Title	A bibliometric study of video retrieval evaluation benchmarking (TRECVid) : a methodological analysis
Author(s)	Thornley, Clare V.; McLoughlin, Shane J.; Johnson, Andrea C.; Smeaton, Alan F.
Publication date	2011-12-19
Publication information	Journal of Information Science, 37 (6): 577-593
Publisher	Sage
Link to online version	http://dx.doi.org/10.1177/0165551511420032
Item record/more information	http://hdl.handle.net/10197/3038

Article

JIS

Journal of Information Science
XX (X) pp. 1-19

DOI: 10.1177/016555150000000
jis.sagepub.com
SAGE

A Bibliometric study of Video Retrieval Evaluation Benchmarking (TRECVid): a Methodological Analysis

Clare V. Thornley
Department of Information Studies, University College London, Gower Street, London, WC1E 6 BT, UK

Shane J. McLoughlin, Andrea C. Johnson
School of Information and Library Studies, University College Dublin, Belfield, Dublin 4, Ireland

Alan F. Smeaton
CLARITY: Centre for Sensor Web Technologies, School of Computing, Dublin City University, Glasnevin, Dublin 9, Ireland.

Abstract
This paper provides a discussion and analysis of methodological issues encountered during a scholarly impact and bibliometric study within the field of computer science (TRECVid Text Retrieval and Evaluation Conference, Video Retrieval Evaluation). The purpose of this paper is to provide a reflection and analysis of the methods used to provide useful information and guidance for those who may wish to undertake similar studies, and is of particular relevance for the academic disciplines which have publication and citation norms that may not perform well using traditional tools. Scopus and Google Scholar are discussed and a detailed comparison of the effects of different search methods and cleaning methods within and between these tools for subject and author analysis is provided. The additional database capabilities and usefulness of "Scopus More" in addition to "Scopus General" is discussed and evaluated. Scopus paper coverage is found to favourably compare to Google Scholar but Scholar consistently has superior performance at finding citations to those papers. These additional citations significantly increase the citation totals and also change the relative ranking of papers. Publish or Perish (PoP), a software wrapper for Google Scholar, is also examined and its limitations and some possible solutions are described. Data cleaning methods, including duplicate checks, expert domain checking of bibliographic data, and content checking of retrieved papers are compared and their relative effects on paper and citation count discussed. Google Scholar and Scopus are also compared as tools for collecting bibliographic data for visualisations of developing trends and, due to the comparative ease of collecting abstracts, Scopus is found far more effective.

Keywords:
Bibliometrics; TRECVid; methodology; research evaluation; visualisation; video retrieval

1. Introduction

This paper presents a methodological analysis and discussion of a bibliometric study [1], examining the scholarly impact of publications arising out of the series of annual TRECVid (TREC Video Retrieval Evaluation) conferences. The purpose of this paper is to provide a reflection and analysis of the methods used to provide useful information and

Corresponding author:
Clare V. Thornley, Department of Information Studies, University College London, London. c.thornley@ucl.ac.uk

guidance for those who may wish to undertake similar studies. As an informative case study of bibliometric evaluation, it is of particular interest to those undertaking scholarly impact studies of fields such as computer science in this case, or also in the humanities [2] where journal publication (and hence comprehensive coverage by the major bibliometric databases) may not be the prime source of publication output. It is also useful as guide to the problem of accurately defining and retrieving publications on a particular subject, as we examine all publications arising out of TRECVid developmental work and not just the actual conference papers. TRECVid started out as part of the large international benchmarking conferences known as TREC (Text Retrieval Evaluation Conferences) which provides test collections for information retrieval (IR) research groups to test and compare their techniques and it became an independent conference in 2003. The results of how particular techniques perform against the test collection i.e. how much better are they at retrieving the correct documents (in this case video shots) to match queries, provides an opportunity to compare alternative approaches and enable improved techniques to be developed.

TRECVid, therefore, is not a 'conventional' conference focussed on disseminating research outputs but can more accurately be described as an enabler of research and development, as without the large scale test collection, it would be very difficult to test and compare approaches as a means for improved techniques to be developed. Our study measured how important TRECVid had been in influencing and progressing the field of video retrieval by examining the publications arising from TRECVid (both actual conference papers and the other publications based on them) and their citation figures. We did this both in terms of the whole series of conferences and also compared, for a selection of participants, the respective citation rates of their TRECVid versus their non TRECVid papers. The results of this are presented in our previous paper [1] where a brief description of our methods is provided.

The methodological issues raised in that work, however, merit a separate and more detailed examination. In particular, we pursued some methods initially which we then abandoned and also found that some choices we became committed to could perhaps have been managed differently. We describe our learning and insights gained to facilitate the effective and efficient design and management of future related studies. Some of the issues based on database coverage are perhaps particularly relevant to computer and information science. The broad range of methodological questions raised, however, when evaluating and analysing a complex scholarly object (TRECVid includes, for example, authors, working papers, published papers, patents, experimental results and trends) are relevant to many disciplines. We illustrate examples based on the original techniques to evaluate, validate, and discuss methods chosen. The results of these examples will not be exactly the same as our original study as search results vary on a day-to-day basis and we also deliberately changed some of the parameters to observe the effects.

The following questions are addressed and discussed: why do a bibliometric study; how accurate are data collection methods; how accurate are data cleaning methods; how do different tools affect author and subject searching; how do different tools affect data collection for visualisations; what conclusions can be drawn to guide further bibliometric studies?

2. Why do a bibliometric study?

The relevance and value of measuring research impact through bibliometric studies, whether on the scale of an individual, research group or even in some cases entire nations, is certainly not unproblematic. It is difficult to get comprehensive and accurate data on publication numbers and even more difficult to get such data on how often these publications have been cited. Within the discipline of Computer Science one of the main issues has been proper recognition of the importance of conferences vs. journal publications [1, 3] where journal publications are easily evaluated through impact factors and the like, whereas the impact of publications in conferences is less easily measured. This has been a focus of attention for some time [4] but it is especially important in more recent times as reported by Freyne et al [5] and Franceschet [3].

The various difficulties of working in the computer science field can be overcome, to an extent, by comprehensive and careful research work but there are still issues of interpreting what the metric data means. If a particular publication or a set of publications by a particular group have received many citations, is this necessarily an indication of 'quality'? This is not a new question and not unique to computer science. Concerns about the meaningfulness of bibliometric indicators have been around for many years such as, for example, discussions about the correlation between citation rates other measures of esteem such as the Nobel prize [6] and more recently as discussed by Moed [7], De Bellis, [8] and Bar-Ilan [9]. Studies are emerging, which examine the 'popularity' of papers and authors in terms of citation counts, versus the 'prestige' of the citations accrued to those papers and authors [10, 11]. Within library and information science Li et al [12] compared 'expert judgement' of esteem in LIS, with citation metrics obtained through WoS, Scopus and Google Scholar, overall showing a strong correlation. Other research is beginning to examine factors such as geographic location, influence of open access publications [13] publication maximisation of research [10] and possible advantages for more senior researchers in terms of accruing citations. There is also the

problem of how one compares different individuals and/or groups, and whether one is fairly comparing 'like' with 'like', which can be particular challenge for small or niche research areas [14]. Some approaches to the issue of comparison for research groups include conceiving a research group as a 'virtual journal' and comparing its impact to journals in its field [15]. Another such example is the work by Raan [16] who sought to examine the bibliometric statistical properties relating to research groups. In terms of our own study, it proved difficult to compare to an 'equivalent' series of conferences as there is no similar scale event on a similar topic over such an extended period of time, almost a decade of annual cycles. We can show, however, how successful the output from TRECVid has been in gaining coverage in other larger scale conferences and other publication venues. As measures of bibliometric impact are increasingly used for promotion and research evaluation purposes, however, so it becoming imperative that all disciplines have a good knowledge of the limits and scope of the bibliometric tools relevant to their area so they can provide informed input into any evaluation work which is done. So in terms of 'why', there may no longer be much choice as the drive for bibliometric evaluation may come from external sources. We would certainly recommend, if possible, that studies of impact are initiated by the individual or research group under evaluation. The coverage and accuracy of the various bibliometric tools vary a great deal for different disciplines [17], perhaps more importantly, different disciplines have different publication cultures and different norms as to what counts as 'scholarly esteem'[18]. While bibliometric work initiated and guided from within may appear to lose some 'objectivity', it ensures that discipline expertise and nuance are incorporated, as far as possible, into the results.

3. How accurate are data collection methods?

We now provide an overview of the issues of coverage and accuracy to give a context to the more detailed discussion in section 4 which compares the results of different bibliometric tools. Our major questions concern the extent of coverage (is this tool going to retrieve everything in our subject) and accuracy (will it only retrieve relevant results and how good is it as eliminating non-relevant results). Even if tools are fairly accurate, cleaning of the results is still required. We used a mixture of methods both involving human input and automation. Expert domain checking which is a subject expert checking through the search results from the databases (examining author, date, title, source data) and deleting those not seen as relevant was used to clean the initial results before using automated duplicate checking primarily used Google Scholar for reasons of coverage [1] and also Scopus for some tasks and in this analysis we tools. For a small sample of documents we also used detailed content checking where a researcher (not a domain expert but with general knowledge of information science) checked for subject relevance by checking the entire document We now proceed to discuss data collection and then in section 4 we examine some of the cleaning techniques.

