promise

PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

Deliverable 6.4
Report on the impact analysis for the CLEF initiative

Version 0.4, 31 August 2013

SEVENTH FRAMEWORK PROGRAMME

Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191

Document Information

Deliverable number:	D6.4
Deliverable title:	Report on the impact analysis for the CLEF initiative
Delivery date:	31/08/2013
Lead contractor for this deliverable	RSLIS
Author(s):	Theodora Tsikrika, Birger Larsen, Georgeta Bordea, Paul Buitelaar
Participant(s):	RSLIS, HES-SO, UNIPD, NUIG
Workpackage:	WP6
Workpackage title:	Evaluation activities
Workpackage leader:	RSLIS
Dissemination Level:	PU – Public
Version:	0.4
Keywords:	Impact analysis, bibliometric study, citation analysis, Google Scholar, Scopus, Saffron

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
0.1	22/07/13	Draft	Theodora Tsikrika (RSLIS), Birger Larsen (RSLIS)	First draft – overall structure, method description, CLEF results
0.2	18/08/13	First version	Theodora Tsikrika (RSLIS), Birger Larsen (RSLIS)	Added ImageCLEF results
0.3	26/08/13	Second version	Theodora Tsikrika (RSLIS), Birger Larsen (RSLIS)	Summaries and conclusions added
0.4	31/08/13	Third version	Theodora Tsikrika (RSLIS), Birger Larsen (RSLIS), Georgeta Bordea (NUIG), Paul Buitelaar (NUIG)	Content analysis added; internal review comments incorporated

Abstract

Evaluation campaigns have been widely credited with contributing tremendously to the advancement of information access systems by providing the infrastructure and resources that support researchers in the development of new approaches. Measuring the impact of such benchmarking activities is crucial for assessing which of their aspects have been successful, and thus obtain guidance for the development of improved evaluation methodologies and information access systems. The goal of this deliverable is to develop methodologies that measure the scholarly impact of evaluation campaigns and to apply the established workflow (i) for assessing the scholarly impact of the first ten years of CLEF activities (2000-2009), and (ii) for extending a previous study on the scholarly impact of ImageCLEF. We choose to measure the success of CLEF by the scientific impact of the research they foster, i.e., the publications derived from it and the citations they receive. We perform a study of the publication output and of the citation impact using data from Scopus and Google Scholar. Our bibliometric analysis of the CLEF 2000–2009 Proceedings indicates a significant impact of CLEF, particularly for its well-established Adhoc, ImageCLEF, and Question Answering labs, and for the lab/task overview publications that attract considerable interest. The high impact of the overview publications further indicates the significant interest in the created resources and the developed evaluation methodologies, typically described in such papers. In addition, the large number of derived publications published at other venues such as conference and in journals receive even more citations on average than official CLEF publications. This indicates the widespread appeal and use of resources built in the context of ImageCLEF activities irrespective of where they have been published. Finally, our analysis has highlighted the differences between the available citation analysis tools, and the difficulties encountered in constructing suitable baselines against which to measure the relative impact of multidisciplinary evaluation campaigns.

Table of Contents

Executive Summary

Motivation

PROMISE organises and supports CLEF, an annual evaluation campaign promoting research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information with various levels of structure. Such campaigns have been widely credited with contributing tremendously to the advancement of information access systems by providing the infrastructure and resources that support researchers in the development of new approaches, and by promoting the exchange of ideas. Measuring the impact of such benchmarking activities is crucial for assessing which of their aspects have been successful, and thus obtain guidance for the development of improved evaluation methodologies and information access systems.

Goals

The goal of this deliverable is to develop methodologies that measure the scholarly impact of evaluation campaigns and to apply the established workflow (i) for assessing the scholarly impact of the first ten years of CLEF activities (2000-2009), and (ii) for extending a previous study on the scholarly impact of ImageCLEF.

Methods

Given that contribution to the field of evaluation campaigns is mainly indicated by the research that would otherwise not have been possible, we choose to measure the success of CLEF by the scientific impact of the research they foster. The scientific impact of research is commonly measured by its scholarly impact, i.e., the publications derived from it and the citations they receive. Existing work in bibliometrics and scientometrics has mainly focussed on assessing the scholarly impact of specific publication venues (e.g., journals and conference proceedings) or of the research activities of individual authors, institutions, countries, or particular domains. Only few studies have examined the scholarly impact of evaluation campaigns. We perform a study of the publication output and of the citation impact using data from Scopus and Google Scholar.

Results

Our bibliometric analysis of the CLEF 2000–2009 Proceedings indicates a significant impact of CLEF, particularly for its well-established Adhoc, ImageCLEF, and Question Answering labs, and for the lab/task overview publications that attract considerable interest. The high impact of the overview publications further indicates the significant interest in the created resources and the developed evaluation methodologies, typically described in such papers. Our analysis of ImageCLEF also includes a detailed study of the associated working notes as well as derived publication at external venues. The results show that there is a significant number of working notes papers and derived publications , but that much higher citation impact is achieved by the CLEF proceedings papers as well as the derived publications, with the latter showing the highest impact (13,6 citations per publication on average). This indicates the widespread appeal and use of resources built in the context of ImageCLEF activities irrespective of where they have been published. Finally, our analysis has highlighted the differences between the available citation analysis tools, and the difficulties encountered in constructing suitable baselines against which to measure the relative impact of multidisciplinary evaluation campaigns. An alternative content-based analysis allows zooming into subfields and examining more closely the topics and issues dealt with in publications that cite CLEF.

1 Introduction

Evaluation campaigns have been widely credited with contributing tremendously to the advancement of information access systems by providing the infrastructure and resources that support researchers in the development of new approaches, and also by promoting the exchange of ideas. Over the years, several large-scale evaluation campaigns in the field of information access have been established at the international level, where major initiatives include TREC[1], CLEF[2], INEX[3], NTCIR[4], and FIRE[5]. PROMISE organises and supports CLEF, an annual evaluation campaign that aims to promote research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information with various levels of structure.

Measuring the impact of such benchmarking activities is crucial for assessing which of their aspects have been successful, and thus obtain guidance for the development of improved evaluation methodologies and information access systems. Given that their contribution to the field is mainly indicated by the research that would otherwise not have been possible, it is reasonable to consider that their success can be measured, to some extent, by the scientific impact of the research they foster. The scientific impact of research is commonly measured by its **scholarly impact**, i.e., the publications derived from it and the citations they receive.

Existing work in bibliometrics and scientometrics has mainly focussed on assessing the scholarly impact of specific publication venues [5] (e.g., journals and conference proceedings) or of the research activities of individual authors [1], institutions, countries, or particular domains [2]. Only few studies have examined the scholarly impact of evaluation campaigns; recent investigations have reported on the scholarly impact of TRECVid[6] [7] and ImageCLEF[7] [8], with the latter being performed by HES-SO in the context of the activities of the PROMISE network of excellence.

The goal of this deliverable is to develop methodologies that measure the scholarly impact of evaluation campaigns and to apply the established workflow (i) for assessing the scholarly impact of the first ten years of CLEF activities (2000-2009), and (ii) for extending the study on the scholarly impact of ImageCLEF [8]. To this end, the rest of the deliverable is organised as follows: Section 2 presents the bibliometric analysis method and tools, together with the bibliographic and citation data considered. Section 3 reports on the results of this analysis for the CLEF initiative, while Section 4 presents the results of this analysis for ImageCLEF. Section 5 presents the results of an alternative content based analysis that identifies the most influential research topics introduced by CLEF, based on an analysis of the full text of documents citing CLEF publications. Section 6 concludes this deliverable.

[1] Text REtrieval Conference (`http://trec.nist.gov/`)

[2] Cross–Language Evaluation Forum (`http://www.clef-initiative.eu/`)

[3] INitiative for the Evaluation of XML retrieval (`http://www.inex.otago.ac.nz/`)

[4] NTCIR Evaluation of Information Access Technologies (`http://ntcir.nii.ac.jp/`)

[5] Forum for Information Retrieval Evaluation (`http://www.isical.ac.in/~clia/`)

[6] TREC Video Retrieval Evaluation (`http://trecvid.nist.gov/`)

[7] CLEF Image Retrieval Evaluation (`http://www.imageclef.org/`)

2 Scholarly Impact Analysis Method

Bibliometric studies provide a quantitative and qualitative indication of the scholarly impact of a research activity by examining the number of publications derived from it and the number of citations these publications receive. Such studies typically follow these three main steps:

1 Publication data collection

2 Citation data collection

3 Data analysis

Sections 2.1-2.3 describe each of these steps, respectively, outline the strategies applied in the past, and present the approaches adopted in this work.

2.1 Publication data collection

The first step for assessing the scholarly impact of an evaluation campaign is to identify the publications associated with it and collect them in a dataset so that their citation data can then be obtained and analysed. An examination of the publications generated as a result of benchmarking activities indicates that there are typically three main types of such publications:

1. **Working Notes (WN)**: publications in the Working Notes (Notebooks) accompanying the workshop organised by each evaluation campaign as a culmination of its activities, where participants present and discuss their findings with other researchers. There are typically three types of Working Notes publications:
   a. *participant* papers where participating research groups describe their techniques and results,
   b. *overview* papers where organisers of evaluation campaigns present the evaluation resources used, summarise the approaches employed by the participating groups, and provide an analysis of the main evaluation results, and
   c. *evaluation* papers reflecting on evaluation issues, presenting other evaluation initiatives, or describing and analysing evaluation resources and experimental data.
2. **Proceedings**: publications in post-workshop Proceedings (if any), where participants publish more detailed descriptions of their approaches and perform more in–depth analyses of the results of their participation, together with further experimentation, while organisers also analyse more thoroughly the constructed evaluation resources and the generated experimental data. The same three types of publications that appear in the Working Notes (i.e., evaluation, overview, and participant) are also encountered in the Proceedings.
3. **Derived**: published in venues (e.g., journals, conferences, and workshops) outside the context of the campaign. There are typically two types of such publications:
   a. **User**: publications where resources developed in the context of the evaluation campaigns are employed for evaluating the research that is carried out,
   b. **Resources**: publications describing the resources (e.g., test collections, evaluation

metrics, etc.) developed in the context of evaluation campaigns and also discussing evaluation issues regarding the campaign in general.

In CLEF, publications of all the above types are generated[8]. In other evaluation campaigns, such as TREC and TRECVID, there are no post-workshop proceedings, but all other types of publications are encountered.

The complete lists of the *Working Notes* and *Proceedings* publications can be automatically and readily obtained from bibliographic data sources, such as DBLP; the rest need to be discovered. Our publication data collection methodology consists of the following steps:

1. Construct an initial "clean" (i.e., manually validated) set of publications *D* associated with an evaluation campaign.
2. Identify candidate publications to be added to *D*; the candidate set *C* is obtained automatically using bibliographic and citation data sources, such as Google Scholar (see Section 2.2 for a discussion):
    a. Add to *C* the publications that are retrieved when querying the data source using the name and/or the URL of the evaluation campaign (e.g., for the case of ImageCLEF use "imageclef", "www.imageclef.org", etc. as queries).
    b. (optionally) Add to *C* the publications that cite those in *D.*
3. Eliminate duplicates in *C* and remove from *C* those already in *D*.
4. Validate the publications in C. To determine whether the publications identified in the previous step can indeed be considered for inclusion in *D*, i.e., that they are actually associated with and derived from the research activities of an evaluation campaign, rather than simply mentioning and/or citing the evaluation campaign in passing, a validation step is required. This validation step is typically performed manually by an expert in the field.
5. Enlarge *D* by adding the validated publications.
6. Repeat steps 2-5 until no new publications are added.
7. Given that evaluation campaigns are typically organised as a series of evaluation *tasks* (also referred to as *labs* or *tracks*), each with a focus on a particular research area, annotate these publications with the task they relate to.

The CLEF (2000-2009) Proceedings and the CLEF (2000-2009) Working Notes publications were obtained automatically through DBLP and through the CLEF initiative website (http://www.clef-initiative.eu/), respectively. CLEF is organised as a series of evaluation *labs* (referred to as *tracks* before 2010), each with a focus on a particular research area, with some labs in turn structured into *tasks*, each with even more focussed research objectives. This organisation is reflected into the associated CLEF Proceedings and Working Notes publication lists, which are structured according to these labs and/or tasks. Therefore, the CLEF Proceedings and CLEF Working Notes publications were automatically annotated with their respective lab(s) and/or tasks(s) by exploiting the structure of the publication lists. These automatic annotations were then manually validated by an expert in

[8] To be accurate, this publication scheme was followed until 2009; in 2010 the format of CLEF changed and the there are no longer any follow–up CLEF proceedings, just the Working Notes.

the field (the first author of this deliverable) for their precision and recall.

The CLEF derived publications can be obtained by applying the iterative process outlined above. This requires the use of the bibliographic and citation data sources, described next.

2.2 Citation data collection

The most comprehensive citation data sources are:

1. Thomson Reuters (formerly ISI) Web of Knowledge (http://wokinfo.com/),
2. Scopus (http://www.scopus.com), and
3. Google Scholar (http://scholar.google.com/).

ISI and Scopus provide citation analysis tools to calculate various metrics of scholarly impact, such as the h–index [3]. Google Scholar, on the other hand, does not offer such capabilities for arbitrary publication sets; citation analysis using its data can though be performed by systems such as the *Online Citation Service* (OCS – http://dbs.uni-leipzig.de/ocs/) and *Publish or Perish* (PoP – http://www.harzing.com/pop.htm). OCS and PoP provide different querying facilities: OCS allows to upload entire publications lists, but lacks keyword-based querying, whereas PoP supports faceted search over a number of fields, but cannot find the citations of a given list of publications.

Each of these sources follows a different data collection policy that affects both the publications covered and the number of citations found. ISI has a complete coverage of more than 10,000 journals going back to 1900, but its coverage of conference proceedings or other scholarly publications, such as books, is very limited or non-existent. For instance, in the field of computer science, ISI only indexes the conference proceedings of the Springer Lecture Notes in Computer Science and Lecture Notes in Artificial Intelligence series. The citations found are also affected by its collection policy, given that in its General Search, ISI provides only the citations found in ISI-listed publications to ISI-listed publications. Scopus aims to provide a more comprehensive coverage of research literature by indexing nearly 18,000 titles from more than 5,000 publishers, including conference proceedings and "quality web sources". In its General Search, it lists citations in Scopus-listed publications to Scopus-listed publications from 1996 onwards. GS, on the other hand, has a much wider coverage since it includes academic journals and conference proceedings that are not ISI- or Scopus-listed, and also books, white papers, and technical reports, which are sometimes highly cited items as well.

