On Synergies Between Information Retrieval and Digital Libraries

Maristella Agosti[(✉)] [iD], Erika Fabris [iD], and Gianmaria Silvello [iD]

Department of Information Engineering, University of Padua, Padua, Italy
{maristella.agosti,erika.fabris,gianmaria.silvello}@unipd.it

Abstract. In this paper we present the results of a longitudinal analysis of ACM SIGIR papers from 2003 to 2017. ACM SIGIR is the main venue where Information Retrieval (IR) research and innovative results are presented yearly; it is a highly competitive venue and only the best and most relevant works are accepted for publication. The analysis of ACM SIGIR papers gives us a unique opportunity to understand where the field is going and what are the most trending topics in information access and search.

In particular, we conduct this analysis with a focus on Digital Library (DL) topics to understand what is the relation between these two fields that we know to be closely linked. We see that DL provide document collections and challenging tasks to be addressed by the IR community and in turn exploit the latest advancements in IR to improve the offered services.

We also point to the role of public investments in the DL field as one of the core drivers of DL research which in turn may also have a positive effect on information accessing and searching in general.

Keywords: Trends in digital libraries (DL) ·
Trends in information retrieval (IR) ·
Emerging interrelationships between DL and IR

1 Introduction

The area of Digital Libraries (DL) is a multidisciplinary research field and Information Retrieval (IR) is a research area which, amongst other things, addresses methods and technologies for accessing digital information persistently preserved in digital libraries and archives. As a matter of fact, preservation would be useless without means and techniques of accessing stored information, to find and extract useful data. There has always been a strong interrelationship among IR and DL, and one significant example of this is the major organization which promotes IR, the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM SIGIR),[1] which is also one of the co-promoters

---

[1] http://sigir.org/.

of one of the major international forums focusing on digital libraries, the Joint Conference on Digital Libraries (JCDL).[2] Moreover, the programs of all conferences on DL also cover IR topics, and many methods and techniques designed and developed by the IR research community are used in the design and building of DL. Another significant example of this synergy is the ideation and development of the first "online public access catalogues" (OPACs) [2,6,11], which made broad use of state-of-the-art IR methods in a typical DL setting. Thus, as stated by Edie Rasmussen in [9], "it would appear that digital library developers have the opportunity to incorporate the best results of information retrieval research".

The goal of this work is to conduct a longitudinal analysis of relevant IR research results with the aim of understanding the influence of IR research on the theory and practice of DL; moreover, we aim to identify, if possible, the topics that seem most important among those that the two areas are going to address together in the future. We are convinced that by studying IR trends we can garner useful information about the future trends of the DL community.

Hence, this study targets two main research questions:

1. Is current IR research relevant to DL?
2. Does DL research affect IR?

In order to answer these questions, we explore the recent evolution of IR research. To do so, we created a corpus containing all papers that have been published in the proceedings of the last 15 years of the ACM SIGIR Conference, one of the most relevant conferences in IR. We extracted useful information from the main text of published papers, the bibliography, topics and keywords. Afterwards, we conduced statistical analyses to identify research trends in IR and discover if there are DL topics discussed by the IR community.

The remainder of the paper is organized as follows: in Sect. 2 we present some related work; in Sect. 3 we describe the process of creating the corpus and present some statistics on the collected data; in Sect. 4 we investigate the relationships between DL and IR; and in Sect. 5 we summarize the results of our study, make some final remarks and reveal our future directions.

2   Related Work

Bibliometrics and topic analysis of corpora of scientific conference papers are quite common especially in information science. These analyses attract the researchers attention since they can offer an interesting viewpoint on the evolution of the topics addressed in the scientific areas related to the conferences.

As an example, in 2003, for the commemoration of the 25th anniversary of ACM SIGIR Conference, Smeaton et al. [10] did a content analysis on all papers presented at the ACM SIGIR Conferences from its beginning up to 2002; Smeaton et al. investigated the evolution of topics over the selected period and provided the most central author in the co-authorship graph. Five years later,

---

[2] http://www.jcdl.org/.

Hiemstra et al. [5] provided further analysis on papers presented in the ACM SIGIR Conferences from 1978 to 2007 focusing on other aspects such as the contributions of countries in terms of number of papers published and common aspects, such as the analysis of the co-authorship graph.