3.1. Coverage: do we have everything we need?

The main tools available for bibliometric analysis are: Google Scholar (GS) (this often includes use of Publish or Perish (PoP) [19], a software wrapper for GS which does some bibliometric calculations on the results); Web of Science (WoS); and Scopus, consisting of Scopus General and Scopus More. These have different coverage both in terms of sources they include and sources from which they are able to identify citations. In our original study we compare them in detail. In our examination of Scopus, we found that results from Scopus More were sometimes particularly good for increasing coverage and citation counts compared to Scopus General results. The purpose of Scopus More is to deliver incomplete bibliographic references from the Scopus database, not otherwise published in the Scopus General results. This is often because it draws on the bibliographic section of articles, which Scopus can access [Scopus 2011, personal communication, February 23rd]. Numerous studies to date have compared Scopus, WoS and GS, but reviewing existing studies, few discuss in detail (e.g. Harzing [17] and Meho and Rogers [20]) if or how Scopus More figured in results. In the current paper, we distinguish between searches using Scopus General or Scopus More. However, we use the phrase 'Scopus Combined' to indicate combined results from Scopus General and Scopus More searches. According to Harzing [17], like Google Scholar, Scopus More can include whitepapers, technical papers and book chapters etc. Book citations are noted as particularly prominent in the Computer Sciences, and from Harzing's cross disciplinary comparisons, Scopus More makes a substantial difference to citation counts with Computer Science publications, which is not necessarily the case with other disciplines [17].

Several studies to date have sought to evaluate the different bibliometric tools available to collect and analyse data related to specific disciplinary areas or how they perform across disciplinary areas, e.g. Harzing [17], Schroeder [21], Jacsó [22], Levine-Clark & Gil [23] and Meho & Yang [24] provide studies on tools such as Scopus and Google Scholar. A common finding in comparative studies is the often complementary nature of these tools. In addition, the changing functionality and indexing capability (as discussed for example by Bar Ilan, [9]) prompts the need for

researchers within their chosen disciplinary areas to not only be aware of the strengths and limitations, but also of new developments which can rapidly change publication and citation counts.

Overall, our review of the literature focused attention on the choice of data collection method for publications and citations being heavily dependent on the subject matter, as different disciplines have different publication, citation patterns, and cultures, and different tools have different strengths in these areas. As TRECVid falls broadly under the subject of computer science, it is probable that many researchers in the field will focus on conference papers as opposed to journal papers thus, it is important to use a data collection method which has good coverage of conferences [3]. We also found that not only was coverage important (i.e. what is in the database) but also its method of searching (i.e. where and how it looks). GS does full-text searching [25] and therefore, it is more likely to have higher recall of relevant documents as we search for papers derived from, or based on, TRECVid work, than on Scopus for certain kinds of searching. Our initial searches in our study showed that Publish or Perish (PoP), a software wrapper for GS, retrieved far more results e.g. 53% more papers in 2007 than Scopus General and Scopus More for a TRECVid search.

For the TRECVid Study [1], we chose Publish or Perish (PoP) to retrieve and parse GS results over other similar tools such as CIDS (Citation Impact Discerning Self Citations) or 'GS Counter' which have more limited functionality, and Scholarometer Sidebar (an extension or plug-in to a web browser which post-processes GS searches) or Pagella, both of which are less flexible in manipulating results. The decision to use Google Scholar and PoP, however, was not without its problems. Whilst it certainly had the coverage, it has much less sophisticated methods of duplicate detection and data analysis than its more established rivals. Thus, there was trade-off to be made between coverage and the ability to easily clean and sort results. As the coverage appeared so much better in this case, this trade-off was acceptable, yet other studies may have different needs. Another problem in general, mirroring a perennial problem in information retrieval itself, was the difficulty of being sure of full recall. How can we be sure that, in our case, we have retrieved all the papers about TRECVid? The answer is, of course, that one cannot be sure of this when examining a particular subject. The best strategy is to use very broad search strategies and use data collection methods with good coverage. In terms of analysing particular authors or smaller research groups; it is possible to check results against, for example, author's CV's, but for a subject search, this task is just too large and ill defined. Another problem with PoP is that one is dependent on what it retrieves from GS, so, even if you know it has missed, for example, one paper, there is no way to add that into the results. So even though PoP's analysis of GS results may be a correct analysis of the GS data in some cases the GS data will be inaccurate. PoP provides a list of results with various calculations, including the h-index and cites per paper. As duplicates or erroneous publications are deleted by the researcher; these figures change to take into account the new refined results. A more flexible option would be to import records into Excel and then create one's own bibliometric calculations within that to allow manual addition of papers, a technique also recommended by Meho & Rogers [20].

3.2. Accuracy: do we only have what we need?

We discuss in section 4 the problem of duplicate checking and its significance for different parts of our study. Firstly, however, it is important to be clear what a duplicate is and why it may or may not matter. Increasingly, authors may 'publish' one document in many perhaps slightly different formats, for example on an institutional repository, as a draft on someone's web page, as a slightly summarised conference presentation and as a paper in a subscription journal. Are these the same paper and how do we count their citations? Generally the convention here is that this is one paper where the author explicitly intends it to be seen as the one paper by using the same paper title, (for discussion of an authors 'expressions' or 'manifestations' of research work, refer to Bar-Ilan [9]) but that if different people (this can be checked by looking at the sources of the citations) have cited different versions then each citation counts rather than being seen as duplicates. Hence if 2 people cited the web page, 4 cited the institutional repository, 1 cited the conference and 5 cited the journal version then the total number of citations to that 'paper' would count as 12 (a total of all combined). In terms of one's h-index [26] which requires information on citations to one's papers and total number of papers, this would then count as one paper that had been cited 12 times rather than 4 papers with fewer citations. There is debate and some concern regarding the role of institutional repositories on citation counts but it would appear that generally they can only help citations as discussed in a range of studies including, Brody, Harnad, and Carr [27], Swan [28] and Garguari et al [13]. Thus, normally, the more one's paper is visible to the research community the more likely it is to accrue citations, and, if it is seen as 'one paper' this will not have a negative impact on one's average cites per paper. We found that in checking a sample, the citations accrued to each duplicate of a paper were nearly always unique and separate. In most cases, apart from very small studies, it may not be possible to check every single citation but a sample should be indicative. In section 4.3, we examine the citations accrued to duplicate papers by an author using both Scopus and GS, in order to verify if they are unique or otherwise.

So in terms of bibliometric research, we need to be careful that we only eliminate duplicates that really are duplicates and that we don't inadvertently deflate the citation count. This is perhaps another argument to use excel as

well as PoP, as discussed in the previous section, as it makes it easier to sum the citations of different versions. Duplicates matter, then, when the paper clearly is the same paper but perhaps with a misspelling of the author in one case but one has to be alert to versions that appear to be duplicates but are not, as this can incorrectly reduce the citation count.

3.3. Conclusions on data accuracy

The importance and role of data cleaning in bibliometric studies has to be taken in the context of the overall nature of the data. Bibliometric data, as has been well documented [29] is nearly always very skewed i.e. a small number of papers are cited a large number of times with a long tail of papers receiving very low or no citations. This can also vary between disciplines and computer science tends to have a particularly high ratio of documents which are not cited at all. For instance, comparing the 254 TRECVid papers which Scopus and Scholar both located in 2007, we found that 1/3 of the Scopus papers had located no citations at all, and Figure 1 shows the dramatic drop in citations per papers for both databases. Normally what matters are the documents that have been cited a large number of times, and it is interesting and important for the field to find out why, what they are about, who wrote them, and where they were published [1, 9]. Thus, in general, more effort should be put into carefully cleaning and checking this data rather than the long tail of non-cited documents These highly cited documents can be analysed in detail. For example, in our TRECVid study we investigated the quality and breadth of the conference and journal venues of highly cited TRECVid papers, as a good indicator of the impact of the research, and related studies have similarly focused on highly cited papers [30].

The next section describes the results and the implications for accuracy of some sample searches and evaluation we carried out using different bibliometric tools.

4. A comparison of bibliometric tools: data sources of Scopus, Scopus More and Google Scholar

This section provides a detailed analysis of subject searching and author searching both in terms of paper count and citation count. We also compare different types of data cleaning. In our original study we cleaned data using a domain expert to check through the initial PoP results (title, author, source) and used their judgement as to whether the papers were definitely 'about' TRECVid (rather than just perhaps tangentially related) as well as checking for duplicates. In a sample discussed here, we also used content checking by reading the entire document to ascertain how much extra certainty this might provide on relevance.

4.1. Subject searching: Scopus versus Google Scholar

Results for Harzing [17] and Thornley et al [1] showed how GS (Google Scholar) coverage for Computer Science publications is substantially greater than Scopus, and Harzing [17] indicated that a 'Scopus More' search is particularly important to increase coverage of Computer Science publications when using Scopus. To examine this issue in greater detail, we examined a number of datasets related to document searches we carried out with Google Scholar (GS), Scopus General, and Scopus More. We chose the year 2007 as a representative year as it is fairly recent whilst not being so recent that it is likely to accrue many more citations than it already has. Papers published in that year are likely, within computer science, to have already received most of the citations they are ever going to get. 2007 is also reasonably important in the progress of TRECVid since it represents a year where the same video data had been used for a few years so performance of participants was peaking and thus it should have formed a good basis for publications. Firstly, we undertook a general search, comparing the results as seen below, and secondly we analysed in more detail the top 20 most cited papers in that year.