As it is evident, these differences in their coverage can enormously affect the assessment of scholarly impact metrics; the degree to which this happens varies among disciplines [1,2]. For computer science, where publications in peer–reviewed conference proceedings are highly valued and cited in their own right, ISI greatly underestimates the number of citations found [5,1], given that its coverage of conference proceedings is very partial. Scopus offers broader coverage, which may though be hindered by its lack of coverage before 1996; this does not affect this study. Google Scholar offers an even wider coverage and thus further benefits citation analyses performed for the computer science field [5, 2]. As a result, this study employs both Scopus and Google Scholar (in particular its OCS and PoP wrappers) for assessing the scholarly impact of CLEF. This allows us to also explore a further goal: to compare and contrast these data sources in the context of such an analysis.

It should be noted that the reliability of Google Scholar as a data source for bibliometric studies is being received with mixed feelings [1], and some outright scepticism [4], due to its widely reported shortcomings [5, 4, 1]. In particular, Google Scholar frequently has several entries for the same publication, e.g., due to misspellings or incorrectly identified years, and therefore may deflate citation counts [5, 4]. OCS rectifies this through multiple matching and PoP through support for manual merging. Inversely, Google Scholar may also inflate citation counts by grouping together citations of different papers, e.g., the journal and conference version of a paper with the same or similar titles [5, 4]. Furthermore, Google Scholar is not always able to correctly identify the publication year of an item [4]. These deficiencies have been taken into account and addressed with manual data cleaning when possible, but we should acknowledge that examining the validity of citations in Google Scholar is beyond the scope of this study.

Once the citation data sources have been selected, the next step is to query them using the publication data as input so as to obtain the citation data. The citations for the CLEF (2000-2009) Proceedings publications were obtained in a 24-hour period in April 2013 as follows:

- *Scopus*: the query "SRCTITLE(lecture notes in computer science) AND VOLUME(*proceedings_volume*)" was entered in the Advanced Search separately for each year and the results were manually cross–checked against the publication list.
- *OCS*: the list of publications was directly uploaded into the system, which matched each publication to one or more Google Scholar entries. The result list consisting of tuples of the form <*input_publication*, *Google_Scholar_match*, *number_of_citations*> was manually refined so as to remove false positive matches. Furthermore, the citations (if any) of publications for which OCS did not find a match were manually added to the list.
- *PoP*: the proceedings title was used in the *Publication* field and the proceedings publication year in the *Year* field. The results were also manually refined by removing false positive matches, merging entries deemed equivalent, and adding the citations of unmatched publications.

The citations for the CLEF (2000-2009) Working Notes publications could only be obtained through Google Scholar, since Scopus does not index these publications. Unfortunately, halfway through our analysis, OCS stopped operating in May 2013 due to unforeseen circumstances, and therefore PoP was the only data source that could be employed. Since PoP does not enable the uploading of entire publication lists, the citations to the Working Notes publications could only be obtained by querying PoP. The following querying strategies were explored: First, the phrase "CLEF Working Notes" was used in the *Publication* field and the publication year in the *Year* field. Since Google Scholar indexes such publications from a variety of sources (e.g., researchers homepages) and not only from the official CLEF initiative website, not all such publications are correctly associated as being part of the CLEF Working Notes. This results in many of them having their *Publication* field empty and therefore this querying strategy yielded incomplete results. Similar incomplete results were obtained when "Working Notes" was used in the *Publication* field, the keyword "clef" in the *All of the words* field, and the publication year in the *Year* field. Therefore, it was decided to simply use the keyword "clef" in the *All of the words* field, and the publication year in the *Year* field. This had the opposite effect and yielded too many results, as "clef" is also a French word included in many French publications. These limitations of the PoP querying capabilities forced to re-think our goals and it was decided to consider only a subset of the CLEF Working Notes publications, and in particular those corresponding to ImageCLEF, one of its most popular labs launched in 2003, which organises the evaluation tasks relevant to the "Visual clinical decision support" PROMISE use case and for which a preliminary bibliometric analysis study had already been performed [8].

The citations for the ImageCLEF (2003-2009) Working Notes publications were thus obtained by querying PoP using the keyword "imageclef" in the *All of the words* field and the publication year in the *Year* field. This yielded the citations for several publications including not only the Working Notes, but also the Proceedings and the ImageCLEF derived publications. The PoP results were manually refined so as to eliminate the Proceedings publications, since these had already been gathered, and to identify the Working Notes and the derived publications by removing irrelevant publications. The ImageCLEF derived publications were also manually annotated with the tasks they relate to by an expert in the field (first author of this deliverable) who performed this annotation by examining their full text.

2.3 Data analysis

The analysis was performed similarly to [8] along several axes, such as the types of publications and the labs and tasks comprising the evaluation campaign while also drilling down the data into time dimension. Furthermore, the necessity of defining a baseline against which to compare the results was identified. There is no straightforward answer in determining such as baseline given the interdisciplinary nature of evaluation campaigns and the significant differences in the publishing and citing norms and practices among the different disciplines[9]. For instance, ImageCLEF focusses on the field of visual media analysis, indexing, classification, and retrieval, and to this end it develops evaluation tasks in various domains, including medical image annotation and retrieval, general image annotation and retrieval from historical archives, news photographic collections, and Wikipedia pages, robot vision, and plant identification. As a result, ImageCLEF participants originate from a number of different research communities, including (visual) information retrieval, cross–lingual information retrieval, computer vision and pattern recognition, medical informatics, and human-computer interaction, and thus their publications can be found in completely disparate "worlds". Given the differences in the publishing and citing practices between e.g., the disciplines of computer science and medicine, it is not trivial to define a baseline against which to compare the results of an ImageCLEF analysis as a whole. One solution would be to perform the analysis on the (relatively homogenous) task level. The publications and citations forming the baseline would then correspond to those in the related fields, e.g., the computer vision and pattern recognition field for the photo annotation task. Moreover, the results of such a bibliometric study could be compared to those obtained for similar evaluation campaigns. However, only one such similar study exists [7] which assesses the scholarly impact of TRECVid, an evaluation campaign which can be considered to focus on a domain similar to that of ImageCLEF. Therefore, it was decided to consider the results of this study as a baseline; see further discussion and a comparison in Section 4.4.

[9] An interesting discussion on this can be found at: `https://wiki.oulu.fi/display/tor/1.3.1.7+Evaluation+of+disciplines+and+research+fields`.

3 The Scholarly Impact of CLEF

The scholarly impact of the CLEF evaluation campaign is assessed by performing a bibliometric analysis of the citations of the CLEF 2000–2009 Proceedings publications collected through Scopus and Google Scholar. This section first provides a brief overview of the CLEF evaluation campaign (Section 3.1), then presents the results of the bibliometric analysis (Section 3.2), and finally provides a few concluding remarks (Section 3.3).

3.1 The CLEF Evaluation Campaign

Evaluation campaigns enable the reproducible and comparative evaluation of new approaches, algorithms, theories, and models, through the use of standardised resources and common evaluation methodologies within regular and systematic evaluation cycles. Motivated by the need to support users from a global community accessing the ever growing body of multilingual and multimodal information, the CLEF annual evaluation campaign, launched in 1997 as part of TREC, became an independent event in 2000 with the goal to promote research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information. To this end, it provides an infrastructure for: (i) the comparative evaluation of multilingual and multimodal information access systems, (ii) the creation of reusable resources for such benchmarking purposes, (iii) the exploration of new evaluation methodologies and innovative ways of using experimental data, and (iv) the exchange of ideas.

CLEF is organised as a series of evaluation *labs* (referred to as *tracks* before 2010), each with a focus on a particular research area, ranging from the core cross-lingual adhoc retrieval (*Adhoc*) to multilingual question answering (*QA@CLEF*), cross-language image retrieval (*ImageCLEF*), and interactive retrieval (*iCLEF*). Some labs are in turn structured into tasks, each with even more focussed research objectives. In 2010, CLEF changed its format by accompanying its labs with a peer-reviewed conference. This deliverable focusses on the first ten years of CLEF and does not consider the changes that took place thereafter.

CLEF's annual evaluation cycle culminates in a workshop where participants of all labs present and discuss their findings with other researchers. This event is accompanied by the **CLEF Working Notes**, where research groups publish, separately for each lab and task, *participant* notebook papers that describe their techniques and results. In addition, the organisers of each lab (and/or each task) publish *overview* papers that present the evaluation resources used, summarise the approaches employed by the participating groups, and provide an analysis of the main evaluation results. Moreover, *evaluation* papers reflecting on evaluation issues, presenting other evaluation initiatives, or describing and analysing evaluation resources and experimental data may also be included. These (non-refereed) CLEF Working Notes papers are available online on the CLEF website.

From 2000 to 2009, participants were invited to publish after each workshop more detailed descriptions of their approaches and more in–depth analyses of the results of their participation, together with further experimentation, if possible, to the **CLEF Proceedings.** These papers went through a reviewing process and the accepted ones, together with updated versions of the overview papers, were published in a volume of the Springer Lecture Notes in Computer Science series in the year following the workshop and the CLEF evaluation campaign.

Moreover, CLEF participants and organisers may extend their work and publish in journals, conferences, and workshops. The same applies for research groups from academia and industry that, while not official participants of the CLEF activities, may decide at a later stage to use CLEF resources to evaluate their approaches. These **CLEF derived** publications are a good indication of the impact of CLEF beyond the environment of the evaluation campaign.

Next the results of our bibliometric analysis are presented, and as discussed in Section 2.2, the focus of this study is on the analysis of the CLEF 2000-2009 Proceedings publications.

3.2 Results of the Bibliometric Analysis

The results of the bibliometric analysis of the citation data found by the three sources (OCS, PoP, and Scopus) for the 873 CLEF 2000-2009 Proceedings publications are presented in Table 3-1. Over the years, there is a steady increase in the number of publications, in line with the continuous increase in the number of offered labs (with the exception of 2007). The coverage of publications varies significantly between Scopus and Google Scholar, with the former indexing a subset that does not include the entire 2000 and 2001 CLEF Proceedings and another four individual publications, and thus contains 92% of all publications, while the latter does not index 22 (0.02%) of all publications. Table 3-2 indicates that Spain is the country that has produced the most CLEF Proceedings publications, with five of its institutions and four of its authors being among the top 10 most prolific. Although the statistics in Table 3-2 are obtained from Scopus, and therefore cover only the years 2002–2009, they can still be considered representative of the whole dataset since they describe over 90% of all publications; OCS and PoP do not readily support such analysis.

The number of citations varies greatly between Scopus and Google Scholar, with the latter finding around ten times more citations than Scopus. Overall, the total number of citations over the 873 CLEF Proceedings publications are 9,137 and 8,878 as found by OCS and PoP respectively, resulting in 10.47 and 10.17 average cites per paper, respectively. This is slightly higher, but in essence comparable to the findings of the studies on the scholarly impact of TRECVid [7] and ImageCLEF [8], with the difference that the former considers a much larger dataset (2,073 publications with 15,828 citations) that also includes TREC–derived papers, while the latter a much smaller one (249 publications with 2,147 citations).

When examining the distributions over the years, OCS and PoP reach their peak in terms of number of citations and h-index values in 2006, while Scopus does so in 2009. The average number of citations per publication peaks much earlier though, indicating that the publications of the early CLEF years have on average much more impact than the more recent ones. This could be attributed to the longer time period afforded to these earlier publications for accumulating citations. Given though the current lack of access to the citing papers through the OCS and PoP systems, only a future analysis that will monitor changes in regular intervals (e.g., yearly) could provide further insights.

Table 3-1: The citations, average number of citations per publication, and h-index of the CLEF Proceedings publications as found by the three sources.

	# labs	# publ.	OCS			PoP			Scopus		
			# cit.	avg.	h-index	# cit.	avg.	h-index	# cit.	avg.	h-index
2000	3	27	501	18.56	15	507	18.78	15	-	-	-
2001	2	37	904	24.43	17	901	24.35	17	-	-	-
2002	4	44	636	14.45	14	634	14.41	14	74	1.68	4
2003	6	65	787	12.11	15	776	11.94	15	87	1.34	5
2004	6	81	989	12.21	17	942	11.63	16	137	1.69	5
2005	8	112	1231	10.99	18	1207	10.78	17	133	1.19	5
2006	8	127	1278	10.06	18	1250	9.84	18	133	1.05	5
2007	7	116	1028	8.86	16	902	7.78	15	119	1.03	5
2008	10	131	1002	7.65	16	989	7.55	16	78	0.60	3
2009	10	133	781	5.87	12	770	5.79	12	144	1.08	5
Total	14	873	9,137	10.47	41	8,878	10.17	41	905	1.04	10

Table 3-2: Top 10 countries, affiliations, and authors of the CLEF 2002–2009 Proceedings publications as found by Scopus.