The general results we report in this paper complement the results presented in [5,10], because we present general results for the period from 2003 to 2017 that is not fully covered in the other two works. Furthermore, in order to conduct our study and answer our two research questions, rather than rely on an already populated database, we chose to create and populate our own database so as to have all the different data and metadata needed to carry out the study. Unlike [5,10] we also analyzed the synergy between IR and DL. Some considerations of interest for this further study are contained in [3], where key trends that can have implications in DL were identified and highlighted, i.e. new emerging technologies such as big data techniques on search, the growth in importance of documents containing information other than articles such as datasets, presentations, the growth in importance of other sources of data such as media and the recent broadening of technology providers of information. Other considerations have been reported in [9] where, by studying the IR research history, the synergy between IR and DL is shown and some IR key challenges that can have an impact on DL research are highlighted, i.e. multilingual and cross-lingual search, retrieval in non-textual formats, use of relevance feedback, comprehensive retrospective research and user interactive IR systems.

3 Dataset Creation and Statistics

The steps needed to gather together and organize the corpus used for the study are:

1. data collection and database design;
2. data processing and data storage.

These two steps are outlined below.

We collected all the ACM SIGIR conference proceeding papers published between 2003 and 2017 by downloading them from the ACM digital library[3] as PDF files. The collection includes long papers, short papers/posters, workshop papers and tutorials. As a result our data set consists of 2974 distinct papers.

After having built the corpus, the next step was to design and build the database for storing all the data needed to conduct the analyses of interest for our study: the general details and affiliations of authors, the denomination and location of institutions, metadata of ACM SIGIR papers (DOI, year of publication, title, abstract, keywords, ACM Computing Classification System – CCS rev. 2012[4] – categories), and lists of references for each paper. It is worth noting that the database keeps track of every change of affiliation of the authors in the years of interest.

---

[3] https://dl.acm.org/.
[4] https://dl.acm.org/ccs/ccs.cfm.

Table 1. Number of papers published in ACM SIGIR proceedings from 2003 to 2017 and number of active authors per year.

Year	Number of papers	Number of active authors
2003	106	252
2004	133	293
2005	138	341
2006	152	335
2007	221	482
2008	206	461
2009	207	501
2010	217	504
2011	238	559
2012	224	562
2013	210	512
2014	229	588
2015	204	524
2016	234	647
2017	255	705

We encoded the PDF files of the corpus by obtaining structured TEI-encoded XML files.[5] This pre-processing step was performed by using the open source GROBID library [1,7].

Subsequently, we parsed the TEI-encoded files by a set of custom methods to extract the raw data about papers and authors and to store them in the database. These methods were designed to allow interactive controls in order to manually solve possible homonyms amongst different authors and to detect possible errors introduced by GROBID.

After this process, we performed a manual inspection of the stored information in order to find and fix possible inconsistencies (e.g. different denominations or abbreviations for the same institution were solved). As a result, we obtained a clean database storing all the information necessary to conduct our analyses.

Before going deeper into the content analysis and examination of synergy between Digital Library and Information Retrieval, we present relevant statistics that provide a visual insight of the information within the database that clearly demonstrates the growth in activity within the IR community over the last 15 years.

Our database stores a total of 2974 papers published in the ACM SIGIR proceedings, 1155 of those are full papers and the other 1819 are short papers, posters, demos, tutorials and workshops. Table 1 compares the evolution in the number of published papers and the number of active authors. The amount of

---

[5] http://www.tei-c.org/index.xml.

publications increased noticeably from 2003 to 2007, then, from 2008 to 2017 there were some minor fluctuations with a minimum of 206 and a maximum of 255 papers per year.

Moreover, the number of active authors per year has been growing considerably suggesting that IR is emerging as a primary research area and that the number of researchers that choose to work in this field is growing.

This consideration is strengthened by the inspection of the distribution of the country contributions which are reported in Fig. 1 and that highlight the worldwide expansion of IR research. The contribution of each country was computed by considering the number of papers in which there is at least one author affiliated with an organization/institution located in that country. The countries with the highest number of publications are the United States, with a percentage of presence in author affiliations of 37%, and the Republic of China, with a percentage of presence of 14%. It is worth noting that located in these two countries are (were) the most active and central organizations in the research area, such as Microsoft and some universities, such as Tsinghua University and the University of Massachusetts Amherst.