4.1.1 All TRECVid papers for 2007

Using Scopus, we carried out a document search for 'TRECVid' for the year '2007'. We carried out four initial searches, firstly searching 'TRECVid' using 'Title, Abstract, Keyword' (TAK) fields, examining 'Scopus General' and 'Scopus More' results. We also searched for 'TRECVid' in 'all fields', which crucially searches each paper's bibliography. It was important to carry out this 'All Fields' search, because 'GS' does full text searches which include 'references' in retrieved results. Thus, although a like for like comparison between a topic search in Scopus and GS is not possible, we can at least ensure a fairer comparison with 'All Fields' searching, retrieving paper results in Scopus with 'TRECVid' mentioned in one or more of their bibliographic references. The results are as follows:

Table 1. Keyword Search for TRECVid limited by year, "2007" (Dec 2010)

	Scopus General		**Scopus More**		**Combined Total**	
	Papers	Cites	Papers	Cites	Papers	Cites
Title, Abstract, Keywords						
Results	144	744	187	246	331	990
Bad data removed*	-		156	232		
Duplicates removed also	127	744	68	232	195	976
Overlaps with Scopus removed			56	232	183	976
Papers not from 2007 removed			48	225	175	969
All Fields						
Results	274	1244	187	246	461	1490
Bad data removed*			156	232		
Duplicates removed	253	1244	68	232	321	1476
Overlaps with Scopus removed			55	232	308	1476
Papers not from 2007 removed			47	225	300	1469

* No author and Title name

A notable finding in this instance was the limitations of 'Scopus More' compared to the 'Scopus General' search, in terms of coverage and duplicates. The additional papers 'Scopus More' principally found were TRECVid workshop or conference proceedings and there were a high degree of duplicates in results. An important consideration here is the issue of citations to duplicate papers, and in other data analysed. The argument discussed earlier in section 3.2, and findings in section 4.2.2, is that whilst the paper is deemed a duplicate, the citations are not, as the cites are from different sources. Next, we looked for papers found in both 'Scopus General' and 'Scopus More' results, and found 12 papers (18%) for 'TAK' search, and 13 papers for 'All fields' which overlapped. For the latter 13 papers, Scopus General picked up 105 citations, and Scopus More picked up 26 citations. There is the potential that some of the additional 26 citations 'Scopus More' picks up are already accounted for in the 105 citations from Scopus General (Scopus 2011, personal communication, February 23rd). However, given that 26 citations amounts to .02% of the total citations, this is of no concern.

To more fairly compare the search results between combined 'Scopus General', 'Scopus More' results, and GS results, we carried out an 'All Fields' search with Scopus Combined [Scopus General and Scopus More combined]. This searches additional fields such as a papers bibliography, but, unlike GS, this does not search the full contents of a document. We find the following:

Table 2. Search for TRECVid limited by year "2007" (Dec 2010).

	Scopus Combined		**Google Scholar**	
	Papers	Cites	Paper	Cites
Total after cleaning	300	1469	460	4962
Total with domain checking	289	1426	424	4488

Crucially, we found that there are 39 papers in Scopus Combined results not available in GS, and vice versa, there are 183 results within GS (201 without domain checking) not available in 'Scopus Combined 'All Fields' search. .Of the 39 Scopus Combined exclusive results, 32 were 'All Field' results accounting for 52 of the 53 citations. The coverage of papers between Scopus Combined and GS is not nearly as significant as differences in citations, as Scopus Combined produced 68% of Scholar's coverage in terms of papers found. The larger publication count with GS may be mainly as a result of Scholar's additional sources, or because GS is searching the full contents of documents. There are

255 papers that both databases pick up, and of these, the Line Chart in Figure 1. Illustrates how GS consistently picks up more citations for each paper than Scopus Combined, in total accumulating over 2 ½ times the number of citations. It also illustrates that occasionally Scopus Combined does do considerably well at matching citations. Interestingly, it shows that the ranking of papers by citations in Scholar does not correlate well with that of Scopus Combined if we focus on more highly ranked papers by citation. A Spearman correlation of 0.785 is observed for the 254 papers which includes poorly cited papers, but if we focus on the greater cited papers, for instance the top 150 papers, this Spearman correlation drops to 0.697 In comparison with other work [5] who, when comparing GS and WoS impact factor measures of journal and conference, found a Spearman rank correlation of 0.88 [5, p.130]. Meho and Rogers [20 ,p. 10] also found a strong correlation between ranking (0.970) of authors (within Human Computer Interaction) by citation when comparing Scopus and Web of Science. It should be noted, however, that they also found Scopus did well at locating more citations for most authors and was also better at showing nuanced differences at either end of the ranking i.e. between highly cited authors and between poorly cited authors. The reasons behind the differences in paper ranking between GS and Scopus are probably due to Scopus and Scholar differing in indexing and access to publications but the impact of different databases on ranking also seems to vary on whether analysing a journal/conference, author, or subject analysis, with future studies needed to understand this further.

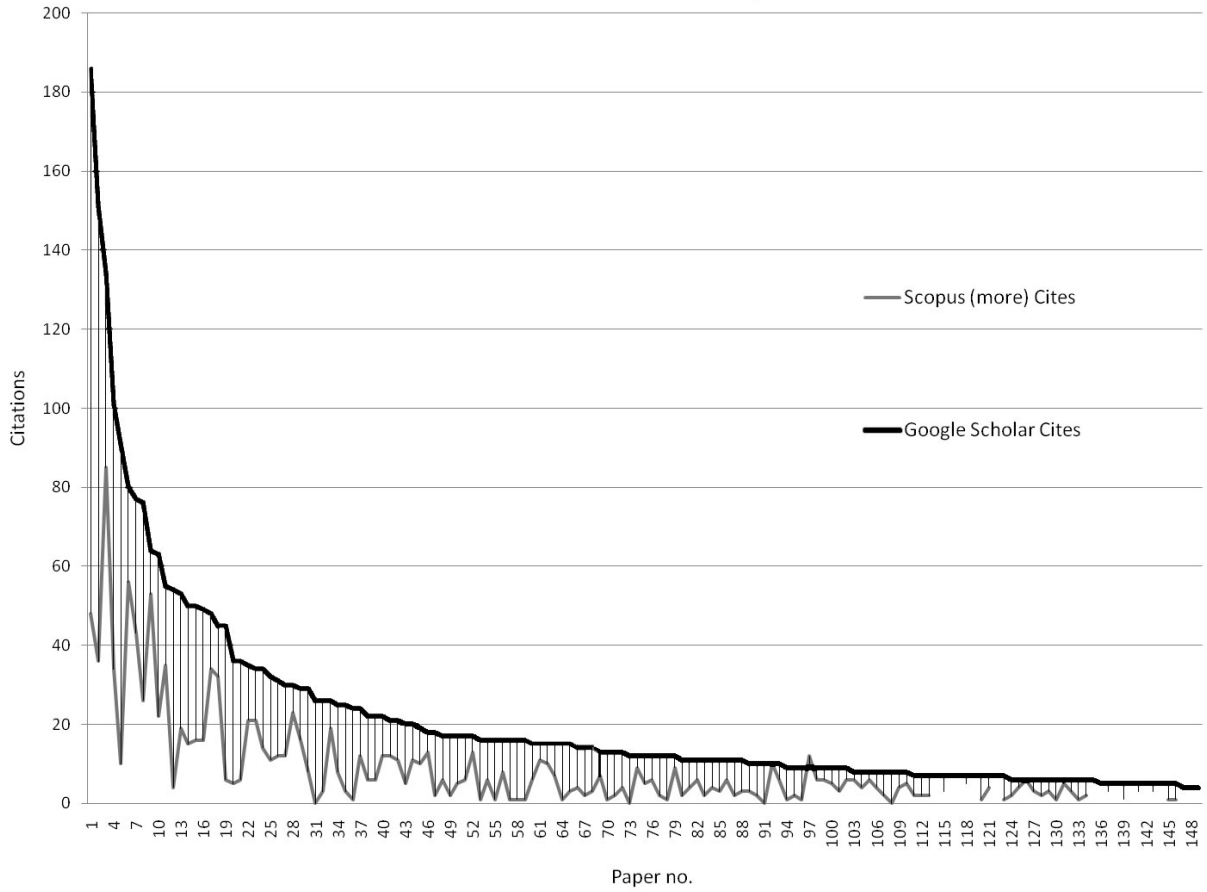

Figure 1. Comparing paper citations, ranked by GS citations per paper

4.1.2 Top 20 TRECVid cited papers for 2007

Next, we examined citations, taking a cross sample of the top 20 cited papers in GS using a keyword search for 'TRECVid' in 2007. We found these cites corresponded to 37% of overall GS results, and 68% of top 20 exclusively retrieved GS results (i.e. when we removed the papers that were also in Scopus). In addition to 'Expert Domain Checking' of titles and authors previously discussed, in this example, we also carried out 'Content Checking' by examining the full content of the document to verify that TRECVid; data, discussion or influence, is present in the paper. When we isolated the top 20 papers which GS exclusively recovered, upon examination we found over half of

the papers were potentially not highly relevant results for our search, 5 of the papers appeared not to have been published in 2007, and 7 papers only fairly briefly referred to TRECVid.