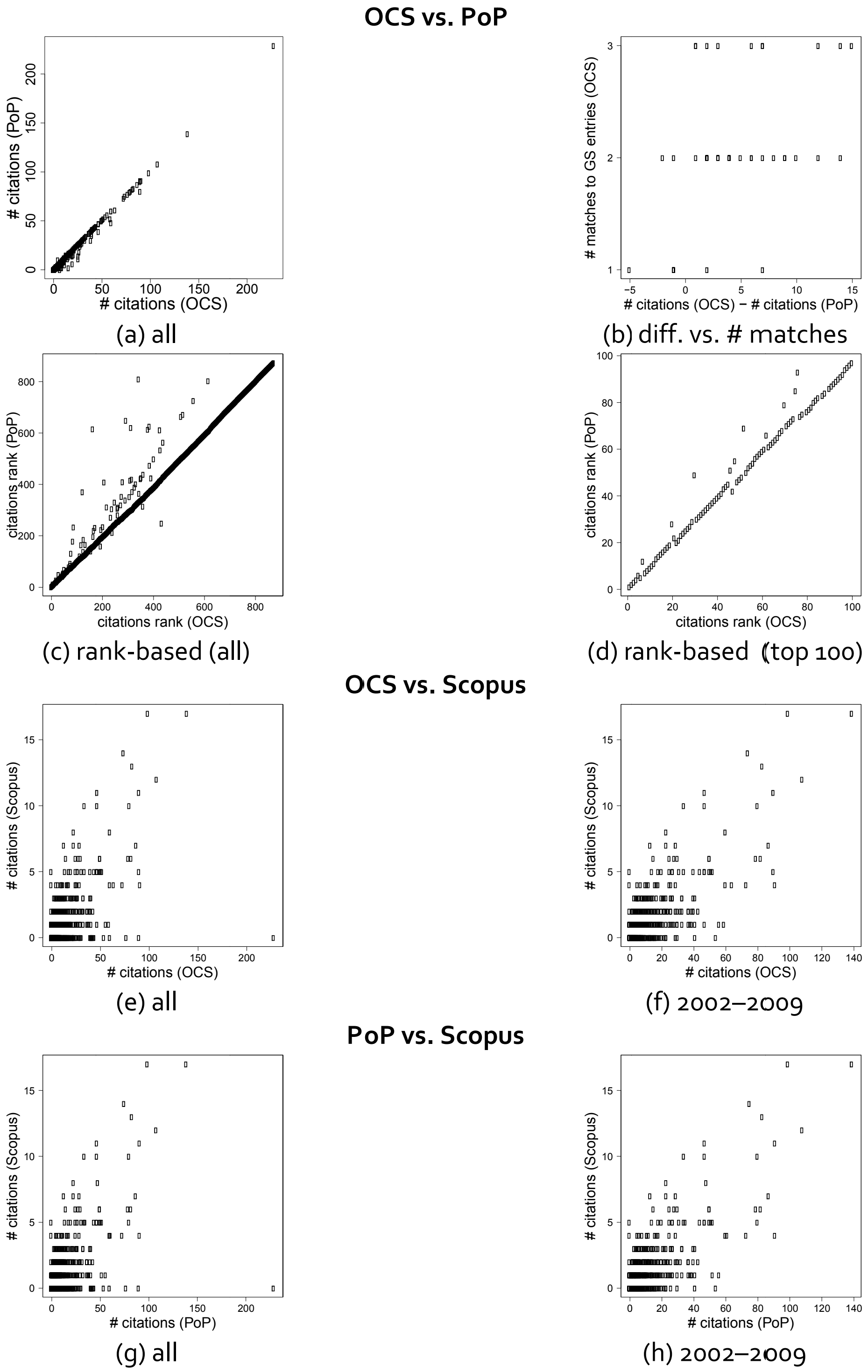
Country		Affiliation		Author	
Spain	178	Universidad de Alicante	44	Jones G.J.F.	29
Germany	105	UNED	33	Mandl T.	25
United States	93	Dublin City University	30	Llopis F.	24
France	67	University of Amsterdam	29	de Rijke M.	24
United Kingdom	61	Universidad de Jaen	27	Garcia-Cumbreras M.A.	20
Italy	55	Universität Hildesheim	25	Urena-Lopez L.A.	20
Netherlands	54	Universidad Carlos III de Madrid	24	Clough P.	19
Switzerland	52	UC Berkeley	23	Penas A.	18
Ireland	41	Universidad Politecnica de Madrid	22	Rosso P.	18
Canada	25	University of Sheffield	21	Leveling J.	17

3.2.1 Comparing Citation Data Sources

The differences between the three data sources are investigated further by examining the correlations of the citations found by them. OCS and PoP differ significantly from Scopus as indicated in Figure 3-1(e)–(f) and Figure 3-1(g)–(h), respectively, which show that there is no correlation between the number of citations each source finds for the same publication over the dataset of considered publications. In addition, the rankings based on citation counts are compared using Kendall's τ. Ties in these rankings are resolved by using the alphabetical order of either the publications' title or their authors' names. In both cases and for both pairs of data sources, Kendall's τ is less than 0.03 (p>0.2), further indicating the lack of correlation between these sources.

On the other hand, the differences between OCS and PoP are much less substantial since both rely on Google Scholar. Figure 3-1(a) shows a strong correlation between the number of citations OCS and PoP find for each publication, particularly for publications with high citation counts. This is further confirmed by Figure 3-1(c)–(d) that show the correlations between the rankings based on the citation counts over all publications and over the 100 most cited publications, respectively. Here, ties in the rankings are resolved using the titles, but similar results are obtained when using the authors' names. Also, the overlap in publications ranked by both in the top k={100,200,300,400,500} is over 96%.

Overall, OCS finds 259 (3%) more citations than PoP. The difference for a single publication ranges from 1 to 15 citations, as illustrated in Figure 3-1(b). Small differences could be attributed to changes in the Google Scholar index that may have taken place during the time period that intervened between obtaining the citation data from each source. Larger differences could be attributed to the different policies adopted by OCS and PoP for matching each input publication to a Google Scholar entry. Figure 3-1(b) plots the differences in citation counts against the number of Google Scholar matches found by OCS; the higher the difference, the more likely that OCS found more matches. This indicates that OCS achieves a slightly higher recall, and therefore OCS data will be used for the analysis performed here, unless stated otherwise.



(a) all

(b) diff. vs. # matches

(c) rank-based (all)

(d) rank-based (top 100)

(e) all

(f) 2002–2009

(g) all

(h) 2002–2009

Figure 3-1: Correlations between the citations found by the different sources.

3.2.2 Citation Distribution

Metrics such as the total number of citations and the average number of citations per publication do not allow us to gauge the impact of individual publications, given that scientific publications are typically cited to a variable extent and citation distributions across such publications are found to be highly skewed [6]. To determine the degree of citation skew and thus gain insights into the variability of the impact of particular publications, the distribution of citations into publication quartiles are examined for each year and overall.

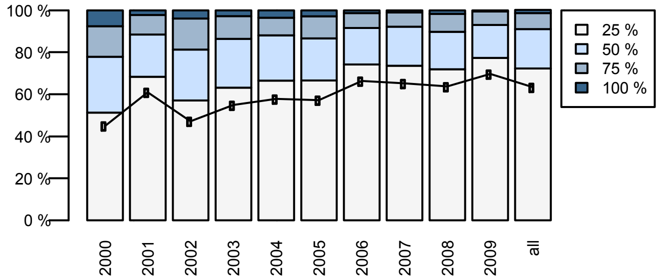


Figure 3-2: The distributions of citations found by OCS (split by quarters) over the years and overall, and the Gini coefficient of these distributions plotted as a line.

Figure 3-2: The distributions of citations found by OCS (split by quarters) over the years and overall, and the Gini coefficient of these distributions plotted as a line. indicates the relative cumulative citation count for each quartile of publications. The 25% of top cited publications account for 50 to 75% of all citations (72% on average), while the bottom 25% of publications merely attract 0.5–7.5% of all citations (1.5% on average). This citation skewness appears to be increasing over the years. For the first three years, the top 25% of publications account for less than 60% of all citations, for the next three years, for around 65% of all citations, while for the last four years, for close to 75% of all citations.

These results are corroborated by also measuring the skewness of the citation distribution using the *Gini coefficient*, a measure of statistical dispersion that reflects the inequality among values of a frequency distribution. The Gini coefficient corresponds to a nonnegative real number, with higher values indicating more diverse distributions; 0 indicates complete equality, and 1 total inequality. Its overall value of 0.63 in CLEF indicates the high degree of variability in the citations of individual publications, and this diversity is continuously increasing as indicated by the values of the Gini coefficient being below 0.5, around 0.55, and over 0.65 for the first three, next three, and final four years, respectively.

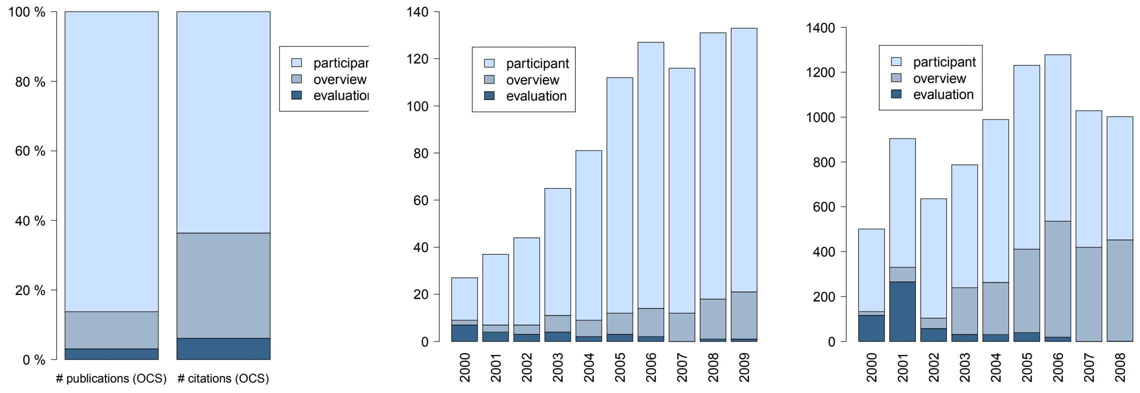
The exception to the above observations is the year 2001, which is more skewed compared to the other early CLEF years; its Gini coefficient is 0.61, while its top 25% publications account for almost 70% of all citations. This high degree of variability is due to the inclusion of two of the top 10 cited publications over all years, listed in Table 3-3, and in particular due to the domination of the most cited publication, a paper by Ellen Voorhees [9] , which achieves around 65% more citations than the second most cited publication. The remaining top cited publications in Table 3-3 are more or less evenly spread across the years.

Table 3-3: Top 10 cited publications as found by OCS: their rank and number of citations by the three sources, and their author(s), title, year, and type (E = evaluation, O = overview, P = participant). Terms in italics denote abbreviations of original title terms.

OCS / PoP / Scopus						Author(s)	Title	Year	Type
rank			# citations						
1	1	-	228	229	-	Voorhees	The Philosophy of Information Retrieval Evaluation.	2001	E
2	2	2	139	139	17	Müller et al.	Overview of the ImageCLEFmed 2006 Medical Retrieval [...]	2006	O
3	3	5	108	108	12	Clough et al.	The CLEF 2005 Cross-Language Image Retrieval Track.	2005	O
4	4	1	99	99	17	Clough et al.	The CLEF 2004 Cross-Language Image Retrieval Track.	2004	O
5	6	290	91	91	4	Vallin et al.	Overview of the CLEF 2005 Multilingual *QA* Track.	2005	O
6	5	6	90	91	11	Chen	Cross-Language Retrieval Experiments at CLEF 2002.	2002	P
7	12	29	90	80	5	Grubinger et al.	Overview of the ImageCLEFphoto 2007 [...] Task.	2007	O
8	7	-	90	90	-	Monz & de Rijke	Shallow Morphological Analysis in Monolingual *IR* [...]	2001	P
9	8	14	87	87	7	Müller et al.	Overview of the CLEF 2009 Medical Image Retrieval Track.	2009	O
10	9	4	83	83	13	Magnini et al.	Overview of the CLEF 2004 Multilingual *QA* Track.	2004	O

3.2.3 Citation Analysis of CLEF Publications Types

Figure 3-3(a) compares the relative number of publications of the three types (*evaluation*, *overview*, and *participant*) with their relative citation frequency. As also listed in the last column of Table 3-4, the participants' publications account for a substantial share of all publications, namely 86%, but only receive 64% of all citations. On the other hand, overview and evaluation publications receive three times or twice the percentage of citations compared to their publications' percentage. This indicates the significant impact of these two types; the significant impact of overview publications is further illustrated in Table 3-3 where 7 out of the 10 most cited publications are overviews, while the impact of evaluation publications can be attributed to a single publication, the Voorhees paper [9].



(a) all (b) # publications (OCS) (c) # citations (OCS)

Figure 3-3: Relative impact of different types of CLEF Proceedings publications.

Table 3-4: Relative percentages of different types of CLEF Proceedings publications and their citations over the years.

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2000–2009
	% publications										
evaluation	25.93	10.81	6.82	6.15	2.47	2.68	1.57	0.00	0.76	0.75	0.03
overview	7.41	8.11	9.09	10.77	8.64	8.04	9.45	10.34	12.98	15.04	0.11
participant	66.67	81.08	84.09	83.08	88.89	89.29	88.98	89.66	86.26	84.21	0.86
	% citations										
evaluation	23.15	29.42	8.96	3.94	3.03	3.17	1.49	0.00	0.10	0.00	0.06
overview	3.39	7.19	7.39	26.43	23.56	30.22	40.45	40.86	45.11	55.44	0.30
participant	73.45	63.38	83.65	69.63	73.41	66.61	58.06	59.14	54.79	44.56	0.64

Figure 3-3(b)–(c) and Table 3-4 drill down from the summary data into the time dimension. During the early years, CLEF Proceedings included several evaluation publications, many of them invited, which attracted a considerable number of citations, with the Voorhees [9] paper in 2001 being the most prominent example. More recently, such publications and consequently their citations have all but disappeared. The number of participants' publications has mostly followed a steady increase both in absolute and in relative terms, reaching almost 90% of all publications for some years. However, such publications manage to attract only between 44% and 74% of all citations, with the exception of 2002, where participants' publications received almost 84% of all citations. This is mostly due to a single participant's publication included among the 10 most cited publications (see Table Table 3-3). Finally, the impact of overview publications has significantly increased during the more recent years, where overviews constitute only 10 to 15% of all publications, but account for 40 to 55% of all citations.

3.2.4 Citation Analysis of CLEF Labs and Tasks

Table 3-1 presents the results of the citation analysis for the publications of the 14 labs and their tasks organised by CLEF during its first 10 years. Two more "pseudo–labs", *CLEF* and *Other* are also listed; these are used for classifying the evaluation type publications not assigned to specific labs, but rather pertaining to evaluation issues related to CLEF or other evaluation campaigns, respectively.