4 Data Analysis

In this section we investigate the research trends in IR and the role of DL research within the IR community. We focus on the evolution of the most used terms, keywords and topics in ACM SIGIR papers from 2003 to 2017.

### 4.1 Most Used Keywords and Cited Terms

We analyzed the evolution of term frequencies in the title and abstract fields and the frequencies of the keywords generated by the authors in the long papers published in the ACM SIGIR Conference between 2003 and 2017.

We used the Terrier IR Platform,[6] an open source search engine which provides indexing and retrieval functionalities [8], to process the text documents of the corpus. Paper titles and abstracts have been extracted by removing stopwords (we used the Terrier default stoplist), then they were indexed and finally word frequencies were calculated.

Table 2 shows the most used words in title and abstract fields alongside the most used keywords extracted from the text of the papers of the corpus.

We can see that the most common and consolidated topics are "Evaluation", "Learning to Rank" or "Web Search", but it is also interesting to note that some topics were preponderant for a certain period of time and then vanished – a noticeable example is "Diversity" which appeared in 2010 and lasted for 4 years – whereas other topics have appeared only in recent years, such as "Twitter" which appeared in 2009 and was the second most used keyword in 2012 although since it has not been used as much in the past (as shown in Fig. 2).

---

[6] http://terrier.org/.
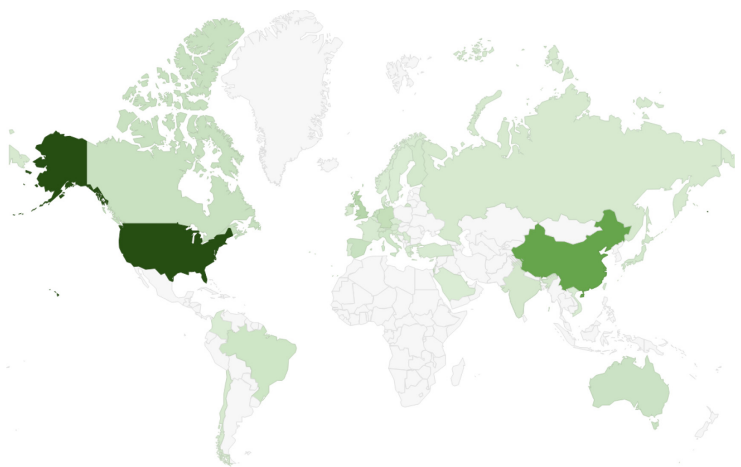
Country	Contribution Percentage
USA	37%
Republic of China	14%
United Kingdom	8%
Canada	4%
Singapore	4%
Spain	3%
The Netherlands	3%
Hong Kong	3%
Australia	3%
Italy	2%
Israel	2%
Japan	1%
Brazil	1%
Switzerland	1%
Japan	1%
India	1%
South Korea	1%
Russia	1%
Taiwan	1%
Finland	1%
France	1%
Other	4%

Fig. 1. Heat map and table showing country contributions on ACM SIGIR Conference between 2003 and 2017.

Table 2. Top 15 most used words in title and abstract and most used keywords in ACM SIGIR papers in the period 2003–2017.