Table 3. Examining highly cited TRECVid related papers limited by year "2007"

	Top 20 Papers cites	**Overall cites**
Overall Scholar Results	1833	4962
Content checking	1641	
Results found exclusively in Scholar	817	1246
Papers not from 2007 removed	678	
Content checking	275	
Scopus Combined (TAK)	524	969
Scopus Combined (All fields)	649	1469
Scopus (All fields) without Scopus (Tak) overlap	465	

From figures presented in Table 3., exclusively retrieved GS results accounted for 25% of the total citations, and only 33% of citations in the top 20 exclusive GS results remained after examination (when we removed all that had any uncertainty over date and level of relevant content). Given that the top 20 exclusive GS results accounted for 68% of its total before examination, we can infer that papers accruing up to 835 citations from Scholars total may not be completely accurate in terms of year or content for this type of subject search. Upon investigating this, we found it a particularly common problem with the interaction of IEEE papers and GS. For instance, Scholar was drawing on IEEE Explore paper's 'Date of Current Issue' rather than 'Issue Date' Information. Upon contacting IEEE Explore, we found that 'Date of Current Issue' is availability on IEEE Explore, not the publication date [IEEE Explore 2011, Personal Correspondence, Feb 24th,]. This suggests that, depending on the importance of dates in the nature of the study, it is important to double check the date of publications, and that the reliability of GS picking up the correct date varies between publication sources as also discussed by Jasco [31] and Bar-Ilan [32]. Detailed content checking in terms of reading every document is clearly not practicable in many cases and it may not be as 'reliable' as subject expert looking at author and title (who for example, despite minimal apparent TRECVid content in paper, may know from background knowledge that it is based on TRECVid work). The way in which relevance judgements about the papers depended both on the level of subject expertise of the researcher and also the amount of the publication data (bibliographic details or full text) that they checked are clearly important factors to consider in future or related studies based on subject areas.

4.2. Author searching: Scopus Combined versus Google Scholar

In this section we discuss the relative results from searching for a comprehensive list of a particular author's publications, and also citations using Scopus and GS. This data was collected to ascertain what percentage of total publication output of TRECVid participants work was based on TRECVid work and if these papers tended to receive different levels of citation than those based on their other research.

4.2.1 Author papers

The original study chose 5 authors involved in TRECVid conferences at various levels of seniority. Names of authors were sampled based on names which were less likely to have similarly-named authors. The purpose was to ascertain how important TRECVid-related papers were to an author's academic career. A domain expert selected 5 researchers involved in TRECVid of varying seniority, ranging from most junior tv1 to most senior tv5, and we examined those author's publication and citation records over a five year period. Similar to the 'Document Search' using the keyword, 'TRECVid', we found significant author paper duplication in the 'Scopus More' results. Compared to our experience with 'document searching' for 'TRECVid' papers, Scopus More performed considerably better than 'Scopus General' in retrieving additional papers and citations when 'author searching'. However, compared to the 'Document Search' results in 'Scopus More', which principally found additional papers not in Scopus General, the author search results in

'Scopus More' (although notably finding many additional papers not in Scopus General) also had considerable overlap with papers retrieved in Scopus General, requiring additional cleaning of combined results.

Table 4. Author Search Limited by years 2003-2007

	Scopus General		**ScopusMore (SM)**		**SM Cleaned**		**Scopus Combined**		**Google Scholar**		**CV**
	papers	cites	papers	cites	papers	cites	papers	cites	papers	cites	papers
Tv1	7	15	29	42	16	42	16	57	12	115	16
Tv2	27	396	104	278	37	278	47	674	48	1440	46
Tv3	47	445	108	239	36	239	63	684	68	1000	55
Tv4	13	20	54	69	30	54	32	74	28	157	26
Tv5	26	120	144	343	52	144	59	264	57	1264	45

To provide external validity for the papers retrieved, and test the extent of coverage, we reviewed publication data published by the authors themselves on their personal websites. We found that in most instances, Scopus Combined and Google Scholar databases provided good coverage of authors' publication records. There is some inconsistency in terms of higher retrieval for papers above the authors' published CV's, and upon investigation we found specifically 'Scopus More' and 'GS' were finding further papers in the samples tested, not listed in the author's CV. In some cases but not others, this was as a result of these databases retrieving papers outside the publication years sought. It is also possible that on author's CV's they tend to exclude 'lower status' papers and our initial review of 'missed' paper suggested that this may be the case. Another interesting finding was that 'Scopus Combined' had better coverage of author publications than Scholar in most cases. The data also shows a high degree of duplicate papers in 'Scopus More' as opposed to Scopus General, requiring careful checking to ensure accuracy of paper count. The comparison chart below demonstrates that for author searching, 'Scopus More' did well in terms of unique papers it retrieved in addition to Scopus General results, which contrasts with our experience with 'Document Searching'.

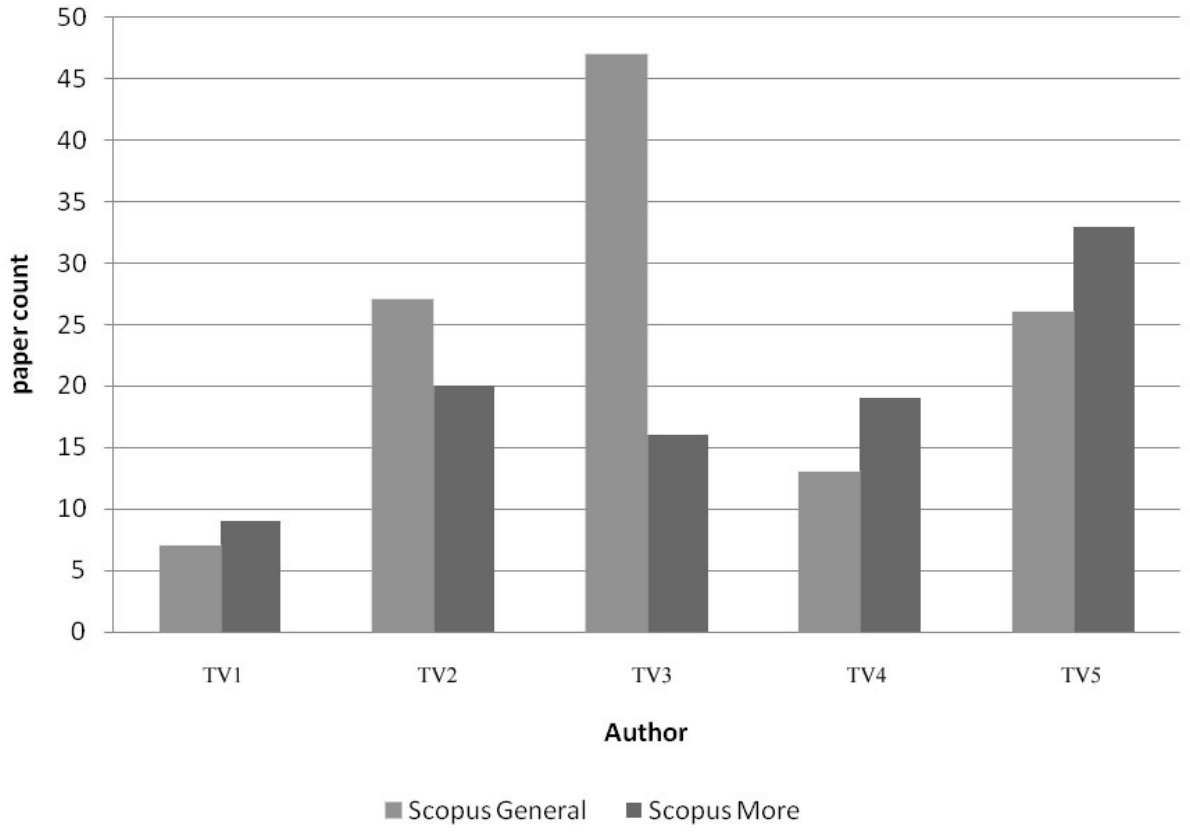

Figure 2 Author Search comparing Scopus General and Scopus More

4.2.2 Citation counts for papers

Finally, we examined the issue of citations that GS and Scopus attribute to papers. We wanted to see if either GS or Scopus would provide any duplicate citations to a paper. For instance, we sampled a recent paper by TRECVid participants; 'A formal Study of Shot Boundary Detection' [33] and examined, in detail, citations from both Scopus Combined and GS. Scopus returned 67 citations in Scopus General, and 1 citation in Scopus More. The article citing the paper in Scopus More was not found in Scopus General. GS returned 104 citations in total. Of the 104 citations, on close inspection, only 1 was a duplicate. Interestingly we found that there are 42 unique citations to the paper in GS not available in Scopus, and found 10 unique citations in Scopus not picked up by GS. Included in GS citations were books, a doctoral proposal paper, several non-English language papers, scholarly papers hosted on repositories, as well as journal and conference papers not picked up by Scopus. The conclusion in this case was that we could validate GS's ability to locate unique and relevant citations of researcher's work. We also concluded that in this case, duplicate citations were nonexistent for Scopus Combined and rare for GS.

