Knowing the past to Plan for the Future
An In-depth Analysis of the First 10 Editions of the WEBIST Conference

Giseli Rabello Lopes[1], Bernardo Pereira Nunes[2], Luiz André P. Paes Leme[3], Terhi Nurmikko-Fuller[4] and Marco A. Casanova[2]

[1]*Computer Science Department, Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil*

[2]*Department of Informatics, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil*

[3]*Computer Science Institute, Fluminense Federal University, Niterói, RJ, Brazil*

[4]*Oxford e-Research Centre, Oxford University, Oxford/OX1 3QG, U.K.*

Keywords: Conference Analysis, Statistical Analysis, Bibliometrics, Social Network Analysis, WEBIST Analysis, Linked Data.

Abstract: Over the last ten years, members of the WEBIST community have dedicated their time and efforts to face a number of research challenges, making significant advances in Information Systems and pointing to new directions for innovation and learning. After ten successful WEBIST conferences and several scientific publications, an extensive analysis of the WEBIST conferences has been carried out (involving authors, publications, conference impact, topics coverage, community analysis and other aspects) to possibly assist us to further advance Information Systems. Thus, in this paper, we present an in-depth analysis of the last ten WEBIST conferences based on social network analysis, bibliometrics and statistical measures and describe a Web-based application built on top of triplified datasets to interactively explore the findings and possibly assist the Information Systems community to reveal new directions.

1 INTRODUCTION

Data analysis has shown its utility for conferences, detecting related research groups, topics of interest, impact of authors and publications in a given field, among others. An example of this is the analysis of a group of four conferences in the field of Human-Computer Interaction (HCI), conducted by Henry et al. (2007). It was based solely on metadata information of publications (such as authors and keywords) and was capable of providing valuable insights into authors' behaviours and research topics investigated in HCI in the last two decades. Blanchard (2012) presented a ten-year longitudinal study over Intelligent Tutoring Systems (ITS) and Artificial Intelligence in Education (AIED) fields. He focused on the analysis of potential cultural biases of the American Psychology Association (APA) in the ITS and AIED fields. Chen et al. (2009) presented a visual analytic approach to the study of scientific discoveries and knowledge diffusion. Their analysis focused on the identification of co-citations clusters where they were classified and used to understand how astronomical research evolved between 1994 and 1998. Another example following the same line was conducted by Gasparini et al. (2013). In their study, they were able to identify central authors and institutions in the HCI field, as well as important trends and topics. As for the Information System (IS), Posada and Baranauskas (2014) analysed a sister event called International Conference on Enterprise Information Systems (ICEIS). They built a roadmap of the IS field based on paper titles and authors from the last three years in ICEIS, and for the last eight years of selected papers published in a Springer series on IS. Chen et al. (2007) performed citation analysis of all papers published in the International Conference on Conceptual Modeling (ER) between 1979 and 2005. The analysis conducted by the aforementioned communities opened up a wide range of opportunities for research agendas and trends as well as supporting the domains introspective analysis.

This paper aims at extending previous analytical methods and providing a comprehensive social analysis of the community of WEBIST. Recently, Zervas et al. (2014) presented a study on research collaboration patterns via co-authorship analysis regarding Technology-enhanced Learning fields. Sim-

Rabello Lopes G., Pereira Nunes B., P. Paes Leme L., Nurmikko-Fuller T. and A. Casanova M..
Knowing the past to Plan for the Future - An In-depth Analysis of the First 10 Editions of the WEBIST Conference.
DOI: 10.5220/0005447704310442
In *Proceedings of the 11th International Conference on Web Information Systems and Technologies* (WEBIST-2015), pages 431-442
ISBN: 978-989-758-106-9

ilar analyses were conducted by Procopio Jr. et al. (2011) regarding Databases fields and by Cheong and Corbitt (2009) regarding Information Systems fields (analysing the Pacific Asia Conference on Information Systems). The analysis of co-authorships in research communities can reveal strong research groups in the area and also enable the creation of links overtime between different groups. Apart from presenting the analysis of the last ten editions of WEBIST, we are also concerned with the publication of the results in a format where they can be replicated and reused in further analysis. For this, we have borrowed Batista and Loscio's approach (Batista and Loscio, 2013) where they used Linked Data (LD) principles to publish conference data.

In this paper we present an in-depth and thorough analysis of the first ten editions (2005-2014) of the WEBIST conference. To briefly summarise, WEBIST has brought together researchers and professionals to develop and advance the IS field. So far, it has already attracted 2,867 researchers and professionals from several institutions as well as published 1,449 papers, which in turn are being cited by other researchers in IS and other fields. Moreover, the conference currently has five main tracks that cover a wide range of aspects involving IS: *Internet Technology*, *Web Interfaces and Applications*, *Society, e-Business and e-Government*, *Web Intelligence* and *Mobile Information Systems*.