	Words	Frequency	Keywords	Frequency
1	Search	4596	Information Retrieval	184
2	Retrieval	3150	Evaluation	153
3	Query	2941	Web Search	95
4	Based	2705	Learning to Rank	73
5	User	2640	Query Expansion	63
6	Results	2094	Personalization	58
7	Web	1904	Recommender Systems	53
8	Model	1889	Collaborative Filtering	52
9	Data	1754	Language Models	46
10	Users	1613	Question Answering	46
11	Document	1597	Diversity	41
12	Queries	1496	Machine Learning	41
13	Documents	1487	Ranking	41
14	Paper	1443	Twitter	38
15	Relevance	1292	Text Classification	36
...	...	...	...	...
105	...	...	Digital Libraries/Digital Library	11
...	...	...	...	...
802	Library/Libraries	71	...	...
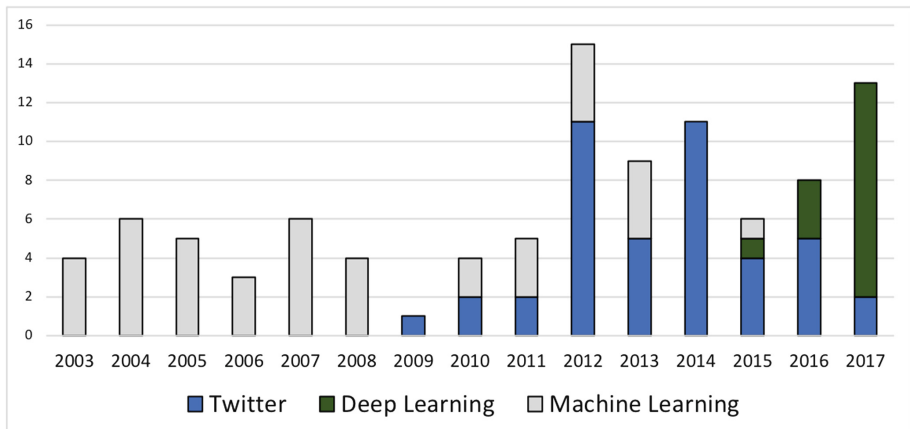
Fig. 2. Evolution of frequencies of usage of "Twitter", "Deep Learning" and "Machine Learning" keywords in the text of the ACM SIGIR papers of the period 2003–2017.
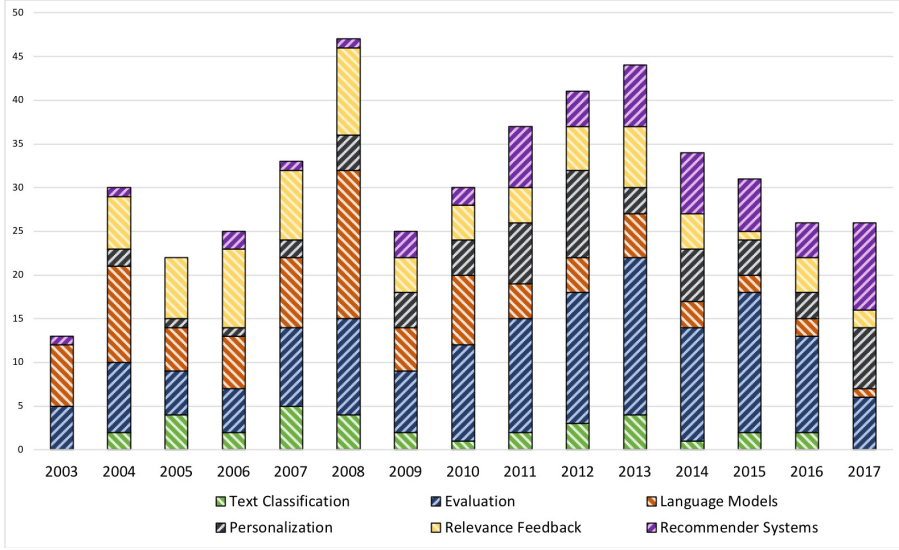
Fig. 3. Evolution of frequencies of usage of "Text Classification", "Evaluation", "Language Models", "Personalization", "Relevance Feedback" and "Recommender Systems", keywords in ACM SIGIR papers in the period 2003–2017.

Another relevant aspect is that in the initial years of the period we have analyzed, the main focus of IR research was on system-oriented topics, including data-centered and information processing aspects – keywords such as "Language Models", "Text Classification" and "Evaluation" were preponderant – whereas in later years the focus progressively shifted towards user behavior and user-system interaction – in this case the most common keywords are "Personalization", "Relevance Feedback", "Query Reformulation" and "Recommender Systems" (Fig. 3).

It is interesting to note that a crucial disruptive topic appeared in 2015: "Deep Learning" which strengthened the "Machine Learning" topic already present in the previous years, even though it was not as central as it is nowadays (Fig. 3).