Three labs, *Adhoc*, *ImageCLEF*, and *QA@CLEF*, clearly dominate in terms of publication and citation numbers; they account for 67% of all publications and for 72% of all citations. They also account for 9 of the 10 most cited publications in Table 3-5. The highest number of citations per publication is observed for the Other evaluation publications, which are highly skewed due to the presence of the Voorhees [9] paper. Excluding these from further consideration, the aforementioned three labs are among the top ranked ones, together with the *Domain–Specific* and *MorphoChallenge*. Overall, the *Medical Retrieval* and *Medical Annotation* ImageCLEF tasks have had the greatest impact among all labs and tasks, closely followed by the main *QA* task and the main *Cross/Mono-lingual* Adhoc task. This also indicates a bias towards older, most established labs and tasks. Finally, the most cited publication in each lab or task is in most cases its overview, further indicating the high impact of such publications.

Figure 3-4 depicts the number of citations for the CLEF labs and tasks over the years. Although it is difficult to identify trends over all labs and tasks, in many cases there appears to be a peak in their second or third year of operation, followed by a decline. Exceptions include the *Photo Annotation* ImageCLEF task, which attracted significant interest in its fourth year when it employed a new collection and adopted new evaluation methodologies, and also the *Cross–Language Speech Retrieval* (CL–SR) lab that increased its impact in 2005 following a move from broadcast news to conversational speech. Such novel aspects result in renewed interest in labs and tasks, and also appear to strengthen their impact.

3.3 Conclusions

This bibliometric analysis of the CLEF 2000–2009 Proceedings has shown the considerable impact of CLEF during its first ten years in several diverse multi-disciplinary research fields. The high impact of the overview publications further indicates the significant interest in the created resources and the developed evaluation methodologies, typically described in such papers. It is necessary though to extend this analysis and include the Working Notes and all derived work. Finally, our analysis has highlighted the differences between the available citation analysis tools: Google Scholar provides a much wider coverage than Scopus, while OCS and PoP are in essence comparable, each with different querying facilities that might prove advantageous in different situations.

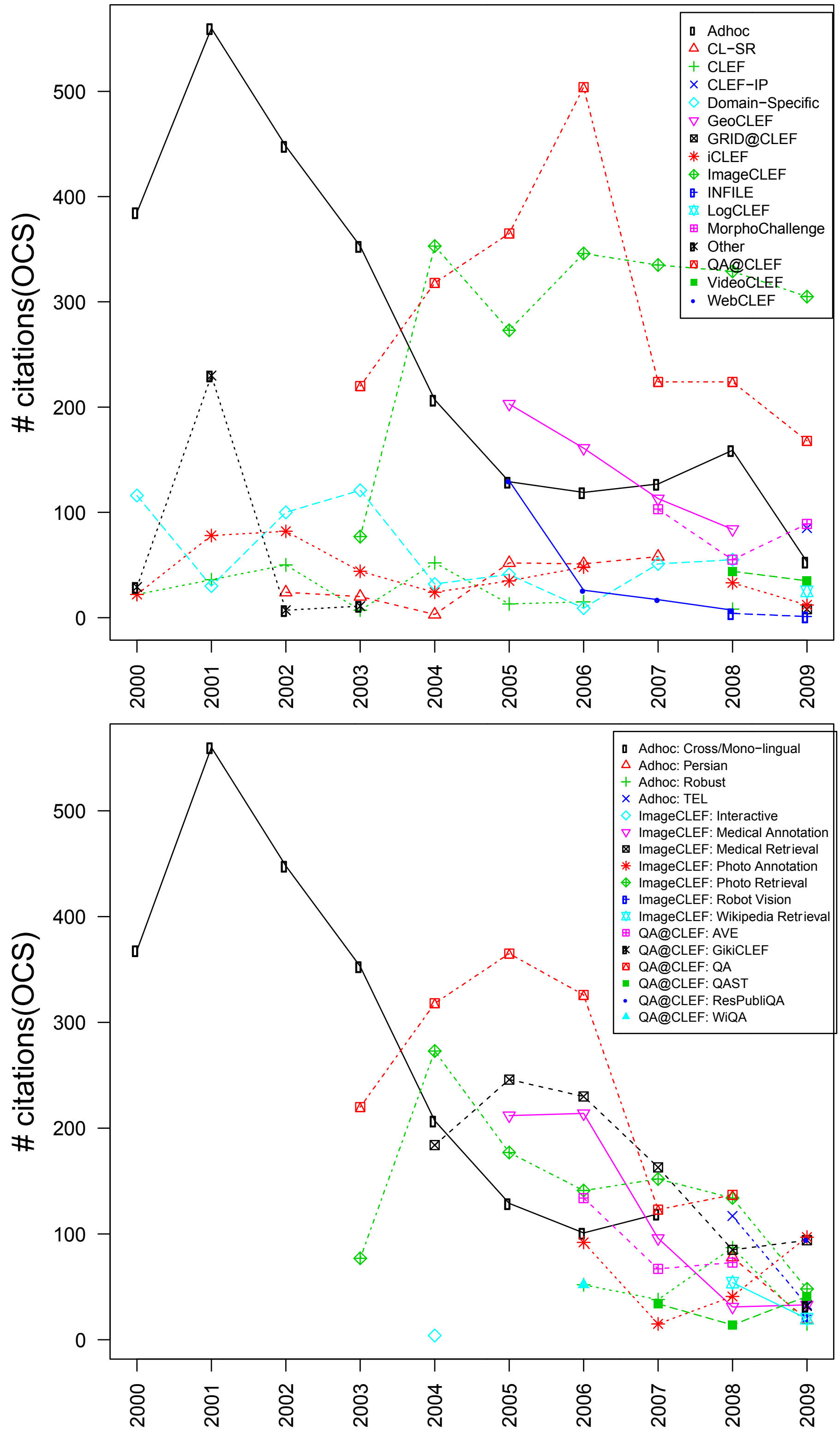


Figure 3-4: The impact of CLEF labs (top) and tasks (bottom) over the years.

Table 3-5: CLEF labs and tasks in alphabetical order, the number of years they have run, their publications, citations, average number of citations per publication, and the type of the most cited publication (E = evaluation, O = overview, P = participant). The number of publications and citations over all tasks for a lab may not sum up to the total listed for all tasks for that lab, since a publication may refer to more than one task. Similarly for the number of publications and citations over all labs

Lab	Task	#years	# publications	# citations	average	most cited
Adhoc	*(all tasks)*	10	237	2540	10.72	P
	Cross/Mono-lingual	**8**	**188**	**2285**	**12.15**	**P**
	Persian	2	11	97	8.82	O
	Robust	4	30	192	6.40	O
	TEL	2	19	150	7.89	O
CL-SR		6	29	208	7.17	O
CLEF		10	23	203	8.83	E
CLEF-IP		1	15	85	5.67	O
Domain-Specific		9	47	555	11.81	P
GeoCLEF		4	58	561	9.67	O
GRID@CLEF		1	3	8	2.67	O
iCLEF		9	41	378	9.22	O
ImageCLEF	***(all tasks)***	**7**	**179**	**2018**	**11.27**	**O**
	Interactive	1	2	4	2.00	P
	Medical Annotation	5	37	586	15.84	O
	Medical Retrieval	6	62	1002	16.16	O
	Photo Annotation	4	21	245	11.67	O
	Photo Retrieval	7	86	1002	11.65	O
	Robot Vision	1	6	23	3.83	O
	Wikipedia Retrieval	2	11	74	6.73	O
INFILE		2	8	5	0.62	O
LogCLEF		1	6	25	4.17	O
MorphoChallenge		3	20	247	12.35	P
Other		5	8	277	34.62	E
QA@CLEF	***(all tasks)***	**7**	**173**	**2023**	**11.69**	**O**
	AVE	3	25	274	10.96	O
	GikiCLEF	1	7	32	4.57	O
	QA	6	114	1489	13.06	O
	QAST	3	11	89	8.09	O
	ResPubliQA	1	10	95	9.50	O
	WiQA	1	7	52	7.43	O
VideoCLEF		2	14	79	5.64	O
WebCLEF		4	28	180	6.43	P
All		10	873	9,137	10.47	E

4 The Scholarly Impact of ImageCLEF

The scholarly impact of ImageCLEF was assessed by performing a bibliometric analysis of the citations of both the ImageCLEF publications in the CLEF 2003–2009 Proceedings and Working Notes and also the ImageCLEF derived publications in other venues. These citations were collected through Google Scholar, and in particular through PoP. First, a brief overview of the ImageCLEF lab is given (Section 4.1), followed by the results of the bibliometric analysis (Section 4.2). Next, these results are compared first to those of a similar analysis of ImageCLEF publications in the CLEF Proceedings that was performed two years ago in 2011 (Section 4.3), and then also to the results of a study that assessed the scholarly impact of TRECVid [7] (Section 4.4). Finally, some conclusions are outlined (Section 4.5).

4.1 The ImageCLEF Evaluation Campaign

ImageCLEF, the cross–language image retrieval annual evaluation campaign, was introduced in 2003 as part of CLEF and forms a natural extension to other CLEF tracks given the language neutrality of visual media. Motivated by the need to support multilingual users from a global community accessing the ever growing body of visual information, the main aims of ImageCLEF are: (i) to develop the necessary infrastructure for the evaluation of visual information retrieval systems operating in both monolingual and cross–language contexts, (ii) to provide reusable resources for such benchmarking purposes, and (iii) to promote the exchange of ideas towards the further advancement of the field of visual media analysis, indexing, classification, and retrieval.

To meet these objectives a number of tasks have been organised by ImageCLEF within two main domains: (i) medical image retrieval and (ii) general (non–medical) image retrieval from historical archives, news photographic collections, and Wikipedia pages. These tasks can be broadly categorised as follows:

- *Ad hoc image retrieval*. This simulates a classic document retrieval task: given a statement describing a user's information need, find as many relevant documents as possible and rank the results by relevance. In the case of cross–lingual retrieval, the language of the query is different from the language of the metadata used to describe the image. Ad hoc tasks have run since 2004 for medical retrieval and since 2003 for non–medical retrieval scenarios.
- *Image Annotation*. Although ad hoc retrieval is a core image retrieval task, a common precursor is to identify whether certain objects or concepts from a pre–defined set of classes are contained in an image (object class recognition), assign textual labels or descriptions to an image (automatic image annotation) or classify images into one or many classes (automatic image classification). Such tasks, including a medical image annotation, a photo annotation, and a robot vision task, have run since 2005.
- *Interactive image retrieval*. Since 2003, a user–centred task was run as a part of ImageCLEF and eventually followed by the interactive CLEF (iCLEF) track in 2005. Interaction in image retrieval can be studied with respect to how effectively the system supports users with query formulation, translation (for cross–lingual IR), document selection and examination.

Table 1 summarises the ImageCLEF tasks that ran between 2003 and 2009 and shows the number of participants for each task along with the distinct number of participants in each year. The number of participants and tasks offered by ImageCLEF has continued to grow steadily throughout the years, from four participants and one task in 2003, reaching its peak in 2009 with 65 participants and seven tasks. The number of participants, i.e., research groups that officially submit their results, is typically much smaller than the number of groups that register and gain access to the data; e.g., in 2010, 112 groups registered, but only 47 submitted results. This shows that there is a considerable interest in gaining access to the data sets of the lab. Given its multi–disciplinary nature, ImageCLEF participants originate from a number of different research communities, including (visual) information retrieval, cross–lingual information retrieval, computer vision and pattern recognition, medical informatics, and human-computer interaction. Further information can be found in the ImageCLEF book [11] describing the formation, growth, resources, tasks, and achievements of ImageCLEF.

Table 4-1: Participation in the ImageCLEF tasks and number of participants by year.

Task	2003	2004	2005	2006	2007	2008	2009
				General images			
Photo Retrieval	4	12	11	12	20	14	19
Interactive image retrieval	1	2	*2*	*3*	-	*6*	*6*
Photo Annotation				4	7	11	19
Wikipedia Retrieval						12	8
Robot Vision							7
				Medical images			
Medical Retrieval		12	13	12	13	15	17
Medical Annotation			12	12	10	6	7
	4	17	24	30	35	45	65

It should be noted that the interactive image retrieval task will not be considered as a separate ImageCLEF task in this analysis for the following reasons. During the two years (2003-2004) that it ran as part of ImageCLEF, it relied on datasets created by the photo retrieval task, and in essence it ran as its subtask; therefore, the publications associated with it can be easily attributed to the photo retrieval task, as all of them also contained experiments for the (ad hoc) photo retrieval task. The iCLEF publications (2005-2006, 2007-2009) that relied on ImageCLEF data are considered as ImageCLEF derived publications.

4.2 Results of the Bibliographic Analysis

The results of the bibliometric analysis of the citation data found by PoP for the various sets of ImageCLEF 2003-2009 publications are presented in Table 4-2. Over the years, there is a steady increase in the number of publications, in line with the continuous increase in the number of offered tasks. In total, there are 179 Proceedings and 221 Working Notes publications; this higher number of Proceedings publications is to be expected mainly for two reasons: (i) Proceedings publications undergo a reviewing process, and thus submissions may be rejected, whereas the Working Notes are unrefereed publications and thus all are accepted, and (ii) research groups typically submit separate Working Notes papers for each task in which they participate, but are encouraged to submit a single publication for inclusion in the Proceedings that describes their participation in all tasks. There are also 219 ImageCLEF derived publications, indicating that a significant number of papers, almost the same as those in the Working Notes, are published in other venues.

Table 4-2: The citations, average number of citations per publication, and h-index of the ImageCLEF publications in the CLEF Proceedings, the CLEF Working Notes, in other venues, and their combinations.