Next, we examined whether duplicate papers in Scopus More had unique citations or duplicate citations also. We examined the citations for the one TRECVid author (tv2) in Scopus More, limiting years between 2003-2007. Of the 278 citations to 104 papers (many of which were duplicates), we found only 1 duplicate citation, verifying that the citations to duplicate papers should be counted when using Scopus. Our experience with GS was that duplicate papers overwhelmingly carried very few additional citations. For example, we also carried out a GS (tv2) search limited to years 2003-2009 and we found that of the 83 papers and 1705 citations, 13 were duplicate papers accruing 8 citations, and of these 8 citations, 3 were duplicate citations. In this case, GS duplicate citations have a much higher tendency to be duplicates and not unique. Our overall experience using Google Scholar was that duplicate papers tended to have so few citations, however, that this was not an important issue.

4.3. Discussion on comparisons

The analysis above raises some caution about the premise of additional article coverage GS actually provides, and serves to focus attention on the ability of GS to accumulate greater citations per paper. As with GS, Scopus More work requires careful cleaning and yet provides exportable data fields for further analysis. The comparison of GS and Scopus

Combined citations to papers, illustrates that there is often no correlation between the number of citations which Scholar attributes to a paper compared to Scopus, and this has implications for the use of impact metrics such as the H-Index etc. The above evaluation also serves to highlight an issue as to whether Scopus should provide additional search capabilities to match GS full text searching, and conversely, that GS should provide refined Field Searching also.

Overall, based on our findings, Scopus Combined does considerably well at matching GS paper coverage, but not its ability to locate citations. Because GS searches the full contents of articles, it means that a direct comparison between certain searches with Scholar, and other kinds of tools, is not possible. We can say though that full content searching by GS would provide the most complete means of tracking the use of a term throughout the literature. Scopus General (not Scopus More) on the other hand, provides abstract, keyword and reference data of use in carrying out Co-Word or Co-Citation Analysis not practically possible with GS and should be the tool of choice where citation counts are of less interest. The provision of abstracts, as discussed in the visualisation section, can be useful for further analysis. In our use of GS, we had to retrieve each abstract separately, a time-consuming exercise, as it is not provided with the results.

It is important to conclude here that our original comparison of authors sought to examine 10 TRECVid authors, five more than the authors listed above. Unfortunately, papers for the additional authors we searched could not be adequately isolated and cleaned because results for a particular author may stretch to thousands, when common/similar names were involved. Thus, the use of Scholar and the PoP tool can be an impractical tool for author citation analysis, though narrowing by additional keyword searching etc. and indeed disciplinary area can overcome some limitations. It must also though be noted that, limiting author searching by disciplinary area in PoP, meant that author papers were being excluded from the results. Overall, there is currently a severe limitation of Google Scholar in terms of author impact analysis despite its demonstrated ability to provide strong paper and citation coverage. In test cases, however, where there is correlation between authors' paper ranking by citation between Scopus Combined and Scholar, we would certainly recommend that GS & PoP be used for more comprehensive citation coverage of an authors most popular papers.

5. Comparing different search techniques and data cleaning methods within PoP (Google Scholar).

This section examines the effects of changing search strings and other parameters in running author and subject searches in PoP. The aim is to show the implication of different search methods and thus provide suggestions for the suitability or otherwise of this for related studies. We also discuss the extent to which different methods of data cleaning impact on the final results.

5.1. Subject searching in PoP (GS)

Our TRECVid study sought to examine the scholarly impact of a TRECVid benchmarking conference through the quantification and patterning of citation data. Using PoP, in the table below we show a sample search showing the relative effects of different search approaches (one limited by discipline and one not) and cleaning methods. The initial clean is the removal of duplicate and patents, and the expert domain checking shows the effect of expert domain checking of the bibliographic records of the remaining results. An example of how we searched and cleaned results is as follows:

Table 5. Search for 'TRECVid' with PoP limited by year, '2007'.

	Paper	**Cites Index**	**H Index**
All	492	4952	33
Cleaned*	460	4932	33
Expert Domain Check	424	4488	29
Limited by Discipline; "Engineering, Computer Science and Mathematics"			
All	449	4796	33
Cleaned*	421	4782	33
Expert Domain Check	390	4359	29

* Duplicates and patents removed from results

We searched for papers in PoP using the 'all the words' field for 'TRECVid'. This is the equivalent to a standard GS search, which matches the search terms occurring anywhere in the searched documents (author, title, source, abstract, references, etc.) and may provide too many non-relevant results for effective results. We then removed patents and duplicates (6% drop) and then used expert domain checking to better ascertain that these papers were as a direct result of TRECVid. For each year we checked the PoP search results (not the detailed content but the bibliographic information from the results) with the criterion for inclusion being 'was this publication a direct result of TRECVid activity?'. By this, we mean that the paper uses TRECVid data or benchmarking criteria or describes a technique tried in TRECVid. The expert domain checking thus aimed to exclude papers which just cited or mentioned TRECVid, as our aim was to include papers truly derived from TRECVid participation. For 2007, this resulted in a further 36 papers (an 8% drop), accounting for 444 cites being removed (9% drop) for the 2007 TRECVid search of all disciplinary areas. For the discipline-specific search, it removed 31 entries and 423 cites. In summary, paper counts were reduced by 14% in 2007, through cleaning (6%) and expert domain checking (8%). This was consistent with the domain experts overall assessment of checking citations in the other years examined that there was approximately 10% redundant papers listed. Citations were reduced by 0.4% by cleaning, but 9% through expert domain checking. Expert domain checking also lowered the h-index by 4 points. The conclusion is that, similar to our findings on individual author searching, removing duplicates and patents will reduce the paper count but minimally affect citation and h-index scores. In our case, standard cleaning of data generally had little effect, a finding consistent with Li et al's [12] experience of LIS publications, where "substantial effort resulted in relatively small changes in results". However, expert domain checking of search results had more of an impact on both paper and citation count, with a resulting increase in accuracy, and is probably a useful technique for other studies to consider.

5.2. Author searching in PoP (GS)

In an author search for papers written by tv1 we found that our initial search retrieved results of 245 papers with an h-Index of 19. When we changed the search string, using a search qualifier, to only retrieve the exact phrase of the authors name we retrieved 79 results and 276 cites. After similar authors and duplicates are manually removed, our initial figures go down to 19 papers, 157 cites and a H-index of 7. We also tried narrowing by disciplinary areas (Engineering, Computer Science and Mathematics) but this underscored recall to 16 papers with 119 citations after cleaning. Thus, duplicates can generally be removed with little impact on citation counts, though a 10% drop in paper count can be observed in this instance, with the existence of similar authors in refined searches causing 73% redundant results. We had similar experiences with other authors, for example, cleaning of tv2 results resulted in a 40 percent drop in papers, 17% drop in citations, and a h-index 3 points lower. The conclusion in this case is that an appropriate search string is vital to ensure coverage, whilst checking of author names will ensure accuracy of paper and citation counts. Duplicate checking ensures accuracy of paper counts, but generally does not affect citation or h-index scores, a similar finding to our subject searching analysis.

6. Tools for visualisation

Our original study also sought to ascertain how topics treated in TRECVid papers have developed and evolved year-on-year since the start of TRECVid and to accomplish this we used word frequency within TRECVid publications to create word visualisations. One task when utilising word frequency occurrence in Journal Papers to track research trends, is to ascertain which words in the paper are actually significant and meaningful in terms of describing trends. For example, widely-known research jargon, stop-words or generic terms specific to the field are not likely to be useful. Thus, such terms are removed from the results as we did in our study excluding terms common and generic to TRECVid. Furthermore, there is the question of what data to draw on, whether titles, abstracts or full text and the related question of how to harvest this content. The first set of publication data we used was the combined title and abstract data of TRECVid workshop papers in order to identify popular topics, methods, technical terms over the period of TRECVid conferences. This data was available on the TRECVid website so no searching was required although it was time consuming to cut and paste and in some cases re-typing was necessary. The method was deemed successful for its purpose, as domain expertise provided confirmation that the trend displayed were indicative of how the field had developed.

The second set of data was the TRECVid derived papers, or paper about TRECVid, and in these cases searching and harvesting the content became more of an issue. The collection of abstracts from the TRECVid derived papers from the

GS search was very time consuming as it required clicking on each paper individually, going to the abstract, and then copying it. For the analysis discussed in this section, we used Scopus General, which was far less time consuming, as it allowed easily obtaining and exporting abstracts, and as we have shown in Section 4, its paper coverage can favourably compare to GS so it is a good tool to easily obtain and export all abstract fields pertaining to TRECVid related searches. In our visualisation study, we were not using citation data so its limitations in collecting citations ere not such an important issue.

According to Hui and Fong [34], "co-word analysis attempts to cluster semantically similar documents. Traditionally, co-word analysis has been mostly used to identify research trends and in automatic thesauri generation". Co-occurrence analysis is the umbrella term for co-word analysis which uses the contents of different fields in bibliographic records to study the structure and relationships among documents, authors, institutions and disciplines [35]. The premise of co-word analysis is that if words or concepts tend to both be present in fields encountered, then some relationship exists between these words or concepts. This allows the researcher to map relationships. Unlike multi-word frequency analysis, it does not mean that words are next to each other, but can mean that they are physically close. Multi-word frequency analysis in information retrieval (IR) on the other hand, seeks to identify frequency with which terms or phrases appear in a document, with the intention of automatically deriving indexable terms and keywords for a select document. Multi-words has been referred to as 'Bigrams', 'Word Sequences', 'Terms', 'Phrases' and 'Word Pairs or Triplets' throughout the literature. However, by mapping term frequency chronologically, we may additionally be able to identify research trends in a given field, observing how terms emerge and die-off. Such terms could refer to research thematic areas, techniques, methods, concepts, theories or philosophical frameworks employed.

