$p(A) \stackrel{\text{def}}{=} \sum_{B \subseteq \Omega} \frac{|A \cap B|}{|B|} m(B).$$

Lastly, if one has additional information about the reliability or trustworthiness of the sources then this can be taken into account before constructing the joint evidence by using so called *discounting* (Smets, 2000):

$$m^d \stackrel{\text{def}}{=} \begin{cases} \alpha m(A), & A \neq \Omega \\ 1 - \alpha + \alpha m(\Omega), & A = \Omega \end{cases}$$

where $\alpha \in [0,1]$ expresses the degree of reliability of the source (0 means completely unreliable and 1 fully reliable).

One important issue when using combination operators, such as in Dempster-Shafer theory, is that the information sources need to fulfil certain types of independence assumptions (Smets, 2007), and in principle information sources should base their ranking on different types of information/features. However, even though such assumptions are not completely fulfilled, good results can be obtained, c.f. naïve Bayes (Russel & Norvig, 2003).

Interactive Individual Ranking

We will illustrate our approach with an example. In order to keep the example easy to comprehend, we restrict it to three different information sources, *A*, *B*, and *C*, and look at the three top ranking papers provided by each source.

Firstly, in order to accommodate the scholar's possible individual judgement about information sources regarding the different attributes, such as trust, comprehensibility and perceived impact, the scholar needs to state the attributes, together with his or her individual values for them, only once as shown in Table 1 where the numbers are to be translated as 1 = low, 2 = medium, and 3 = high.

Attribute Source	*Trust*	*Comprehen sibility*	*Impact*
A	1	3	1
B	2	2	3
C	3	1	2

Table 1 Individual attribute assignment to information sources

Secondly, the scholar's search is run on all three information sources in parallel, which results in three ranked lists of papers as shown in Table 2.

Source Ranking	A	B	C
1	*a*	*b*	*d*
2	*b*	*c*	*a*
3	*c*	*d*	*b*

Table 2: Ranked papers for the three information sources.

The rankings can now be translated into the mass functions m_A, m_B, and m_C over the frame of discernment containing the four found papers, $\Omega = \{a, b, c, d\}$. How this translation is done can be regarded as a research topic in itself. We will here use a simple and intuitive translation where the first ranked paper receives the mass 0.45, the second mass 0.30, the third ranked paper the mass 0.15 and the remaining mass of 0.1 is assigned to the frame of discernment. When the situation occurs that there are more papers found than rankings available, as shown in this example, where four relevant papers were found, the mass of 0.1 is assigned to the paper that an information source has not found. After that, the mass function is renormalized. The result for the example is:

$m_A(a) = 0.43$	$m_B(a) = 0.05$	$m_C(a) = 0.28$
$m_A(b) = 0.28$	$m_B(b) = 0.43$	$m_C(b) = 0.14$
$m_A(c) = 0.14$	$m_B(c) = 0.28$	$m_C(c) = 0.05$
$m_A(d) = 0.05$	$m_B(d) = 0.14$	$m_C(d) = 0.43$
$m_A(\Omega) = 0.10$	$m_B(\Omega) = 0.10$	$m_C(\Omega) = 0.10$

Combining these mass functions with Dempster's rule of combination, we receive a new mass function m_{ABC}, which is the basis of the combined ranking of all found papers as shown in Table 3 in the second column.

Rank	General	Trust	Comp.	Impact
1	b	d	a	b
2	a	a	b	c
3	d	b	c	d
4	c	c	d	a

Table 3. Ranking after different attributes

In order to let the scholar decide after what attribute the result should be ranked, e.g. trust, comprehensibility, or perceived impact, the system will translate the scholar's individual attribute values into discounting factors. The mass functions are then discounted accordingly before the combination is done. Discounting after the attribute trust, we assign a discounting factor of 1 to the most trusted source (*C*), a discounting factor of 0.5 to the medium trusted source (*B*) and a discounting factor of 0.25 to the least trusted source (*A*). After that, the discounted mass functions are combined, which results in a new ranking, as shown in Table 3 in the third column. The same can be done for the attributes comprehensibility and perceived impact, with the ranking results shown in Table 3 in the fourth and fifth column, respectively.

Figure 1 provides an overview of all four papers with regard to each individual ranking attribute. From the figure it can, e.g., be seen that paper *b* is the one believed to have the highest perceived impact. Paper *a* is believed to be most easy to comprehend, but has the lowest perceived impact and paper *d* is highly trusted, but believed to be difficult to comprehend.