	# tasks	Proceedings				Working Notes (WN)				Derived			
		# publ.	# cit.	avg.	h-index	# publ.	# cit.	avg.	h-index	# publ.	# cit.	avg.	h-index
2003	1	5	74	14.80	4	5	3	0.60	1	1	5	5.00	1
2004	2	20	340	17.00	10	19	39	2.05	4	11	274	24.91	7
2005	3	22	265	12.05	8	27	98	3.63	6	29	418	14.41	13
2006	4	23	344	14.96	8	25	81	3.24	5	43	733	17.05	14
2007	4	29	291	10.03	9	30	107	3.57	6	42	473	11.26	11
2008	5	40	318	7.95	8	55	190	3.45	7	40	602	15.05	12
2009	6	40	305	7.63	7	60	118	1.97	5	53	474	8.94	12
Total	6	179	1,937	10.82	22	221	646	2.92	11	219	2,979	13.60	28

	# tasks	Proceedings + WN				Proceedings + Derived				Proceedings + WN + Derived			
		# publ.	# cit.	avg.	h-index	# publ.	# cit.	avg.	h-index	# publ.	# cit.	avg.	h-index
2003	1	10	77	7.70	4	6	79	13.17	5	11	82	7.45	5
2004	2	39	379	9.72	10	31	614	19.81	13	50	653	13.06	14
2005	3	49	363	7.41	10	51	683	13.39	15	78	781	10.01	15
2006	4	48	435	9.06	11	66	1077	16.32	16	91	1168	12.84	17
2007	4	59	398	6.75	10	71	764	10.76	14	101	871	8.62	14
2008	5	95	508	5.35	10	80	920	11.50	14	135	1110	8.22	16
2009	6	100	423	4.23	7	93	779	8.38	15	153	897	5.86	15
Total	6	400	2,583	6.46	23	398	4,916	12.35	35	619	5,562	8.99	35

The number of citations varies greatly between the Proceedings and the Working Notes publications, with the former having around three times more citations than the latter (1,937 vs. 646 citations), resulting in 10.82 and 2.92 average cites per paper, respectively. This could be attributed to several reasons. First of all, the Working Notes and the Proceedings publications are in essence quite similar, where the former could be seen as "preliminary" versions of the latter, which are the most definite and complete works. Therefore, a Proceedings publication is more likely to attract more citations than its Working Notes counterpart. Furthermore, it has been observed [5, 4] that Google Scholar (and thus PoP) may inflate citation counts by grouping together citations of different papers with the same or similar titles. Our analysis has indicated that this occurred very frequently for our dataset, given that many research groups use identical or near-identical titles for the Proceedings versions of their Working Notes publications. In particular, 35% of the Working

Notes publications had similar titles with their Proceedings counterparts and Google Scholar could not distinguish them. As a result, their citations were conflated and attributed in all cases to the Proceedings publication. This attribution to the Proceedings rather than to the Working Notes publication could be due to the fact that the former is obtained from Springer's website, a very reputable publisher, which is probably considered more trustworthy by Google Scholar, compared to the various websites from where it obtains the Working Notes publications.

The ImageCLEF derived publications have around 50% more citations than the Proceedings publications (2,979 vs. 1,937 citations), resulting in 13.6 average cites per paper, compared to their 10.82. This indicates the widespread appeal and use of resources built in the context of ImageCLEF activities, as it is also evident in Table 4-3 that lists the top most cited ImageCLEF papers, irrespective of where they have been published. Six out of them (indicated by the type User (U) or Resource (R)), including the most cited, have been published in venues other than the CLEF Proceedings and Working Notes. In particular, the top cited paper has almost 70% more citations than the second most cited one.

Table 4-3: Top 10 cited publications in ImageCLEF: the task(s) they relate to (MA = Medical Annotation, MR = Medical Retrieval, PA = Photo Annotation, PR = Photo Retrieval, RV = Robot Vision, WR = Wikipedia Retrieval), the number of their citations found by PoP, their author(s), title, year, and type (E = evaluation, O = overview, P = participant, R = resources, U = user).

	Task						# cit.	Author(s)	Title	Year	Type
	MA	MR	PA	PR	RV	WR					
1	X						234	Deselaers et al.	Features for image retrieval: an experimental comparison	2008	U
2	X	X					139	Müller et al.	Overview of the ImageCLEFmed 2006 medical retrieval and medical annotation tasks	2006	O
3				X			130	Grubinger et al.	The IAPR TC-12 benchmark: A new evaluation resource for visual information systems	2006	R
4				X			111	Kang et al.	Correlated label propagation with application to multi-label learning	2006	U
5		X	X	X			108	Clough et al.	The CLEF 2005 cross-language image retrieval track	2005	O
6		X		X			99	Clough et al.	The CLEF 2004 cross-language image retrieval track	2004	O
7		X		X			89	Braschler et al.	Cross-language evaluation forum: Objectives, results, achievements	2004	R
8		X					87	Müller et al.	Overview of the CLEF 2009 medical image retrieval track	2009	O
9	X						85	Keysers et al.	Deformation models for image recognition	2007	U
10				X			84	Liu et al.	Semi-supervised multi-label learning by constrained non-negative matrix factorization	2006	U

Next, the set of all ImageCLEF publications is analysed. One approach would be to consider that this set consists of all three subsets, i.e., the Proceedings, the Working Notes, and the ImageCLEF derived publications. However, it could be argued that a Working Notes publication and its Proceedings counterpart could be considered as a single piece of work, given that the former could be seen as a "preliminary" version of the latter, and as such, it is likely for both together to attract the number of citations that a single work would attract. In that case, the corresponding Working

Notes and Proceedings publications would be considered as one publication (where possible) with their citations aggregated. However, this requires a significant amount of work for identifying such corresponding publications and therefore a simpler approach would be to just not consider the Working Notes publications altogether. This slightly underestimates the total number of citations, but increases considerably the averages. In particular, when all (Proceedings, Working Notes, and ImageCLEF derived) publications are taken into account, there are 619 papers with 5,562 citations in total and 8.99 average cites per paper. On the other hand, when only the Proceedings and the ImageCLEF derived publications are analysed, there are 398 papers with 4,916 citations in total and 12.35 average cites per paper. Both these sets are considered in the remainder, but the emphasis is placed on one or the other depending on the context of our analysis.

The distributions of citations over the years reach their peak in terms of number of citations in 2006 for the Proceedings publications, the ImageCLEF derived ones, and their aggregation, with or without the Working Notes in this aggregation. Similar to the analysis in Section 3.2, the average number of citations per publication peaks much earlier though in 2004, indicating that the publications of the early CLEF years have on average much more impact than the more recent ones. This could be attributed to the longer time period afforded to these earlier publications for accumulating citations and also to inclusion of some highly cited papers (see in Table 4-3) among the relatively smaller size of publication sets of these early years.

4.2.1 Citation Analysis of ImageCLEF Publications Types

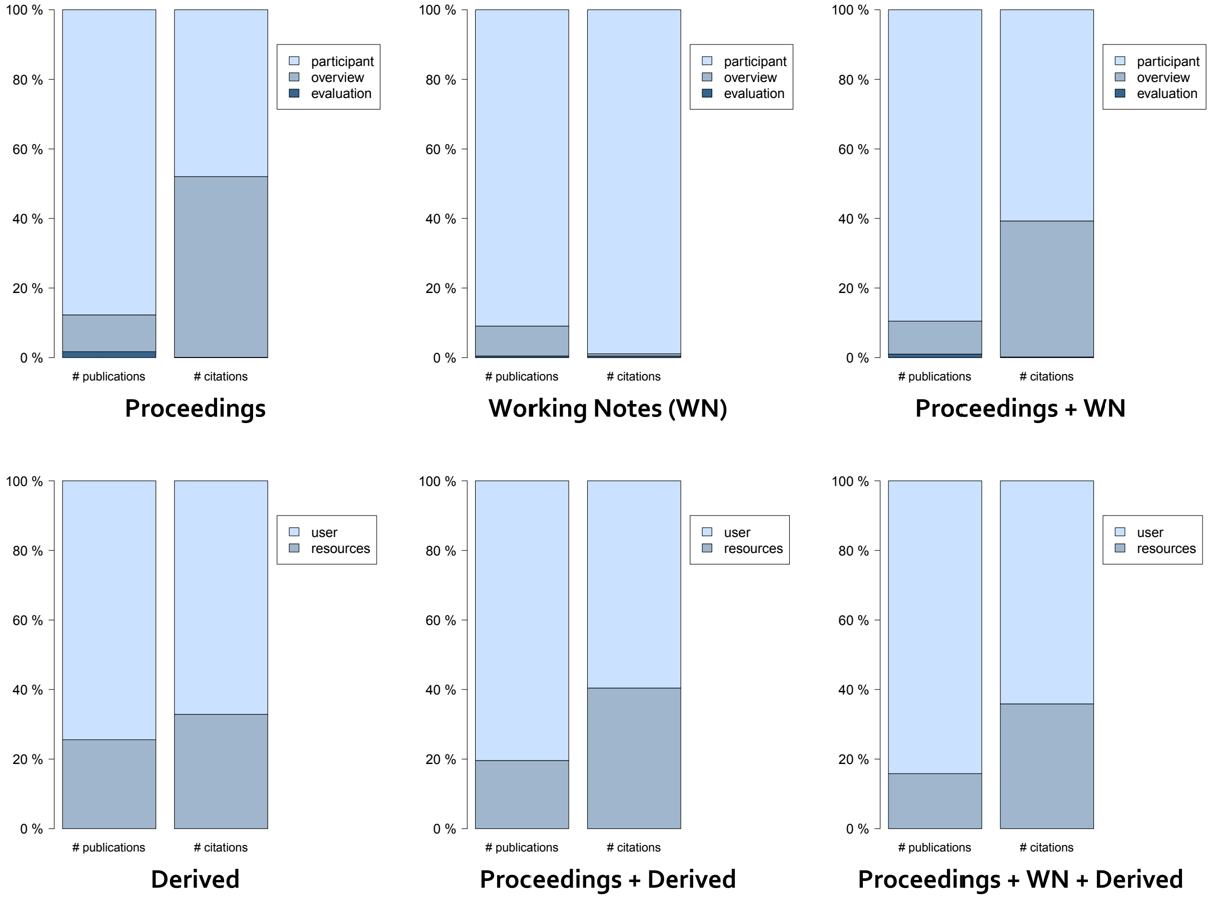
The relative number of publications of the various types are compare with their relative citation frequency in Figure 4-1 and Table 4-4.

For the Proceedings publications, the participants' papers account for a substantial share of all publications, namely 88%, but only receive 48% of all citations. On the other hand, overview publications receive fives times the percentage of citations compared to their publications' percentage. This indicates their significant impact, which is further illustrated in Table 4-3 where 4 out of the 10 most cited publications are overviews. The impact of evaluation type publications is very small given that they account for around 2% of all publications and only attract around 0.5% of all citations.

For the Working Notes, the situation is different: around 9% of all publications are overviews, but they receive less than 1% of all citations. As discussed above, this is due to Google Scholar conflating publications with similar title, a particularly frequent case for overview papers since 90% of them have identical, or near identical titles in their Proceedings and Working Notes instances; therefore any citations to either of them are attributed to their Proceedings instance. As a result, almost all citations to Working Notes publications are received by the participant papers.

For the ImageCLEF derived publications, the situation is again different without any major differences in the relative percentages. In particular, resources (user) papers account for 26% (74%) of such publications and receive 33% (67%) of all citations. Overall, the resources publications have a slightly higher impact, but not significantly so.

To analyse the publication sets consisting of both Proceedings and ImageCLEF derived publications, with or without the Working Notes, the number of publication types is reduced to two: *user* now also encompasses the participant publications and *resources* now also now includes the overview and evaluation publications. In this case, resources publications receive twice the percentage of citations compared to their publications' percentage: 20% vs. 20% when the Working Notes are not included, and 16% vs. 36% when the Working Notes are taken into account. The significant impact of such publications is also evident in Table 4-3 where six out of the 10 most cited publications are of this type.



Proceedings

Working Notes (WN)

Proceedings + WN

Derived

Proceedings + Derived

Proceedings + WN + Derived

Figure 4-1: Relative impact of different types of ImageCLEF publications sets.

Table 4-4: Relative percentages of different types of ImageCLEF publications over the years.

Proceedings								
	2003	2004	2005	2006	2007	2008	2009	2003–2009
	% publications							
evaluation	0.00	0.00	4.55	0.00	0.00	2.50	2.50	1.68
overview	20.00	5.00	4.55	8.70	10.34	12.50	15.00	10.61
participant	80.00	95.00	90.91	91.30	89.66	85.00	82.50	87.71
	% citations							
evaluation	0.00	0.00	0.38	0.00	0.00	0.31	0.00	0.10
overview	67.57	29.12	40.75	57.85	47.77	64.47	67.54	51.94
participant	32.43	70.88	58.87	42.15	52.23	35.22	32.46	47.96

Working Notes								
	2003	2004	2005	2006	2007	2008	2009	2003–2009
	% publications							
evaluation	0.00	0.00	3.70	0.00	0.00	0.00	0.00	0.45
overview	20.00	5.26	3.70	8.00	10.00	9.09	10.00	8.60
participant	80.00	94.74	92.59	92.00	90.00	90.91	90.00	90.95
	% citations							
evaluation	0.00	0.00	3.06	0.00	0.00	0.00	0.00	0.46
overview	0.00	0.00	0.00	0.00	3.74	0.00	0.00	0.62
participant	100.00	100.00	96.94	100.00	96.26	100.00	100.00	98.92

Derived								
	2003	2004	2005	2006	2007	2008	2009	2003–2009
	% publications							
resources	100	63.64	24.14	25.58	45.24	10.00	13.21	25.57
user	0	36.36	75.86	74.42	54.76	90.00	86.79	74.43
	% citations							
resources	100	89.78	26.08	35.61	45.24	10.13	17.72	32.9
user	0	10.22	73.92	64.39	54.76	89.87	82.28	67.1

4.2.2 Citation Analysis of ImageCLEF Tasks

Table 4-5 presents the results of the citation analysis for the publications in each of the six tasks organised by ImageCLEF during its first seven years.

In the Proceedings, the publications associated with the Medical Retrieval task receive the most citations, followed by the Photo Retrieval, the Medical Annotation, and the Photo Annotation tasks. On average, though, the situation is different, with the Medical Retrieval task publications receiving more citations on average and still ranked first, but followed by the Medical Annotation and Photo Annotation tasks, and then by the Photo Retrieval task.

In the Working Notes, the ranking in terms of the number of citations received is reversed for the top two positions, with the Photo Retrieval task ranked first and the Medical Retrieval ranked second. Regarding the average cites per publication though, the Medical Retrieval task is ranked first as before, followed by the Photo Retrieval, Medical Annotation, and Photo Annotation tasks.