Using Scopus General, we obtained datasets of TRECVid related papers from the years 2007-2009, the purpose of which was firstly to validate the use of title and/or abstract fields to analyse data, and secondly to examine advantages of multi-word frequency analysis. We compared keywords, titles and abstracts using Taporware [36] to analyse data. 'Taporware Keyword Finder' returns the top 20 single words, and each of the top 10 pairs and triplets. We then filtered from the source document irrelevant words, and reran the Taporware tool, until results were meaningful. Next, we removed single words which had approximately the same occurrences in word pairs, and word pairs with similar word frequency in triplet. Thus, 'Shot', was combined into 'Shot boundary', and 'Shot Boundary' into 'Shot Boundary Detection' etc. We found many single terms could be excluded as they were generic terms, or could be explained better using word pairs or triplets. We included the remaining top 3 single word terms for each year. Finally, we removed word triplets that only appeared once, and colour coded terms to easily glean patterns from the table.

	Keywords		Abstract		Titles	
126 Papers 2007	fusion	12	fusion	52	summarization	10
	semi-supervised	6	query	49	query	5
	interactive	6	framework	45	automatic	5
	concept detection	9	video search	31	video search	14
	video summarization	7	concept detection	23	news video	10
	video search	7	news video	21	video annotation	9
	video annotation	6	concept detectors	20	concept detection	8
	semantic video	6	average precision	17	semantic video	7
	supervised learning	6	supervised learning	16	interactive video	7
	video indexing	5	active learning	13	active learning	4
	semantic video retrieval	6	semantic concepts	13	video indexing	4
	shot boundary detection	5	mean average precision	12	news video retrieval	7
	video concept detection	4	shot boundary detection	12	semantic video search	4
	semantic concept detection	3	video concept detection	6	video concept detection	4
	multimedia information systems	3	semantic concept detection	6	shot boundary detection	4
	retrieval average precision	2	video retrieval systems	5	semantic concept detection	3
	video retrieval average	2	support vector machines	5	interactive video search	3
	near duplicate detection	2	video search engine	4	near duplicate keyframe	3
	local interest point	2	support vector machine	4	semantic video retrieval	2
			near duplicate keyframe	4	semantic news video	2
			news video retrieval	4	efficient near duplicate	2
128 Papers 2008	fusion	11	fusion	53	fusion	9
	image	10	baseline	34	temporal	8
	event	6	training	27	multimodal	7
	video summarization	18	concept detection	25	video search	12
	video annotation	7	video search	22	concept detection	12
	concept detection	7	semantic concept	20	video summarization	8
	information retrieval	7	video summarization	18	semantic video	8
	video search	6	large scale	17	video event	6
	data mining	5	video annotation	16	video annotation	6
	video indexing	5	video summary	14	semantic concept	6
	near duplicate	4	news video	14	video concept	5
	video concept detection	3	visual features	12	image retrieval	4
	kernel fisher discriminant	2	semantic concept detection	12	semantic concept detection	6
	mid level features	2	video concept detection	6	video search reranking	3
	shot boundary detection	2	mean average precision	5	video event detection	3
	fuzzy decision trees	2	video search reranking	5	semantic video indexing	3
	summarization algorithm fusion	2	support vector machines	4	video concept detection	3
	semantic concept detection	2	context dependent fusion	4	semantic video annotation	2
	multimedia information retrieval	2	video content analysis	4	semantic video event	2
	high level feature	2	single level emd	3	multimedia semantic concept	2
			semantic word similarity	3		
			support vector machine	3		
84 Papers 2009	fusion	9	event	33	features	8
	motion	6	model	32	visual	7
	event	5	similarity	29	domain	6
	concept detection	11	video search	21	concept detection	12
	video search	6	concept detection	18	video search	5
	feature extraction	6	semantic concept	10	semantic concept	5
	video indexing	5	feature value	8	large scale	4
	supervised learning	5	visual features	8	semantic video	3
	video annotation	5	interactive video	8	story segmentation	3
	multiple correspondence	4	negative examples	8	real time	3
	shot boundary detection	5	shot boundary detection	9	shot boundary detection	4
	multiple correspondence analysis	4	interactive video retrieval	7	video concept detection	3
	high level feature	3	feature value pairs	7	semantic concept detection	3
	level feature extraction	3	multiple correspondence analysis	5		
	semantic concept detection	3	association rule mining	5		
	video concept detection	2	rough set theory	4		
	association rule mining	2	uni modal detectors	4		
	video information retrieval	2	semantic concept detection	4		
			video retrieval evaluation	3		
			dimensional feature space	3		

Figure 3. Comparing bibliometric fields for term frequency analysis.

The results indicate that multi-word frequency provides a promising means to track research trends in certain disciplinary areas by chronologically identifying frequency of terms. Comparing bibliographic data fields such as 'Keywords', 'Title' and 'Abstract we can see for instance that 'Concept Detection', 'Video Search' and 'Shot Boundary Detection' holds similar ranking in frequency across fields. (Title field data appeared moderately good at indicating terms, though abstract data should still be the preferred option. Combining title and abstract data, and focusing on a greater degree of word pairs would provide a middle ground approach. In order to visualise trends, a programming interface such as "R" [37] has become a popular choice in data mining and extraction research fields.

The methodological issues arising from our visualisation work in terms of choice of bibliometric tools strongly favours the use of Scopus in terms of ease of collection of abstracts if these are required. In our case, abstracts were particularly important when examining actual TRECVid workshop papers, as in those cases, titles tended to be uninformative of content e.g. 'the University of x at TRECVid 2008'. In terms of similar future studies, a small pilot run on title and a comparative one including abstracts which is then checked with a domain expert is a good indicator of whether abstracts are necessary.

7. Conclusions and Future Directions

We have discussed and compared a range of different bibliometric tools used in a particular study within computer science and shown how the use of different search and data cleaning techniques can impact on the accuracy of the results. What conclusions, can be drawn about which is the 'best' bibliometric tool to use and what new insights can this particular study bring to existing evaluative research of the various bibliometric tools and possible guidelines for related studies. These finding are probably most relevant to related computer science studies but also for any academic discipline which, for reasons of publication and citation patterns, may not be suited to a bibliometric analysis using only the large commercial databases. They are also, due to the range of evaluative and analysis tasks that we carried out, likely to be useful to anyone aiming to carry out such a comprehensive and multi-faceted scholarly impact study.

7.1. Which bibliometric database is the best?

Our analysis broadly confirms previous studies (e.g. Harzing, [17], Li et al, [12]) that Google Scholar is a very comprehensive bibliometric tool for computer science. It further develops this work by detailed comparisons of both paper counts and citations counts between Scopus General and Scopus More, and their combined results against Google Scholar. The main finding is that in addition to Scopus General, Scopus More was particularly important for paper searching by author, and that these combined results often surpassed Google Scholar in terms of papers it could retrieve, but that Google Scholar was far more successful at retrieving citations. We did however find that GS could be an impractical tool for author searching, given that half of our selected test authors' work could not be practically or reliably ascertained. This can be a severe limitation in methodological studies seeking to compare author's work between various databases, as well as in practical use.

Comparing the citations of 255 papers that both Scopus Combined and GS retrieved for a given search, we found that Scholar consistently retrieved greater citations, in total accumulating 2 ½ times that of Scopus Combined. In terms of comparison of different databases, we also found an inconsistent correlation between papers retrieved by each database when ranked by citation. So, depending on which database is used, not only may individuals or groups have different citation levels, but they may come in different places in any citation ranking. This is an important issue and suggests that, certainly within computer science and other disciplines, any comparison of citations levels between different authors or research groups should be approached with caution and a thorough study of the coverage of the database used. For related studies on this issue, Li et al (12) have examined this issue with LIS researchers. In CS, Meho & Rogers [20] looked at h-index score rankings across databases finding high correlation and Franceschet [38] compared WoS and GS finding high correlation between cited paper rankings. Future work should further focus on this issue in different disciplines, and in different contexts such as author ranking, paper ranking, and research team ranking.

In further development from existing studies we sampled whether citations accruing to duplicate papers in GS were unique citations to that paper, and we found that in nearly all cases they were valid. Our general experience with duplicate papers and their citations was that they had little to no impact on metrics. Future studies could therefore focus on highly cited paper's duplicate copies, and whether their accrued citations are unique. Overall, we conclude that if the main interest is to accurately locate papers then Scopus and Scopus More are a good choice but GS is better if comprehensive citation metrics of; authors, papers, or kinds of 'topic searching' are sought, though author searching may well require delimiting by keywords or disciplinary area as well as time consuming cleaning of data. One caveat

for GS must be, however, the problem of inaccuracy of its date fields, which has shown in other studies [31, 32]. This varies a lot between different publication sources so would need to be checked for carefully in other studies to locate sources within that subject area that may have this problem.