The analysis of word frequencies underlines the presence of the same words appearing as the most used words in the paper title and abstract sections over all the study years; this was expected because a compiled stop-word list provided by the Terrier tool was used and not a customized stop-word list. However, this analysis is not without worth, because it supports the previous result on the transition from a data-centered to user-centered focus of the IR community that is underlined by the increasing frequency of the words "User" and "Users".

4.2   CCS Categories

It is worth noting that the results presented in Sect. 4.1 reported on the evolution of core topics by analyzing frequent words and keywords that are informally defined by authors without any specific standardization.

By contrast, good indicators of the topics that appeared and those that were left out over the last 15 years in the IR area can be derived from the analysis of the categories associated to ACM SIGIR papers. In fact all the categories associated to the papers of the ACM SIGIR collection were updated and revised in 2012 using the categories of the ACM CCS rev. 2012, which is a standard classification system. This is one of the reasons why we decided to rely on the ACM digital library, because it provides a standard updated classification system applied to all papers of all years and revised in 2012. Therefore, we analyzed the evolution of these categories, which increased the validity of our investigation, and found some particular and noticeable trends. Figures 4 and 5 show the most interesting CCS category frequencies on the ACM SIGIR Conference long papers between 2003 and 2017.

Some IR topics are consolidated and have been relatively stable in the past 15 years, such as "Document Representation", "Retrieval Model and Ranking", "Retrieval Task and Goal", "Evaluation of Retrieval Results". Other categories have been introduced in recent years such as "Query Log Analysis" and "Query Reformulation" which appeared only in the last two years, while "Search Personalization", "Search Interfaces", "Sentimental Analysis", "Retrieval on Mobile Devices", "Specialized Information Retrieval", "HCI Design and Evaluation Methods" started to play a preponderant role only in 2016. This means that IR is constantly expanding its application domain, and the expansion is driven by new and emerging technologies and changes in social lifestyle, growth both in more specific applications and in human-centered systems. Moreover, IR reacts to the effect of the dizzying growth rate of the new computer science areas: this emerges from the appearance of the use, only since 2016, of "Machine Learning Algorithms" and "Machine Learning Theory" categories.

4.3   Focused Analysis on Digital Library Topics

For the sake of our study, it is worth noting the presence of "Digital Libraries and Archives" both as a subcategory of "Applied Computing" and a subcategory of "Information Systems". Figure 4 reports that 10 long papers are classified under the "sub-tree" of the first subcategory and Fig. 5 reports that 11 long papers are classified on the other sub-tree of the classification system. Thus, we extended our analysis to all types of papers (long, short, demos, poster, ...) and we considered the presence of the "digital library" keyphrase in the abstract field. We found that the presence of "digital library" in the categories does not necessarily mean the presence of the term "digital library" in the abstract field and vice versa.

We found that a total number of 31 ACM SIGIR 2003–2017 papers are categorized with at least one "Digital Libraries and Archives" category (11 of those are long papers). Figure 6 shows the evolution of the number of papers (long papers, short papers, posters, workshops, demos) classified with "Digital Libraries and Archives" and shows a peak of 6 papers which present DL in the CCS Categories in 2007, about 3% of the total amount of ACM SIGIR papers published in that year, and a peak of 5 papers dealing with DL in the abstract field in 2013, that is about 2.5% of the total number of ACM SIGIR

papers published in that year. Further investigation revealed that there were 81 authors for these papers, most of whom came from the USA and the Netherlands (as shown in Table 3). Moreover, it is worth noting that all of these authors are academic researchers, with no affiliations other than universities.