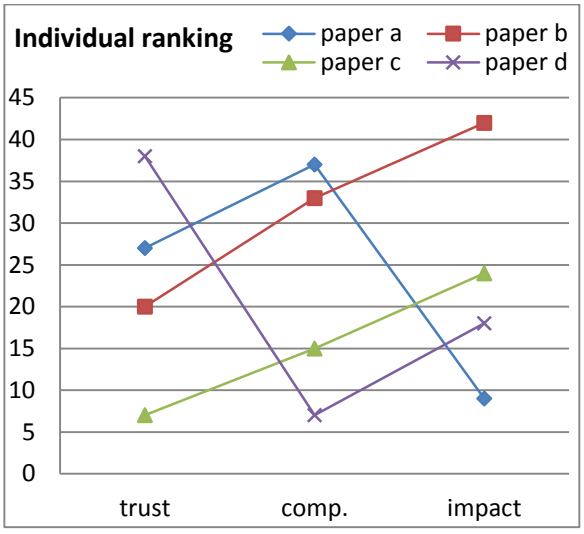

Figure 1. Comparison of papers after individual ranking attributes.

Further, the values of upper and lower bound on the probability can be used to provide more information about the certainty of the combined ranking. If, for example, two search results are combined, where their ranking differ very much, the result will be less certain as compared to when the two sources rank the papers equally. The uncertainty can be displayed by the lower and upper probabilities. Figure 2 shows the intervals for paper *a*, regarding the combined search results for trust, comprehensibility and perceived impact, depicted as vertical lines. The triangle on each line corresponds to the pignistic probability.

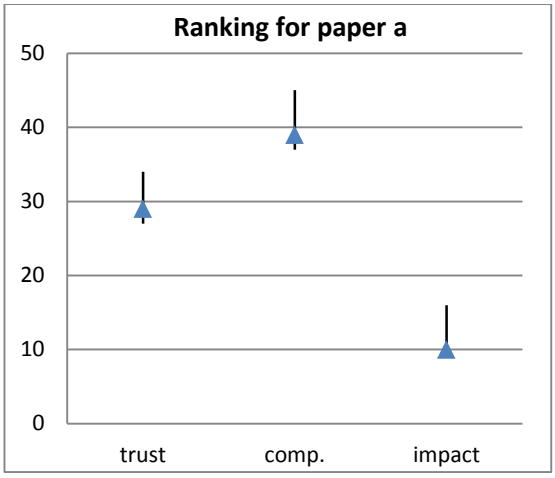

Figure 2. Combined ranking for paper a, showing the interval between upper and lower probability

Concluding remarks

The translation of the ranks into mass functions uses a very simple qualitative approach. Usually one does not know how the internal ranking process for each information source works. Therefore, it can't be known how close to each other the ranked papers are. There might be a huge gap between two closely listed papers or they might be (almost) equally ranked. The reliability of our approach would improve if the internal ranking of each information source would be known, so that the mass functions could be adjusted accordingly.

Also the discounting factors are simply provided by the three values low, medium, and high. If a finer resolution would be used, the reliability of the result should improve accordingly. This step is of particular interest when a larger number of information sources is included.

A further interactive feature to our approach would be to give the scholar the choice to manually input other sources of information that are not found on the net, e.g., a list of literature provided from a colleague.

This approach can take care of many information sources, and a vast list of ranked papers from each source, simultaneously. In order not to include every paper, a criterion needs to be implemented, either how many papers from each source are to be included.

ACKNOWLEDGMENTS

This research has been supported by Region Västra Götaland, and universities of Borås, and of Skövde. Thanks also to Gartner IT Glossary for permission to reproduce their definition of big data.

REFERENCES

Bernardo, J. M. & Smith, A. F. M. (2000). *Bayesian Theory*. John Wiley and Sons.

Borgman, C. L. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge and London: MIT Press.

Bossé, É., Guitouni, A., & Valin, P. (2006). An Essay to Characterise Information Fusion Systems. *Proceedings of the 9th International Conference on Information Fusion*.

Case, D. O. (2007). *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior*. Second ed. Amsterdam: Elsevier.

Connaway, L.S. and Dickey, T.J. (2010). The Digital Information Seeker: Report of the Findings from Selected OCLC, RIN, and JISC User Behaviour Projects, 61 pp. Report. *Higher Education Funding Council for England (HEFCE) and Joint Information Systems Committee London, UK Internet:* http://www.jisc.ac.uk/media/documents/publications/reports/2010/digitalinformationseekerreport.pdf (accessed 2014-05-12).

Cronin, B. (2005). *The Hand of Science: Academic Writing and Its Rewards* Lanham: Scarecrow Press.

Dempster, A. P. (1969). A generalization of Bayesian inference. *Journal of the Royal Statistical Society*, *30*(2), 205-247.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2005). *The Sage Handbook of Qualitative Research* (Third ed.). Thousand Oaks, London, New Delhi: Sage Publications.