In both the Proceedings and the Working Notes, the impact of the other two tasks, Wikipedia Retrieval and Robot Vision, is substantially smaller given the much shorter lifetime of these tasks within ImageCLEF compared to the other more established tasks.

Outside the context of CLEF, the impact of the Medical Annotation task has been particularly significant as it manages to rank second in terms of number of citations, after the Photo Retrieval task and before the Medical Retrieval task, and to rank first in terms of average cites per publication. The impact of the Wikipedia Retrieval task is also particularly notable, given that it ranks second in terms of average cites per publication.

Overall, the Photo Retrieval task publications receive the highest number of citations, followed by the Medical Retrieval, Medical Annotation, and Photo Annotation tasks. On average, though, the medical tasks receive more citations per publication, with the Medical Annotation task having the highest impact overall and outside CLEF in particular, while the Medical Retrieval task has the highest impact within CLEF. The high impact of the medical tasks is also evident in Table 4-3 where seven of the top 10 cited publications are in some way associated with them.

Figure 4-2 depicts the number of citations for the ImageCLEF tasks over the years. Although it is difficult to identify trends over all tasks, it appears that in all cases (with the exception of the Photo Annotation task), the second year of operation has a higher impact than the first. Peaks in years further down the task's lifetime appear to coincide with novel aspects in these tasks, such as the introduction of new datasets, e.g., in 2006 for the Photo Retrieval task and in 2009 for the Photo Annotation task. This trend was also observed in the case of all CLEF labs and tasks (see Section 3.2.4) where it appeared that novel aspects result in renewed interest in labs and tasks, and also appear to strengthen their impact.

Table 4-5: ImageCLEF tasks in alphabetical order, the number of years they have run, their publications, citations, average number of citations per publication, and the type of the most cited publication (E = evaluation, O = overview, P = participant, R = resources, U = user). The number of publications and citations over all tasks may not sum up to the total listed for all tasks, since a publication may refer to more than one task.

Proceedings					
Task	#years	# publications	# citations	average	most cited
Medical Annotation	5	37	566	15.30	O
Medical Retrieval	6	62	1047	16.62	O
Photo Annotation	4	21	238	11.33	O
Photo Retrieval	7	87	958	11.01	O
Robot Vision	1	6	23	3.83	O
Wikipedia Retrieval	2	11	74	6.73	O
Total	7	179	1,937	10.82	O

Working Notes					
Task	#years	# publications	# citations	average	most cited
Medical Annotation	5	42	117	2.79	P
Medical Retrieval	6	72	238	3.31	P
Photo Annotation	4	28	74	2.64	P
Photo Retrieval	7	105	338	3.22	P
Robot Vision	1	6	11	1.83	P
Wikipedia Retrieval	2	21	39	1.86	P
Total	7	221	646	2.92	P

Derived					
Task	#years	# publications	# citations	average	most cited
Medical Annotation	5	50	943	18.86	U
Medical Retrieval	6	83	895	10.78	R
Photo Annotation	4	20	158	7.90	R
Photo Retrieval	7	95	1280	13.47	R
Robot Vision	1	-	-	-	-
Wikipedia Retrieval	2	5	79	15.80	U
Total	7	219	2,979	13.60	U

Proceedings + Working Notes + Derived					
Task	#years	# publications	# citations	average	most cited
Medical Annotation	5	129	1626	12.60	U
Medical Retrieval	6	218	2180	10.00	O
Photo Annotation	4	69	470	6.81	O
Photo Retrieval	7	287	2576	8.98	R
Robot Vision	1	12	34	2.83	O
Wikipedia Retrieval	2	37	192	5.19	U
Total	7	619	5,562	8.99	U

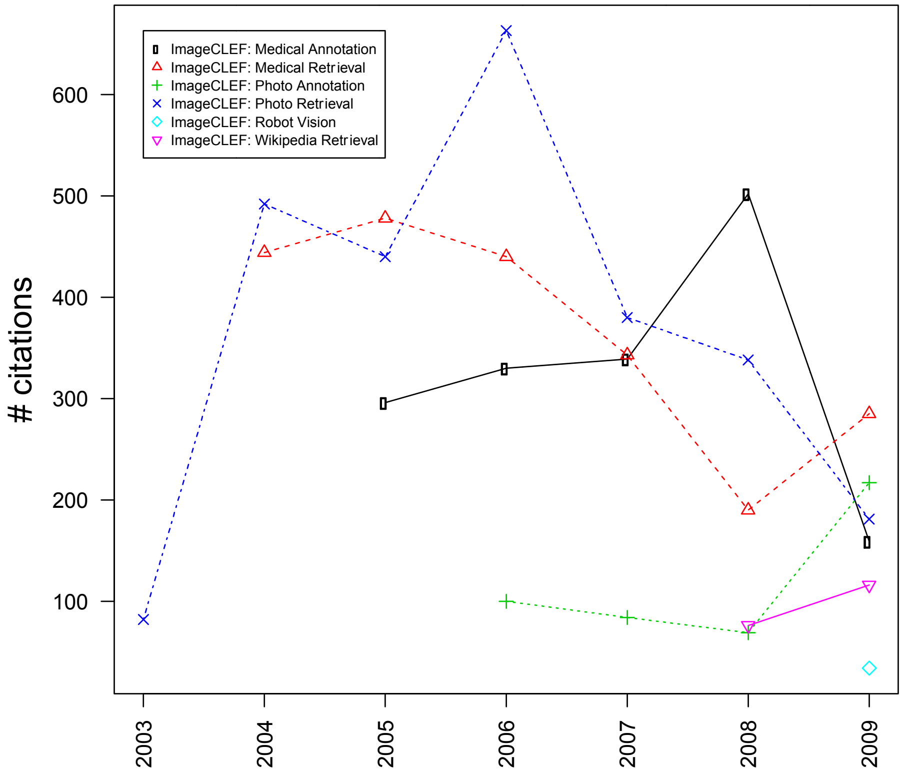


Figure 4-2: The impact of ImageCLEF tasks over the years in terms of number of citations received by their associated publications in the CLEF Proceedings, CLEF Working Notes, and in other venues.

4.2.3 Publication venues for ImageCLEF derived papers

Table 4-6 indicates that most ImageCLEF derived publications, almost half of them, appear in conferences, about a fifth in journals and another fifth in workshops, while a significant percentage (around 6%) are theses of all levels (BSc, MSc, and PhD). On average, though, journal publications have a much higher impact, 28.07 average cites per paper, compared to 12.09 for conference and 9.57 for workshop publications. This is in line with other bibliometric studies within computer science, where it has been observed that conferences are significantly more important in terms of the overall numbers of publications, but that journals, are, in fact, more important in terms of citations received on average per paper [7].

Table 4-6: Types ImageCLEF derived publications.

	Derived				
	# publications	# citations	average	% publications	% citations
Book chapter	1	0	0.00	0.46%	0.00%
Conference	101	1221	12.09	46.12%	40.99%
Demo / Panel	2	0	0.00	0.92%	0.00%
Journal	42	1179	28.07	19.18%	39.58%
Newsletter	5	13	2.60	2.28%	0.44%
Technical Report	3	10	3.33	1.37%	0.34%
Thesis (PhD/MSc/BSc)	14	68	4.86	6.39%	2.28%
Workshop	42	402	9.57	19.18%	13.49%
iCLEF	9	86	9.56	4.11%	2.89%
Total	219	2,979	13.60	100.00%	100.00%

Table 4-7 lists the venues with the highest number of ImageCLEF derived publications. The most popular corresponds to a series of ImageCLEF-focussed workshops organised in conjunction with CLEF, just before the main event. This is followed by the "Pattern Recognition Letters" journal, which was the host of a special issue dedicated to the Medical Annotation task, various PhD Theses, and the iCLEF publications in the CLEF Proceedings and Working Notes. The rest are well known conferences in the area of information retrieval.

Table 4-8 lists the venues with the highest number of ImageCLEF derived publications were published and Table 4-8 lists venues where ImageCLEF derived publications were mostly cited. These correspond to conferences and journals with high impact factors in the area of information retrieval, with the exception of "OntoImage" a workshop that is ranked high due to the Grubinger et al., "The IAPR TC-12 benchmark: A new evaluation resource for visual information systems", 2006 publication that appeared in it and that is ranked third in Table 4-3, as it has received 130 citations.

Table 4-7: Top 10 venues where ImageCLEF derived papers were mostly published.

Derived			
Venue	Type	# pub.	# cit.
MUSCLE/ImageCLEF & THESEUS/ImageCLEF workshops	workshop	13	65
Pattern Recognition Letters	journal	10	137
PhD Theses	thesis	10	68
iCLEF	CLEF	9	86
SIGIR: annual international ACM SIGIR conference on research and development in information retrieval	conference	9	139
ECIR: annual European Conference on Information Retrieval	conference	6	83
AIRS: Asian Information Retrieval Societies conference	conference	5	33
ICME: IEEE International Conference on Multimedia and Expo	conference	5	35
CORIA: COnférence en Recherche d'Information et Applications	conference	4	2

Table 4-8: Top 10 venues where ImageCLEF derived publications were mostly cited.

Derived			
Venue	Type	# cit.	# pub.
Information Retrieval	journal	323	2
SIGIR: annual international ACM SIGIR conference on research and development in information retrieval	conference	139	9
Pattern Recognition Letters	journal	137	10
OntoImage	workshop	131	2
CVPR: IEEE Conference on Computer Vision & Pattern Recognition	conference	121	2
Computerized Medical Imaging and Graphics	journal	91	2
CIVR: ACM International Conference on Image & Video Retrieval	conference	89	3
iCLEF	CLEF	86	9
IEEE Transactions on Pattern Analysis and Machine Intelligence	journal	85	1
AAAI Conference on Artificial Intelligence	conference	83	1

4.3 Assessing the Impact of ImageCLEF in 2011 and in 2013

A previous study [8] assessed the scholarly impact of ImageCLEF by performing a bibliometric analysis of the ImageCLEF publications in the CLEF 2003-2009 Proceedings; their citation data were collected in April 2011 through Scopus and PoP. Table 4-9 compares and contrasts the results of this earlier study with the results of this work using the same data sources two years later. The earlier study also took into account iCLEF publications that relied on ImageCLEF datasets or were otherwise closely related to ImageCLEF. However, the impact of these additional publications is negligible, since their citations account for less than 0.04% of all citations; these two results sets can be viewed as being comparable.

Table 4-9: Bibliometric analyses of the ImageCLEF publications published in the CLEF 2003–2009 Proceedings performed in 2011 and in 2013 using Scopus and PoP.

		#publications		# citations		average		h-index	
		2011	2013	2011	2013	2011	2013	2011	2013
Scopus	2003	5	5	13	14	2.60	2.80	2	3
	2004	20	20	50	64	2.50	3.20	4	5
	2005	25	22	24	30	0.96	1.36	3	3
	2006	27	23	25	38	0.93	1.65	2	3
	2007	29	29	18	34	0.62	1.17	3	3
	2008	45	40	14	34	0.31	0.85	2	3
	2009	44	40	38	59	0.86	1.48	4	5
	Total	195	179	182	273	0.93	1.53	6	7
PoP	2003	5	5	65	74	13.00	14.80	3	4
	2004	20	20	210	340	10.50	17.00	8	10
	2005	25	22	247	265	9.88	12.05	7	8
	2006	27	23	259	344	9.59	14.96	7	8
	2007	29	29	249	291	8.59	10.03	7	9
	2008	45	40	284	318	6.31	7.95	7	8
	2009	44	40	259	305	5.89	7.63	7	7
	Total	195	179	1,573	1,937	8.06	10.82	18	22

Because of the exclusion of iCLEF publications from the 2013 data collection, the number of publications is slightly lower (179 vs. 195). Despite of this there is a considerable increase in the number of citations over these two years: 364 (+23%) more citations are found by PoP and 91 (+50%) by Scopus. For PoP, most citations are added to the 2004 and 2006 publications, while for Scopus to the 2007–2009 ones. Overall, the impact of ImageCLEF tasks appears to increase several years after they took place, however further analysis is needed to determine whether these citations originate from papers published over these two years, or from papers simply added to the sources' indexes during this time.

4.4 Comparison to TRECVID

Assessments of the scholarly impact of other evaluation campaigns have only been performed for TRECVid (2003–2009) [7], where a list containing both the TRECVid Notebook papers and the TRECVid–derived publications was analysed. TRECVid and ImageCLEF can be considered as focussing on similar research domains. Table 4-10 compares the results of the TRECVid analysis to the results of the analysis of the CLEF proceedings and all ImageCLEF publications; this latter set consists of the ImageCLEF Proceedings publications, considered equivalent to the TRECVid NoteBook papers, and also the ImageCLEF derived publications, similarly to the TRECVid publication set.

Table 4-10: Bibliometric analyses of CLEF proceedings, all ImageCLEF, and all TRECVid [7] publications using PoP.

	CLEF Proceedings				ImageCLEF Proceedings + Derived				TRECVID Notebook + Derived			
	# publ.	# cit.	avg.	h-index	# publ.	# cit.	avg.	h-index	# publ.	# cit.	avg.	h-index
2003	65	776	11.94	15	6	79	13.17	5	64	1,066	16.66	18
2004	81	942	11.63	16	31	614	19.81	13	158	2,124	13.44	24
2005	112	1207	10.78	17	51	683	13.39	15	225	2,537	11.28	28
2006	127	1250	9.84	18	66	1077	16.32	16	361	4,068	11.27	30
2007	116	902	7.78	15	71	764	10.76	14	382	3,562	8.97	28
2008	131	989	7.55	16	80	920	11.50	14	509	1,691	3.32	16
2009	133	770	5.79	12	93	779	8.38	15	374	780	2.09	12
Total	765	6,836	8.94	38	398	4,916	12.35	35	2,073	15,828	7.63	52

Overall, there are about three times more TRECVid publications than CLEF Proceedings ones, but receive on average less citations. It is difficult though to draw conclusions given the multidisciplinary nature of CLEF coupled with the different citation practices in different domains. The number of TRECVid publications is four times that of ImageCLEF publications, but the latter receive on average many more citations per publication, 12.35 vs. 7.63. Based on these results, both appear to have had significant impact, but further investigation is needed for reaching more reliable conclusions.