CCS Name	Level	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	TOT
information systems	1	43	58	63	72	80	81	76	80	102	94	70	78	67	62	74	1100
information systems applications	2	9	12	9	8	10	16	6	8	13	10	6	7	4	4	13	135
digital libraries and archives	3	0	1	0	1	2	1	0	1	0	2	1	1	0	0	0	10
computational advertising	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
world wide web	2	1	2	2	4	1	2	4	6	3	4	2	0	3	10	16	60
web searching and information discovery	3	1	0	0	0	0	0	0	0	0	0	0	0	1	5	10	17
information retrieval	2	38	54	60	67	79	78	71	75	98	90	65	75	61	59	65	1035
information retrieval query processing	3	10	8	9	9	16	18	18	15	27	24	23	18	12	7	10	224
query log analysis	4	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1	4
query suggestion	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
query reformulation	4	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	5
users and interactive retrieval	3	0	0	0	0	0	0	0	0	0	0	0	0	1	10	15	26
personalization	4	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	4
task models	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
search interfaces	4	0	0	0	0	0	0	0	0	0	0	0	0	0	5	2	7
collaborative search	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
retrieval models and ranking	3	14	15	22	17	21	23	17	22	19	23	17	12	14	25	25	286
learning to rank	4	0	0	0	0	0	0	0	0	0	0	0	0	0	10	11	21
information retrieval diversity	4	0	0	0	0	0	0	0	0	0	0	0	0	0	5	2	7
retrieval tasks and goals	3	7	15	9	9	11	17	13	11	24	19	14	12	9	16	21	207
sentiment analysis	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2
search engine architectures and scalability	3	4	3	6	6	7	1	6	1	2	1	1	2	0	3	7	50
search engine indexing	4	2	3	4	5	5	1	3	0	2	1	1	1	0	2	1	31
retrieval on mobile devices	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
adversarial retrieval	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
specialized information retrieval	3	0	0	0	0	1	0	0	1	0	1	0	0	0	10	5	18

Fig. 4. CCS category frequencies on ACM SIGIR Conference long papers between 2003 and 2017 (part one).

CCS Name	Level	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	TOT
human-centered computing	1	1	4	6	4	3	3	3	5	6	7	3	3	3	8	7	66
human computer interaction (hci)	2	0	4	5	3	3	3	2	3	2	4	1	2	2	5	5	44
hci design and evaluation methods	3	0	0	0	0	0	0	1	0	0	2	1	1	1	4	4	14
theory of computation	1	1	1	3	1	4	2	0	1	5	0	1	5	1	5	5	35
theory and algorithms for application domains	2	0	0	0	1	0	0	0	0	1	0	0	1	0	3	2	8
machine learning theory	3	0	0	0	0	0	0	0	0	1	0	0	0	0	3	2	6
database theory	3	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	2
semantics and reasoning	2	0	1	2	0	1	2	0	1	4	0	1	2	1	0	0	15
program reasoning	3	0	1	2	0	1	2	0	1	4	0	1	2	1	0	0	15
computing methodologies	1	13	10	14	9	8	13	13	13	18	6	3	8	3	13	16	160
machine learning	2	6	4	6	6	3	7	7	7	10	1	1	1	1	9	13	82
machine learning approaches	3	1	3	4	4	0	3	3	4	2	1	0	1	1	3	8	38
machine learning algorithms	3	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	3
applied computing	1	3	6	10	7	6	5	4	5	6	4	3	3	3	3	3	71
law, social and behavioral sciences	2	0	0	0	0	0	0	1	2	3	1	1	1	1	2	1	13
computers in other domains	2	0	1	0	1	2	1	0	1	0	2	1	1	0	1	0	11
digital libraries and archives	3	0	1	0	1	2	1	0	1	0	2	1	1	0	1	0	11
document management and text processing	2	2	4	7	5	1	3	1	1	1	0	0	0	1	0	0	26

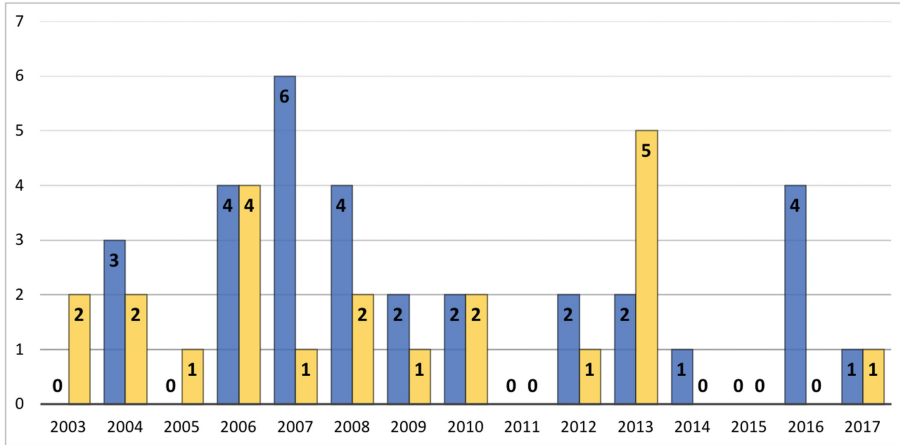Fig. 5. CCS category frequencies on ACM SIGIR Conference long papers between 2003 and 2017 (part two).