We found that PoP is useful tool for searching and retrieving the data from GS though it has limited flexibility for some tasks (e.g. the addition of papers that have been missed in the search) and, in some cases, exporting results to excel may be more flexible option and it should be noted that, due to expert domain input, we were able to notice 'missing' papers which may not have been clear as an issue in other studies. Despite this, however, PoP has all the advantages (mainly coverage) of GS and also the weaknesses (incomplete bibliographic data, some inaccuracy in dates) though one very important point in its favour is the fact that it is free.

Our study involved both visualisation of research trends as well as citation analysis and it became clear that different bibliometric tools were best suited to these two different objectives which has not been observed in similar studies. In terms of comprehensive citation data GS is the best but for our work on visualisation Scopus was a superior choice due the ease of collecting and exporting abstracts

7.2. Which data cleaning method is best?

We compared a number of different methods of cleaning the results of our searches in particular the subject search for TRECVid 'derived' papers. Expert domain checking involved a subject expert manually going through GS search results and excluding those that did not adequately meet the criteria of being 'derived from TRECVid'. For a smaller sample a researcher (not a domain expert) examined the actual content of some of the papers and this was found to reduce the paper count further. The relative merits of expert domain checking of search results which is time consuming but manageable versus detailed content checking, which can only be manageable for small data sets, does not seem to have been studied previously. In general both of these cleaning methods had a significant impact on papers and citation counts (unlike automatic tools which mainly reduced paper count) and thus further research on their role, particularly with regards to subject searches, may be useful. In terms of automatic tools for duplicate checking excel was a very good tool which is not a new finding but its usefulness in allowing the insertion of 'missing papers' unlike in the PoP calculations has not been noted before.

7.3. Future developments and improvements

The tools discussed and analysed in this paper as well as other studies have different strengths and weakness for different bibliometric analyses. The weaknesses can, to a greater or lesser extent, be ameliorated by different levels and type of cleaning and checking of their results. It is still the case, however, that they are a fairly blunt instrument and we discuss some developments which may begin to allow a more nuanced picture of academic quality or scholarly impact.

One potential for new and perhaps more effective means of ascertaining scholarly impact is the analysis of viewing habits of journal articles themselves. Already, Digital Libraries can produce evidence on the usage patterns of academic articles [39]. As cloud-based Personal Information management (PIM) and Virtual Research Environment (VRE) tools proliferate, the ability to be able to monitor and analyse usage patterns of journal articles has become feasible. For instance, Mendeley.com already provides readership statistics for journal articles based on how many users have copies of an article in their personal collections. PIM and VRE tools will soon be able to provide more detailed quantitative understanding of impact, based on number and length of times an article is viewed, and even how much of an article tended to be annotated, and which sections etc. Semantic Web projects are also attempting to create technologies that assist researchers in marking up documents at the time of creation in RDF etc. Thus, the ability for scholarly tools to immediately ascertain trends and divergences in clusters of research will be possible. Hence, overall, we will likely have new rich data in the near future to compare and validate citation counts, as either a normative theory of citation or a social constructivist theory of citation [40] [41] as well as an additional means to assess scholarly impact and trends. Furthermore, as e-reader devices and PIM and VRE technologies proliferate, the relationship between author and reader will change enabling dynamic dialogue between a community of researchers on a paper and its author, undoubtedly leaving evidence along the way of a paper's impact, which maybe used for analysis.

One of the limitations we encountered in the TRECVid study with using tools such as PoP is the inability to clearly distinguish authors of the same name, resulting in manual filtering of results. Given that we specially chose 5 authors in our TRECVid study with distinguishable names, it is likely that the need for checking of results will be even more vital in many cases. It must be noted that there are efforts underway to ameliorate this problem. Ten years ago, Cronin [42] asked whether common standards would emerge for "identification, formatting and labelling" as well as identifying other kinds of 'data objects' [42]. The recent non-profit and open source ORCHID project (Open Researcher and Contributor ID) [43] is attempting to provide a solution to part of this problem similar to DOI. In August 2010, it became a non-profit organisation seeking to create a registry for author identification with data sources

from Thomson Reuters, Scopus, Crossref and others. Thus, we will likely see accurate and better representation of authors' work in the near future. Much of the literature indicates a need for a wider understanding of impact, and better metrics for assessing scholarly work and impact. This might include assigning contribution to mentoring, blogs, presentations and datasets etc. A recent initiative, 'Datacite' [44], is a consortium set up in December 2009 to increase access and increase identification of scientific datasets of research, with Datacite members assigning DOI numbers to their datasets.

As new tools and techniques emerge which extend the ability to locate scholarly work and attributed citations, addressing the issue of quality, and assigning merit emerges [45]. A citation to some kinds of Scholarly Data Objects may not carry the same merit, as for instance, a peer-reviewed Scholarly Journal Article [42]. For example, a Library and Information Science (LIS) search of authors by Meho & Yang [24] in their study of citations, examined the top 31 combined Journal and Conference publications which Scholar could exclusively retrieve compared to WoS or Scopus. They found that with the exception of one title, none of the publications ranked nearly as well as top citation ranked LIS publications. Importantly, as the amount of open access materials increases, the chances of citations being accrued to easily available, yet non peer-reviewed scholarly material may increase also. With offerings such as GS, there becomes an overlap between the goals of what Bibliometrics and Webometrics tries to accomplish, and more attention by the researcher is needed to filter out data from results obtained. Scholar and Scopus More may increase the coverage of potentially cited material, but it also introduces and intertwines citations accrued to different kinds of data objects. In our analysis of top-scoring versus low-scoring teams at TRECVid, we showed how top-performing methods, or techniques that were successful on sample datasets at TRECVid, went on to perform significantly better than middle or low scoring teams with regard to published paper citation counts [1]. In this sense, papers reports on 'successful' techniques also went on to be highly cited themselves, providing a different form of evidence of the value of citation data, in addition to studies correlating 'Peer Review' or 'Expert Judgement' [12] against citation data etc.

Our original study utilised excel to work with datasets of bibliometric data obtained from POP. We used excel to group article titles in order to remove duplicates, group by publication venues in order to ascertain top venues for publication, and search and highlight authors belonging to top research teams etc. An issue with the data from PoP was not only the existence of duplicates, but also incomplete data about publication venues which required individually searching article titles online and then correcting accordingly. Two recent offerings; 'Google Refine' [46], and Stanford's 'Data Wrangler' [47] are applications more suited to messy bibliometric datasets, and for manipulating and transforming bibliometric datasets using criteria, facets and filters such as text filters.

As metrics become increasingly influential in assigning funding and career advancement, their development, an understanding of their limitations, and proper use, become increasingly important. Recent popular examples of unease with the current state of bibliometric include a recent letter in JASIST, Põder [48] and opinion piece in NATURE [49]. At the same time, a survey of current studies in Library and Information Science related journals would highlight the increased activity given by researchers to this area. The analysis presented in this paper has highlighted some practical methodological issues concerning the strengths and weakness of various techniques for ascertaining scholarly impact.

8. Acknowledgements

This material is based upon work supported by Science Foundation Ireland under Grant No. 07/CE/I1147. With thanks to Julia Barrett and Joseph Greene of UCD Library for their invaluable assistance to this project.

9. References

[1] Thornley CV, Johnson AC, Smeaton AF, Lee h. The Scholarly Impact of TRECVid (2003 – 2009). Journal of the American Society for Information Science and Technology. 2011 ;

[2] Harzing AW, Wal R van der. A Google Scholar h-index for journals: An alternative metric to measure journal impact in economics and business [Internet]. Journal of the American Society for Information Science and Technology. 2009 ;60(1):41–46.[cited 2011 Jun 22] Available from: http://onlinelibrary.wiley.com/doi/10.1002/asi.20953/full

[3] Franceschet M. The role of conference publications in CS [Internet]. Communications of the ACM. 2010 Dec ;53(12):129.[cited 2011 Feb 28] Available from: http://portal.acm.org/citation.cfm?doid=1859204.1859234

[4] Moed HF, Visser MS. Developing bibliometric indicators of research performance in computer science: An exploratory study [Internet]. CWTS, Leiden. Available from: http://ict.nwo.nl/files.nsf/pages/NWOA_78NJ63/$file/CWTS_Computer_Science_Study.pdf