European Commission. (2008). *A more research-intensive and integrated European Research Area: Science, Technology and Competitiveness, key figures report 2008/2009* (No. EUR 23608 EN). Luxembourg, Belgium.

Gartner IT Glossary. (2013). Big Data. Retrieved May 14, 2014, from http://www.gartner.com/it-glossary/big-data

Jansen, B. J., A. Spink, och T. Saracevic. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management* 36 (2):207-227.

Karlsson, A. (2010), 'Evaluating Credal Set Theory as a Belief Framework in High-Level Information Fusion for Automated Decision-Making', PhD thesis, Örebro University, School of Science and Technology.

Karlsson, A.; Johansson, R. & Andler, S. F. (2011). Characterization and Empirical Evaluation of Bayesian and Credal Combination Operators. *Journal of Advances in Information Fusion*, 6(2), 150-166.

Layder, D. (1998). *Sociological Practice: Linking Theory and Social Research*. London: SAGE Publications Ltd.

Liggins, M. E., Hall, D. L., & Llinas, J. (Eds.). (2009). *Handbook of Multisensor Data Fusion: Theory and Practice* (2 ed.). Boca Raton: CRC Press.

Meyer, E. T., Schroeder, R., & Taylor, L. E. M. (2013). Big Data: Rewards and Risks for the Social Sciences. Retrieved April 2, 2014, from http://www.oii.ox.ac.uk/events/?id=557

Russel, S. & Norvig, P. (2003). *Artificial Intelligence: A modern Approach*. Second Edition, Prentice Hall.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, *314*(7079), 498-502.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton, University Press.

Silverman, D. (2005). *Doing qualitative research: a practical handbook*. London: Sage Publication Ltd.

Smets, P. & Kennes, R. (1994). The transferable belief model. *Artificial Intelligence*, 66, 191-234.

Smets, P. (2000). Data Fusion in the Transferable Belief Model, *proceedings of the 3rd International Conference on Information fusion*.

Smets, P. (2007). Analyzing the combination of conflicting belief functions. *Information Fusion*, 8, 387-412.

Steinberg, A. N. & Bowman, C. L. (2009). Revisions to the JDL Data Fusion Model, in Martin E. Liggins; David L. Hall & James Llinas (ed.), *Handbook of Multisensor Data Fusion*, Second Edition (pp. 45-68). CRC Press.

van Raan, A. F. J. (2006). Statistical properties of Bibliometric indicators: Research group indicator distributions and correlations. *Journal of the American Society for Information Science and Technology, 57*(3), 408-430.

Walley, P. (2000). Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24, 125-148.

Yin, R. K. (2003). *Case Study Research: Design and Methods* (Third ed. Vol. 5). Thousand Oaks, California: Sage Publications.

Curriculum Vitae

Dr. Nasrine Olson is a senior lecturer (and researcher) at the Swedish School of Library and Information Science. She received her Ph.D. from Gothenburg University in 2010. Her research interests include the constructive nature of decision making, duality of structure and agency, mechanisms of control, social media, and societal implication of interactions with information technologies. She is a member of research utilisation group and a co-leader of the social media studies research program at the University of Borås.

Dr. H. Joe Steinhauer is an assistant professor in computer science at the University of Skövde, Sweden. She received her Ph.D. from Linköpings University, Sweden, in 2008. Dr. Steinhauer's main research interests are information fusion, artificial intelligence, qualitative reasoning and cognitive modelling.

Dr. Alexander Karlsson is a senior lecturer in computer science at the University of Skövde, Sweden. He received his PhD in computer science from Örebro University, Sweden, in 2010. Dr. Karlsson's main research interest is information fusion.

Gustaf Nelhans is a lecturer at the Swedish School of Library and Information Science (SSLIS) at University of Borås and has recently defended his PhD thesis in Theory of Science at the University of Gothenburg, Sweden. His research focuses on the performativity of scientometric indicators as well as on the theory, methodology and research policy aspects of the scholarly publication in scientific practice using a science and technology studies (STS) perspective.

Prof. Göran Falkman is an associate professor of computer science with specialization in interactive knowledge systems at the University of Skövde, Sweden. He received his PhD at Chalmers University of Technology, Sweden, in 2003. Prof. Falkman's main research interests are information fusion, artificial intelligence, visual analytics and decision-support systems.

Jan Nolin is a professor at the Swedish School of Library and Information Science at the University of Borås. He received his PhD in theory of science from the University of Gothenburg. Current research interests focus on the changing role of information practices and information institutions given the character of emerging information technologies. He is the co-leader of the social media studies research program at the University of Borås.