Fig. 6. Evolution of the number of papers classified with "Digital Libraries and Archives" CCS Category (left bars) and containing "digital library" keyphrase in the abstract field (right bars).

We therefore investigated what the topics of these 31 papers were and found that they mostly present, elaborate and evaluate new approaches and techniques to be applied in Digital Libraries, such as XML retrieval, natural language processing, text mining, recommendation methods, federated text retrieval, duplicate detection and methods for managing large collections.

5   Final Remarks

Discussion. In this work we analyzed the evolution of IR research topics in the last 15 years with a specific focus on DL-related aspects.

The first results we would like to emphasize are how the IR research community has been growing in recent years and how it has adapted to the most recent trends in computer science; this is clearly highlighted by the growing importance of topics such as "Deep Learning" and "Social Media" (expressed with different categories and keywords such as "Twitter" in 2012).

Another relevant aspect is the growing attention towards user-centered aspects of search, such as human-computer interaction, user studies, understanding and modeling how users search and what the most relevant tasks are nowadays. This aspect is strictly connected with the DL research field since the role of users in DL has always been preponderant, as highlighted by Rasmussen in [9]. The synergy between DL and IR foreseen in [9] can also be found in the most relevant topics addressed by ACM SIGIR papers, especially with regard to the use of relevance feedback and cross-lingual search. Digital libraries also played a relevant role within the ACM SIGIR community because they provided (and still provide) interesting search tasks and challenging collections to work with.

Table 3. Country contribution percentages on ACM SIGIR papers classified with "Digital Libraries and Archives" CCS Category.

Country	Contribution percentage
USA	23.4%
The Netherlands	20.2%
Germany	13.8%
United Kingdom	11.7%
China	8.5%
Australia	7.4%
Brazil	3.2%
Italy	3.2%
India	1.1%
Qatar	1.1%
Portugal	1.1%
Denmark	1.1%
Japan	1.1%
France	1.1%
Singapore	1.1%

Most of the ACM SIGIR papers categorized with "Digital Library" deal with: searching documents in semi-structured formats such as XML, identifying math formulas and retrieving formulas from textual documents, cooperative search done by expert users, finding translations or searching scanned book collections, federated document retrieval and using user feedback to improve search results. All these aspects are clearly related to DL since they provide wide and challenging document collections that need to be accessed and efficiently searched; in past years XML was a central format for document exchange and encoding, now RDF is growing and the use of semantics and relations between entities and documents are gaining traction. DL on the one hand provides the data and user requirements that can be addressed by IR research, while on the other hand it employs the results that the IR community produces. The relation between IR and DL constitutes a positive and factual technology transfer channel between the two disciplines.

Other aspects that relate DL to IR are user profiling and the definition of specialized search services for different user types such as general users, expert users, domain experts and so on. This is an aspect that has always been central in the DL context and that is slowly gaining traction also in IR. IR and recommendation systems are increasingly intertwined [14] as they become increasingly integrated into DL services [4,13].

Conclusion. It is important to highlight the role of public and private investment in DL research; indeed, in the first decade of the century both the European Commission and the National Science Foundation in the USA invested conspicuously in DL-related research projects. This resulted in a growth in the interest of researchers in the DL area with positive spill-overs into other fields as shown by a growing number of DL-related papers published at ACM SIGIR in 2006–2007. In recent years, research investments have shifted to other topics, thus reducing the interest in DL topics which translated into a smaller number of researchers dedicated to this field; this is also evident when considering ACM SIGIR publications in the last three years, in fact there were almost no DL-related paper presented at the conference.