4.5 Conclusions

The detailed analysis of ImageCLEF 2000-2009 shows that the number of participants and tasks offered by ImageCLEF has continued to grow steadily throughout the years, reaching its peak in 2009 with seven tasks and 65 participants from a number of different research communities. Over the years, there is a steady increase in the number of publications, in line with the continuous increase in the number of offered tasks. The number of citations received varies greatly between the Proceedings and the Working Notes publications, with the former having around three times more citations than the latter. In addition to these official publications, a large number of derived publications are published outside official CLEF publications at conferences (41%), in journals (40%) and at workshops (14%). These receive 50% more citations than the Proceedings publications and also receive more citations on average (13.6 average cites per paper vs. 10.82). On average, though, journal publications have a much higher impact, 28.07 average cites per paper, compared to 12.09 for conference and 9.57 for workshop publications. This indicates the widespread appeal and use of resources built in the context of ImageCLEF activities irrespective of where they have been published. Again we see that overview and evaluation publications receive a large share of the citations. Many of the highly cited papers are the Medical Retrieval and Photo Retrieval tasks (including their respective annotation tasks).

Analysing temporal developments in impact it appears that the second year of operation has a higher impact than the first. Peaks in years further down the task or lab's lifetime appear to coincide with novel aspects in these tasks, such as the introduction of new datasets resulting in renewed interest in labs and tasks, and also appear to strengthen their impact. Comparing the number of citations collected in 2011 to 2013 shows a considerable increase in the number of citations over these two years: 364 (+23%) more citations are found by PoP and 91 (+50%) by Scopus. Finally, comparing the results of a similar study of TRECVid publications to those of ImageCLEF we observe that TRECVid published many more papers, but that the average number of citations per paper was higher for ImageCLEF.

5 Content analysis of citing documents

The most widely accepted approaches for measuring and analysing scientific research rely on publication metadata, focusing on publication counts or the number of citations. However, textual descriptions of scientific research, such as publication titles, abstracts and, increasingly, full-text content call for methods that allow a deeper content-based analysis of scientific output. As discussed above in Section 2.3, establishing a baseline for multidisciplinary evaluation campaigns such as CLEF is a significant challenge that needs further research. The bibliometric analysis showed widespread use of the CLEF resources and a large number of citations to CLEF publications. As an alternative to comparing the impact to a benchmark, we study the scientific topics expressed in documents citing CLEF publications. This type of analysis would allow us to gain detailed insight about the topics being worked with the documents that cite CLEF publications. The goal is to estimate knowledge transfer and uptake by identifying influential topical areas that are often mentioned by other communities.

Content analysis of citing publications is performed using Saffron[10], a system that provides insight into a research community or organisation by extracting its main topics of investigation and the relations between them. Saffron was previously used to analyse interdisciplinarity in the WebSci community [11], a goal that is similar to ours. Saffron makes use of statistical measures that are sensitive to the amount of text that is available about a domain of interest. Therefore we need to identify a significant number of full content publications. Due to copyright restrictions, many of the citing documents are not readily available for this analysis. As a proof of concept, we choose to identify as many documents as possible that cite CLEF publications from CiteSeerX. CiteSeerX[11] is an autonomous citation indexing service that crawls the web to identify freely available scientific

[10] http://saffron.deri.ie/

[11] http://csxstatic.ist.psu.edu/

documents. CiteSeerX identifies publications often in PDF format, cashes them, and extracts metadata and citation information to build a citation index. There is a good chance that CiteSeerX will include CLEF publications and those citing them as CiteSeerX has a good coverage of computer science. CiteSeerX regularly releases their entire citation index including full text extracted from the PDFs[12]. We work on the latest dataset released from June 2012 containing 2.118.180 documents, and investigate the following two questions:

- Which research communities are influenced by CLEF?
- Which are the most influential research topics introduced by CLEF?

5.1 Related work

According to [12] a bibliometric map can be constructed by analysing various types of items including journals, papers, authors, and descriptive terms. The work presented in this paper is based on a basic assumption in bibliometric mapping [12], which states that a research field can be described by a list of important keywords. While previous work made use of author assigned key phrases and already built domain taxonomies [13], we applied an automatic method [14] for the extraction of domain terms as such resources are not readily available for our dataset.

Implicit relations between the extracted topical descriptors can be discovered and described through word co-occurrence analysis, a content analysis technique that was effectively applied to analyse interactions in different scientific fields [15, 13]. This technique was applied to analyse the interconnections between a main field, i.e., fuzzy logic theory, and other computing techniques [16], a setting that is similar to our analysis of CLEF citing publications. A more recent work on co-word analysis [17] outlines several limitations related to the use of keywords and proposes a method to integrate expert knowledge into the process. A main issue with this approach is that it requires a considerable amount of human intervention for the construction of domain specific thesauri. We alleviate this challenge by completely automating the process of identifying topical descriptors and by automatically constructing a topical hierarchy.

5.2 Methods

5.2.1 Data gathering

We used two strategies to identify documents that cite CLEF publications (*citing documents*). First, we extracted a list of titles of all official CLEF publications (*source publications*) based on the data collected above in Section 2.1. Both Proceedings and working notes were included. We matched these one by one against directly against the metadata in CiteSeerX. Using the CiteSeerX citation index we then identified citing documents. Secondly, we searched the titles of the bibliographical references for the source publication titles to identify any additional citing documents that were not linked to the source publications. As CiteSeerX is based on automatically extracted data from papers found on the Internet, in many cases the source publications could not be matched directly because of small differences in e.g. paper titles. Manual inspection of non-matching indicates that more sophisticated cleaning and matching techniques are needed to increase recall. We identified 998 citing documents and choose to focus our efforts in examining the suitability of Saffron for this type of analysis, rather than to attempt achieving a more comprehensive coverage. The results in

[12] See http://csxstatic.ist.psu.edu/about/data

Table 3-1 above showed that Google Scholar is able to identify approximately 9000 documents that cite CLEF publications. Google Scholar can be expected to have a somewhat better coverage than CiteSeerX, but it is likely that recall from CiteSeerX can be significantly improved with more effort. Given that we did exact match on titles our 998 document subset is small, but also of high quality with low risk of false matches included.

5.2.2 Data Processing using Saffron and Gephi

The full text of the identified citing documents was extracted from the CiteSeerX dataset and processed by Saffron. We used the topic extraction component, analysing multiword topics of up to 5 words. We used the ACM Subject Classification to build linguistic patterns for terms in Computer Science. The Saffron analysis yielded 97,458 candidate phrases, with an average of 98 candidates per document. Only the best ranked 10% of these terms are considered in our analysis. This threshold was necessary because the quality of terms influences the quality of the topic hierarchy: it is important to choose meaningful terms before analysing the relations between them.

The most representative 20 terms are the following:

- information retrieval
- search engine
- image retrieval
- retrieval system
- information retrieval systems
- language model
- QA system
- content-based image retrieval
- natural language processing
- natural language
- question answering system
- information sources
- training data
- image retrieval systems
- target language
- machine translation
- query expansion
- retrieval task
- data set
- document retrieval

As you can see, even a relatively short list such as this one is difficult to analyse, because many of the topics are closely related and redundant. Take for example the terms "QA system" / "question answering system", and "natural language "/"natural language processing". Topical hierarchies provide a more succinct summary of research topics, organising them from broader concept to more specific ones.

The index used in co-word analysis to measure the strength of relationships between two research terms is defined as:

$$I_{ij} = D_{ij} / (D_i D_j)$$

where D_i is number of articles that mention the term T_i in our corpus, D_j is number of articles that mention the term T_j, and D_{ij} is the number of documents in which both terms appear.

Edges are added in the research terms graph for all the pairs that appear together in at least 10 documents, in a window of 10 words. Saffron uses a generality measure to direct the edges from generic concepts to more specific ones. This step results in a highly dense, noisy directed graph that is further trimmed using an optimal branching algorithm. An optimal branching is a rooted tree where every node but the root has in-degree 1, and that has a maximum overall weight. This algorithm was successfully applied for the construction of domain taxonomies in [18]. This yields a tree structure where the root is the most generic term and the leaves are the most specific terms.

We used a network graph tool, Gephi, to build a graph showing links between terms: nodes are extracted terms and edges are the relations between them. This allows us to identify 'clusters' of closely related terms. We used the Yifan Hu algorithm to layout the graph, and eccentricity to weight node importance. The eccentricity of a node is higher for central nodes, that are farthest away from the leaves of the graph. We filtered all the nodes that have a degree smaller or equal than 2, for a cleaner visualisation of central nodes. This resulted in a hierarchy with 1406 nodes and 1342 edges.

5.2.3 Results

The entire topical hierarchy extracted from publications citing CLEF is visualised in Figure 5-1. The most generic term is "information retrieval", which is the root of the hierarchy. The topical hierarchy reflects the main contributions of the CLEF campaigns, "test collections" and "retrieval tasks", with a focus on "European languages". Prominent subfields include Question Answering, Image Retrieval, and Machine Translation. Next, we take a closer look at each of these fields. Figure 5-2 shows a close-up of the Question answering subfield. Here the main identified nodes are concerned with semantic representations, networks and relations, different types of answers as well as information extraction. The close-up of the Image retrieval subfield in Figure 5-3 shows a focus on issues related to the ImageCLEF evaluation campaign as well as techniques, feature extraction appropriate datasets. Figure 5-4 indicates that query translation, translation systems and models as well as differences between monolingual, bilingual and multilingual settings to be important for the machine translation subfield.

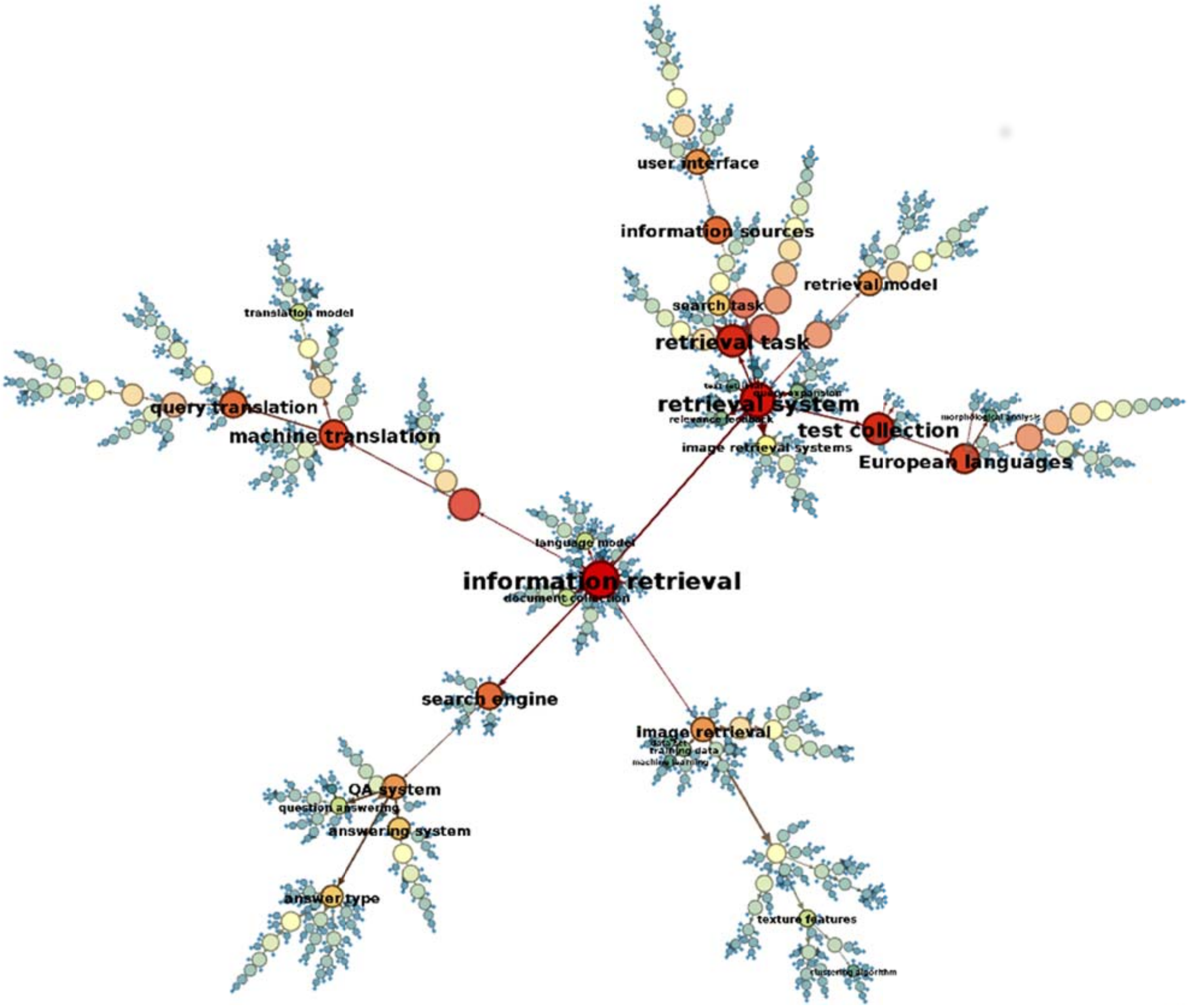


Figure 5-1: Topical hierarchy extracted for CLEF citing publications.