[5] Freyne J, Coyle L, Smyth B, Cunningham P. Relative status of journal and conference publications in computer science [Internet]. Communications of the ACM. 2010 Nov ;53(11):124.[cited 2010 Nov 3] Available from: http://doi.acm.org/10.1145/1839676.1839701
[6] Garfield, E. Do Nobel prize winners write citation classics? Current Contents (23): 182-187
[7] Moed, H.F. Citation analysis in research evaluation. Dordrecht. 2005: The Netherlands; Springer.
[8] De Bellis N. Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics. 2009: USA; Scarecrow Press.
[9] Bar-Ilan J. Web of Science with the Conference Proceedings Citation Indexes: the case of computer science [Internet]. Scientometrics. 2010 Jan 12;83(3):809-824.[cited 2011 Jun 13] Available from: http://www.springerlink.com/index/10.1007/s11192-009-0145-4
[10] Ding Y, Cronin B. Popular and/or prestigious? Measures of scholarly esteem [Internet]. Information Processing & Management. 2011 Jan ;47(1):80-96.Available from: http://linkinghub.elsevier.com/retrieve/pii/S0306457310000087
[11] Franceschet M. The difference between popularity and prestige in the sciences and in the social sciences: A bibliometric analysis [Internet]. Journal of Informetrics. 2010 Jan ;4(1):55-63.Available from: http://linkinghub.elsevier.com/retrieve/pii/S1751157709000698
[12] Li J, Sanderson M, Willett P, Norris M. Ranking of library and information science researchers: Comparison of data sources for correlating citation data, and expert judgments [Internet]. Journal of Informetrics. 2010 ;4554-563.[cited 2011 Jun 13] Available from: http://linkinghub.elsevier.com/retrieve/pii/S1751157710000593
[13] Gargouri Y, Hajjem C, Larivière V, Gingras Y, Carr L, Brody T, et al. Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research [Internet]. PLoS ONE. 2010 Oct 18;5(10):e13636.[cited 2010 Oct 18] Available from: http://dx.plos.org/10.1371/journal.pone.0013636
[14] Henderson M, Shurville S, Fernstrom K. The quantitative crunch: The impact of bibliometric research quality assessment exercises on academic development at small conferences [Internet]. Campus-Wide Information Systems. 2009 ;26(3):149-167.[cited 2010 Oct 14] Available from: http://www.emeraldinsight.com/10.1108/10650740910967348
[15] Ciber. Evaluating the usage of E-Journals in the UK. 2008 ;(November):1-18.
[16] Raan AFJ. Performance-related differences of bibliometric statistical properties of research groups: Cumulative advantages and hierarchically layered networks [Internet]. Journal of the American Society for information science and technology. 2006 Dec ;57(14):1919–1935.[cited 2011 Jun 13] Available from: http://onlinelibrary.wiley.com/doi/10.1002/asi.20389/full
[17] Harzing A. Citation analysis across disciplines: The Impact of different data sources and citation metrics [Internet]. 2010 ;[cited 2011 Feb 23] Available from: http://www.harzing.com/data_metrics_comparison.htm
[18] RIA. The appropriateness of key performance indicators to research in arts and humanities disciplines. Ireland's contribution to the European Debate . [Internet]. Royal Irish Academy. 2011 Available from: http://www.ria.ie/getmedia/2d1c1172-fc9d-4492-aa3b-97581f10c035/Key-Performance-Indicators-2011-Full-PDF.pdf.aspx.
[19] Harzing, A. Publish or Perish, 2010: version 3.0, available from www.harzing.com/pop.htm
[20] Meho LI, Rogers Y. Citation counting, citation ranking, and h-index of human-computer interaction researchers: A comparison of Scopus and Web of Science [Internet]. Journal of the American Society for Information Science and Technology. 2008 ;59(11):1711–1726.[cited 2011 Jun 14] Available from: http://onlinelibrary.wiley.com/doi/10.1002/asi.20874/full [21]
[21] Schroeder R. Pointing Users Toward Citation Searching: Using Google Scholar and Web of Science [Internet]. portal: Libraries and the Academy. 2007 ;7(2):243-248.[cited 2011 Feb 21] Available from: http://muse.jhu.edu/content/crossref/journals/portal_libraries_and_the_academy/v007/7.2schroeder.html
[22] Jacsó P. Google Scholar revisited [Internet]. Online Information Review. 2008 ;32(1):102–114.[cited 2011 Feb 23] Available from: http://www.emeraldinsight.com/journals.htm?articleid=1711361&show=abstract
[23] Levine-Clark M, Gil E. A comparative analysis of social sciences citation tools [Internet]. Online Information Review. 2009 ;33(5):986-996.[cited 2011 Feb 17] Available from: http://www.emeraldinsight.com/10.1108/14684520911001954
[24] Meho LI, Yang K. Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar [Internet]. Journal of the American Society for Information Science and Technology. 2007 Nov ;58(13):2105-2125.Available from: http://doi.wiley.com/10.1002/asi.20677
[25] Emerald. Academic search engines Part: 3 [Internet]. 2009 ;[cited 2011 Feb 23] Available from: http://www.emeraldinsight.com/librarians/info/viewpoints/search_engines.htm?part=3&

[26] Hirsch JE. An index to quantify an individual□s scientific research output. [Internet]. Proceedings of the National Academy of Sciences of the United States of America. 2005 Nov ;102(46):16569-72.Available from:http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1283832&tool=pmcentrez&rendertype=abstract
[27] Brody T, Harnad S. Earlier web usage statistics as predictors of later citation impact [Internet]. Journal of the American Society for. 2006 ;440(8):1060-1072.[cited 2011 Jun 14] Available from: http://onlinelibrary.wiley.com/doi/10.1002/asi.20373/full
[28] Swan A. the open access citation advantage [Internet]. 2010. [cited 2011 Jun 21] Available from: http://eprints.ecs.soton.ac.uk/18516
[29] Franceschet M. The skewness of computer science [Internet]. Information Processing & Management. 2011 Jan ;47(1):117-124.[cited 2011 Feb 28] Available from: http://linkinghub.elsevier.com/retrieve/pii/S0306457310000233
[30] Wainer J, Przibisczki de Oliveira H, Anido R. Patterns of bibliographic references in the ACM published papers [Internet]. Information Processing & Management. 2011 Jan ;47(1):135-142.[cited 2011 Feb 16] Available from: http://linkinghub.elsevier.com/retrieve/pii/S0306457310000610
[31] Jacso P. Deflated, inflated and phantom citation counts [Internet]. Online Information Review. 2006 ;30(3):297-309.[cited 2011 Jun 22] Available from: http://www.emeraldinsight.com/10.1108/14684520610675816
[32] Bar-Ilan J. Which h-index? — A comparison of WoS, Scopus and Google Scholar [Internet]. Scientometrics. 2007 Nov 28;74(2):257-271.[cited 2011 Jun 19] Available from: http://www.springerlink.com/index/10.1007/s11192-008-0216-y
[33] Yuan J, Wang H, Xiao L, Zheng W, Li J, Lin F, et al. A Formal Study of Shot Boundary Detection [Internet]. IEEE Transactions on Circuits and Systems for Video Technology. 2007 Feb ;17(2):168-186.Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4079667
[34] Hui SC, Fong a CM. Document retrieval from a citation database using conceptual clustering and co-word analysis [Internet]. Online Information Review. 2004 ;28(1):22-32.[cited 2011 Feb 4] Available from: http://www.emeraldinsight.com/10.1108/14684520410522420
[35] Sugimoto CR, McCain KW. Visualizing changes over time: A history of information retrieval through the lens of descriptor tri-occurrence mapping [Internet]. Journal of Information Science. 2010 Jun ;36(4):481-493.[cited 2011 Jan 13] Available from: http://jis.sagepub.com/cgi/doi/10.1177/0165551510369992
[36] Taporware [hompage on the Internet]. 2011. Available from: http://taporware.mcmaster.ca/
[37] The R Project for Statistical Computing' [homepage on the Internet]. 2011. Available from: http://www.r-project.org/
[38] Franceschet M. A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar [Internet]. Scientometrics. 2010 Jun 10;83(1):243–258.[cited 2011 Jun 22] Available from: http://www.akademiai.com/index/t5444739wt4pv550.pdf
[39] Thelwall M. Bibliometrics to webometrics [Internet]. Journal of Information Science. 2008 Jun ;34(4):605-621.[cited 2011 Feb 16] Available from: http://jis.sagepub.com/cgi/doi/10.1177/0165551507087238
[40]Bornmann L, Mutz R, Neuhaus C, Daniel H. Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results [Internet]. Ethics in Science and Environmental Politics. 2008 Jun ;893-102.Available from: http://www.int-res.com/abstracts/esep/v8/n1/p93-102 /
[41] Neuhaus C, Daniel H-D. Data sources for performing citation analysis: an overview [Internet]. Journal of Documentation. 2008 ;64(2):193-210.[cited 2011 Jan 17] Available from: http://www.emeraldinsight.com/10.1108/00220410810858010
[42] Cronin B. Bibliometrics and beyond: some thoughts on web-based citation analysis [Internet]. Journal of Information Science. 2001 Feb ;27(1):1-7.[cited 2010 Nov 21] Available from: http://jis.sagepub.com/cgi/doi/10.1177/016555150102700101
[43] Open Researcher and Contributor ID [homepage on the Internet]. 2011. Available from: http://www.orcid.org/
[44] 'Datacite' [homepage on the Internet]. 2011. Available from: http://www.datacite.org/
[45] Kousha K, Thelwall M, Rezaie S. Using the Web for research evaluation: The Integrated Online Impact indicator [Internet]. Journal of Informetrics. 2010 Jan ;4(1):124-135.[cited 2011 Jun 14] Available from: http://linkinghub.elsevier.com/retrieve/pii/S1751157709000777
[46] Google Refine [homepage on the Internet]. 2011. Available from: http://code.google.com/p/google-refine/
[47] Data Wrangler [homepage on the Internet]. 2011. Available from: http://vis.stanford.edu/wrangler/
[48] Põder E. Let□s correct that small mistake [Internet]. Journal of the American Society for Information Science. 2010 ;61(12):2593-2594.[cited 2011 Feb 23] Available from: http://onlinelibrary.wiley.com/doi/10.1002/asi.21438/abstract
[49] Lane J. Let□s make science metrics more scientific. [Internet]. Nature. 2010 Mar ;464(7288):488-9.Available from: http://www.ncbi.nlm.nih.gov/pubmed/20336116