By conducting this study we realized that DL is a field which still has great potential because it has unique collections of textual and non-textual documents and a wide and heterogeneous user base with all kinds of access and search tasks. Nevertheless, in order to keep up the high level of research activity in this field it is necessary to maintain a highly multidisciplinary profile that can continue to attract researchers from other fields and that can "feed" other related research fields with challenging data and tasks to be addressed. This synergy, if constantly sustained and nurtured, has a true potential for improving DL services as well as search and access methods in general.

Future Directions. For the presented analysis we relied on a text mining approach; on the other hand, a bibliometrical approach would allow us to better investigate the interrelationships among papers and authors of the IR and DL fields, and how these two fields influence each other. This would also allow us to identify which authors are peculiar to the DL and who instead belong to both the IR and DL communities. Thus, we plan to use a bibliometric approach to undertake further analyses in the future, and we plan to extend the analysis considering in addition the impact of the public and private funding initiatives in the DL community. To this end, we could leverage on DBLP-NSF [12], that connects computer science publications extracted from DBLP to their NSF funding grants, by extending it also to European and National funding agencies.

References

1. GROBID (2008–2018). https://github.com/kermitt2/grobid. Accessed 16 Aug 2018
2. Agosti, M., Masotti, M.: Design of an OPAC database to permit different subject searching accesses in a multi-disciplines universities library catalogue database. In: Belkin, N.J., Ingwersen, P., Pejtersen, A.M. (eds.) Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, 21–24 June 1992, pp. 245–255. ACM (1992). https://doi.org/10.1145/133160.133207
3. Appleton, G.: Future trends in digital libraries and scientific communications. Procedia Comput. Sci. **38**, 18–21 (2014). https://doi.org/10.1016/j.procs.2014.10.004
4. Beel, J., Aizawa, A., Breitinger, C., Gipp, B.: Mr. DLib: Recommendations-as-a-Service (RaaS) for Academia. In: 2017 ACM/IEEE Joint Conference on Digital Libraries, JCDL 2017, pp. 313–314 (2017)
5. Hiemstra, D., Hauff, C., de Jong, F., Kraaij, W.: SIGIR's 30th anniversary: an analysis of trends in IR research and the topology of its community. SIGIR Forum **41**(2), 18–24 (2007). https://doi.org/10.1145/1328964.1328966
6. Hildreth, C.: The Online Catalogue: Developments and Directions. Library Association, London (1989)
7. Lopez, P.: GROBID: combining automatic bibliographic data recognition and term extraction for scholarship publications. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 473–474. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04346-8_62
8. Macdonald, C., McCreadie, R., Santos, R.L., Ounis, I.: From puppy to maturity: experiences in developing Terrier. In: Proceedings of OSIR at SIGIR, pp. 60–63 (2012)
9. Rasmussen, E.: Information retrieval challenges for digital libraries. In: Chen, Z., Chen, H., Miao, Q., Fu, Y., Fox, E., Lim, E. (eds.) ICADL 2004. LNCS, vol. 3334, pp. 95–103. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30544-6_10
10. Smeaton, A.F., Keogh, G., Gurrin, C., McDonald, K., Sødring, T.: Analysis of papers from twenty-five years of SIGIR conferences: what have we been doing for the last quarter of a century? SIGIR Forum **36**(2), 39–43 (2002). https://doi.org/10.1145/945546.945550
11. Walker, S.: Improving subject access painlessly: recent work on the Okapi online catalogue projects. Program **22**, 21–31 (1988)
12. Wu, Y., Alawini, A., Davidson, S.B., Silvello, G.: Data citation: giving credit where credit is due. In: Das, G., Jermaine, C.M., Bernstein, P.A. (eds.) Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, pp. 99–114. ACM Press, New York (2018). https://doi.org/10.1145/3183713.3196910
13. Yi, K., Chen, T., Cong, G.: Library personalized recommendation service method based on improved association rules. Libr. Hi Tech **36**(3), 443–457 (2018)
14. Zamani, H., Croft, W.B.: Joint modeling and optimization of search and recommendation. In: Alonso, O., Silvello, G. (eds.) Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems (DESIRES 2018), CEUR Workshop Proceedings, Bertinoro, Italy, 28–31 August 2018, vol. 2167, pp. 36–41. CEUR-WS.org (2018). http://ceur-ws.org/Vol-2167