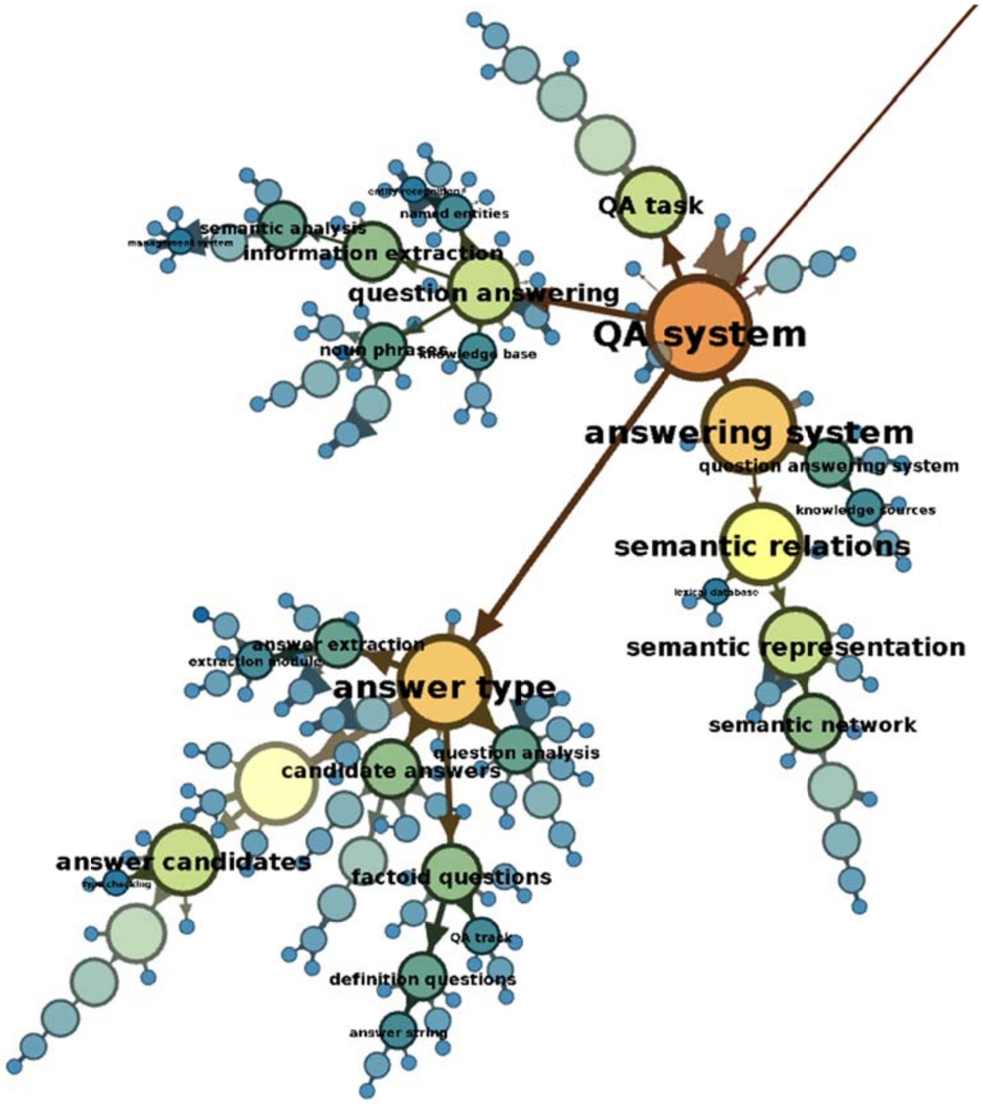


Figure 5-2: Question answering subfield.

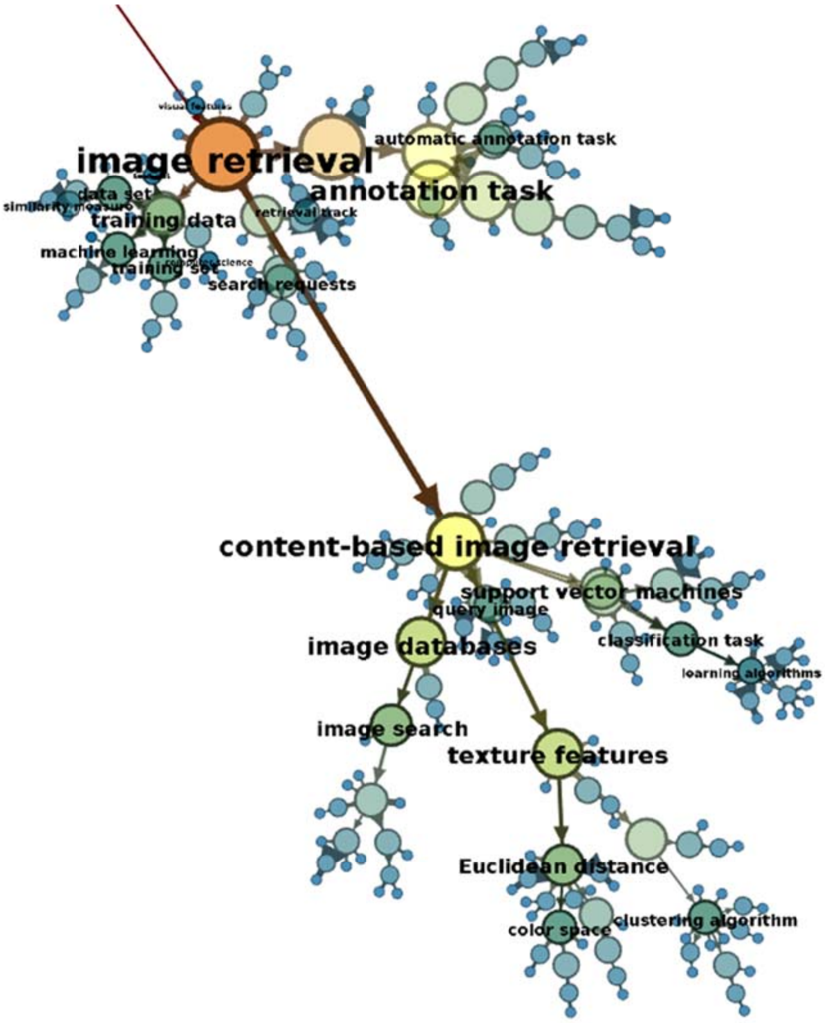


Figure 5-3: Image retrieval subfield.

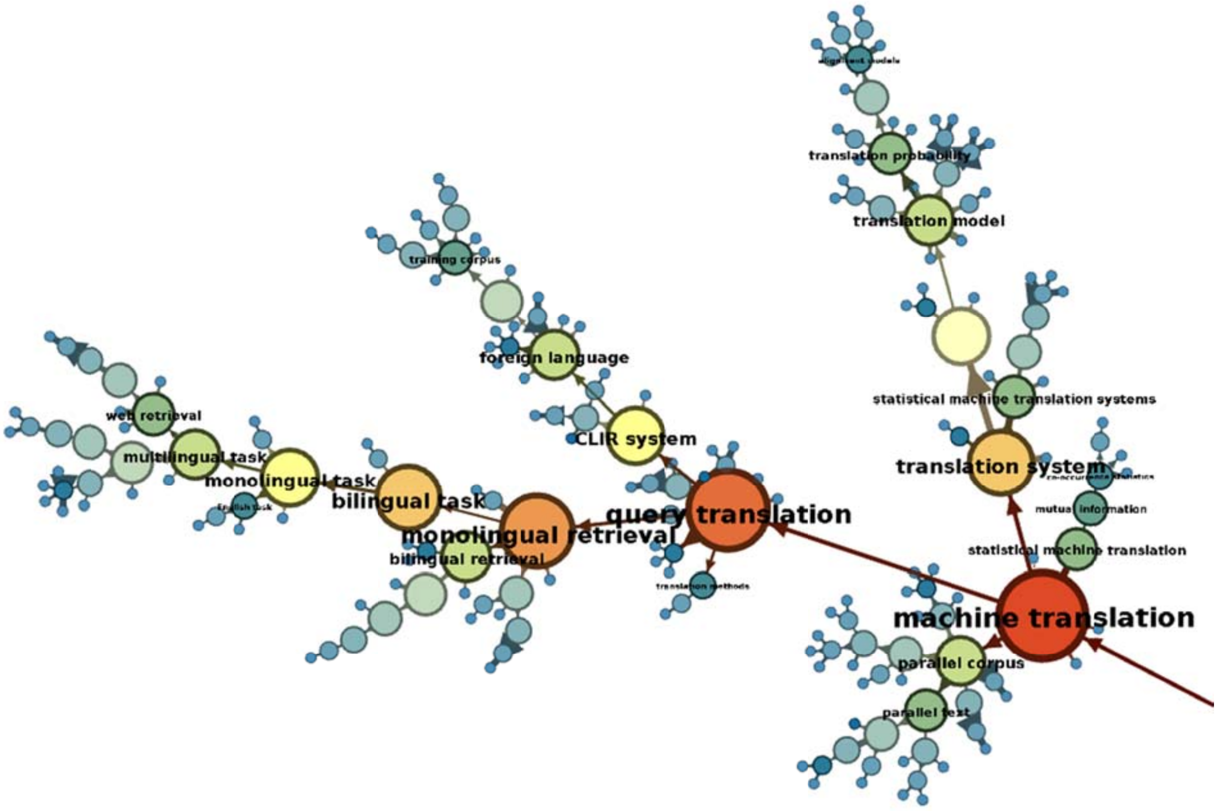


Figure 5-4: Machine translation subfield.

5.2.4 Conclusions

The overall topical hierarchy extracted using Saffron produces structures that correspond very well to the distribution of tasks and labs in CLEF (see e.g. Table 3-5), and provides insights into the focus of these. The results demonstrate that this type of content-based analysis supplements bibliometric results very well. A clear advantage of the topical hierarchy is that it allows zooming into subfields and examining more closely the topics and issues dealt with in publications that cite CLEF. It should be noted that in this initial analysis we have not excluded CLEF publications as citing publications. The produced topical hierarchy thus reflects both CLEF publications themselves as well as CLEF derived publications published elsewhere. A next step could be putting more effort into identifying a larger proportion of citing documents and then compare topical hierarchies of CLEF publications vs. derived publications. The high quality of the hierarchy produced in this initial analysis indicates that interesting results are likely.

6 Conclusions

Measuring the impact of evaluation campaigns may prove useful for supporting research policy decisions by determining which aspects have been successful, and thus obtaining guidance for the development of improved evaluation methodologies and systems. Our bibliometric analysis of the CLEF 2000–2009 Proceedings indicates a significant impact of CLEF, particularly for its well-established Adhoc, ImageCLEF, and Question Answering labs, and for the lab/task overview publications that attract considerable interest. The high impact of the overview publications further indicates the significant interest in the created resources and the developed evaluation methodologies, typically described in such papers.

Our analysis of ImageCLEF also includes a detailed study of the associated working notes as well as derived publication at external venues. The results show that there is a significant number of working notes papers and derived publications , but that much higher citation impact is achieved by the CLEF proceedings papers as well as the derived publications, with the latter showing the highest impact. This indicates the widespread appeal and use of resources built in the context of ImageCLEF activities irrespective of where they have been published. It is worth mentioning that from 2010 onwards only working notes papers are available as CLEF no longer produces proceedings after this date. It will be interesting to compare this change in future work to see whether the impact remains relatively stable.

Our analysis has highlighted the differences between the available citation analysis tools, and the difficulties encountered in constructing suitable baselines against which to measure the relative impact of evaluation campaigns. This is a significant challenge because of the multidisciplinary nature of evaluation campaigns. As an alternative we carried out an initial content-based analysis of the documents that cite CLEF publications. The results demonstrate that this type of content-based analysis supplements bibliometric results very well, as it allows zooming into subfields and examining more closely the topics and issues dealt with in publications that cite CLEF.

7 References

[1] J. Bar-Ilan. Which h-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2):257–271, 2008.

[2] A.-W. Harzing. Citation analysis across disciplines: The impact of different data sources and citation metrics, 2010. Retrieved from http://www.harzing.com/data_metrics_comparison.htm.

[3] J. E. Hirsch. An index to quantify an individualâ€™s scientific research output. *Proceedings of the National Academy of Sciences (PNAS)*, 102(46):16569â€“–16572, 2005.

[4] P. Jacsó. Deflated, inflated and phantom citation counts. *Online Information Review*, 30(3):297–309, 2006.

[5] E. Rahm and A. Thor. Citation analysis of database publications. *SIGMOD Record*, 34:48–53, December 2005.

[6] P. O. Seglen. The skewness of science. *JASIS*, 43(9):628–638, 1992.

[7] C. V. Thornley, A. C. Johnson, A. F. Smeaton, and H. Lee. The scholarly impact of TRECVid (2003–2009). *JASIST*, 62(4):613–627, 2011.

[8] T. Tsikrika, A. G. Seco de Herrera, and H. Müller. Assessing the scholarly impact of ImageCLEF. In *Proceedings of the 2nd CLEF conference*, pages 95–106, 2011.

[9] E. M. Voorhees. The philosophy of information retrieval evaluation. In *Proceedings of the 2nd CLEF Workshop*, pages 355–370, 2002.

[10] H. Müller, P. Clough, T. Deselaers, and B. Caputo, editors. ImageCLEF: Experimental Evaluation in Visual Information Retrieval. Springer, 1st edition, 2010.

[11] C. J. Hooper, G. Bordea, and P. Buitelaar. *Web Science and the Two (Hundred) Cultures: Representation of Disciplines Publishing in Web Science*. (2013).

[12] K. Börner, C. Chen, and K. W. Boyack. (2003), Visualizing knowledge domains. *Ann. Rev. Info. Sci. Tech.*, 37: 179–255. doi: 10.1002/aris.1440370106

[13] N. Coulter, I. Monarch and S. Konda. (1998), Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science*, 49: 1206–1223.

[14] G. Bordea and P. Buitelaar. 2010. DERIUNLP: A context based approach to automatic keyphrase extraction. In: *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10).* Association for Computational Linguistics, Stroudsburg, PA, USA, 146-149

[15] M. Callon J. P. Cortial, W. A. Turner and S. Bauin: From Translations To Problematic Networks – An Introduction To Co-Word Analysis, *Social Science Information Sur Les Sciences Socials* 22(2) (1983), 191–235.

[16] A. G. Lopez-Herrera, M. J. Cobo, E. Herrera-Viedma and F. Herrera. (2010). A bibliometric study about the research based on hybridating the fuzzy logic field and the other computational intelligent techniques: A visual approach. *International Journal of Hybrid Intelligent Systems*, 17, 17–32.

[17] Z.-Y. Wang, G. Li, C.-Ya. Li, A. Li (2012). Research on the semantic-based co-word analysis. *Scientometrics*, 90, 855-875.

[18] R. Navigli, P. Velardi, and S. Faralli. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In: *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - (IJCAI'11)*, Toby Walsh (Ed.), Vol. 3. AAAI Press 1872-1877.