The analysis presented in this paper relies on techniques borrowed from social network analysis (Wasserman and Faust, 1994), bibliometrics and traditional statistical measures. We also created a Web-based application that enables users to interactively explore WEBIST data. As the WEBIST data is published following LD principles (Berners-Lee, 2006), we also provide a SPARQL endpoint where other researchers can extend our analysis. The importance of such analysis goes beyond the analysis and possible interpretation of data and represents a milestone achieved by the IS and WEBIST community so far.

The remainder of this paper is organised as follows. Section 2 overviews metrics and measures used in the analysis of the last ten WEBIST conferences. Section 3 details the extraction, enrichment and publication process of raw WEBIST data into RDF data and presents a visualisation tool specifically created to manipulate and possibly assist users in finding new research groups, topics and insights. Section 4 presents several analysis conducted with the WEBIST tool. Finally, Section 5 concludes the work with remarks and future directions.

2 BACKGROUND

This section provides the necessary background information required to understand the analysis conducted with the data produced over the last decade by the WEBIST community. We review metrics and methods of statistical analysis, social network analysis and bibliometric indices.

2.1 Classical Statistical Measures

Pearson's Correlation Coefficient (Rodgers and Nicewander, 1988), often denoted by the letter r, measures the strength and direction of the linear correlation between two variables X and Y. Pearson's coefficient (see Equation 1) can be defined as the covariance of the variables divided by the product of their standard deviations to measure their dependence:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{Y})^2}} \quad (1)$$

An r value between +1 and -1 indicates the degree of linear dependence between X and Y, r=1 indicates a total positive correlation between the two variables and, finally, r=-1 indicates a total negative (inverse) correlation. For instance, as X values increase, Y values linearly decrease.

Lorenz Curve (Gini, 1912) represents the cumulative distribution of a probability density function. Such a function is built as a ranking of the members of the population disposed in ascending order of the amount being studied. The percentage of individuals is plotted on the x-axis and the percentage of the variable values on the y-axis. The distribution is perfectly equalitarian when every individual has the same variable value; a 45-degree line represents the perfect equality. On the other hand, the perfectly unequal distribution is the one in which only one individual has all the variable value, the curve is y=0 for all x<100%, and y=100% when x=100%, known as the perfect inequality line. This curve was initially created to study the social inequality of wealth and income distributions for a population, but can be applied to analyse other distributions (Lopes et al., 2012). We used the Lorenz curve (Section 4) to study the distribution of papers by author.

Gini Coefficient (Gini, 1912) is a measure of statistical dispersion indicating the inequality among values of a frequency distribution. It is graphically represented as the area between the perfect equality line and the observed Lorenz curve.

Robin Hood Index (Hoover, 1941), also called Hoover index, is used to measure the fraction of the total variable value that must be redistributed over the population to become a uniform distribution. It is graphically represented as the longest vertical distance between the Lorenz curve and the perfect equality line.

2.2 Social Network Analysis

Before introducing social network metrics and concepts (Wasserman and Faust, 1994; Freeman, 1979; Hoser et al., 2006; Marsden, 2002; Newman, 2001, 2003), it is convenient to represent a social network as a graph structure $G = (N,E)$, where N is the set of nodes, where $n_i \in N$ represents an actor of the network, and E is the set of edges, where $e_i \in E$ represents a relational tie between a pair of actors.

Density is calculated as the number of the actual existing edges of a graph, divided by the maximum number of edges the graph can have. A density value equal to 1 indicates an entirely connected network while 0 indicates a disconnected network. Considering an undirected graph where the possible number of connections between each two nodes is 1, the density can be calculated as:

$$D = \frac{2|E|}{|N|\,(|N|-1)} \tag{2}$$

where $|E|$ is the cardinality of the set of edges and $|N|$ is the cardinality of the set of nodes.

Modularity is a measure of the structure of networks and estimates the strength of division of a network into communities (groups). It is often used in optimisation methods for detecting community structure in networks. A high modularity value indicates a network having dense connections between the nodes within the communities, but sparse connections between nodes in different communities. Modularity can be calculated as (Newman and Girvan, 2004):

$$Q = \sum_i (e_{ii} - a_i^2) \tag{3}$$

where e_{ij} is the portion of edges connecting nodes from the community i to nodes from the community j; $a_i = \sum_j e_{ij}$ is the portion of edges with at least one node from the community i. Each edge contributes once in the count (the contribution must be divided by half, each halve for e_{ij} and the other for e_{ji}).

Giant Component (also named main component) is the connected component which contains most of the nodes in the graph.

Giant Coefficient is based on the size of the giant component G' of a graph G. It is calculated as the number of nodes N' in the giant component divided by the total number of nodes N in the entire graph:

$$GC = \frac{|N'|}{|N|}, \text{ where } N' \subseteq N \tag{4}$$

Diameter is associated with graph distance. It is calculated as the maximum value among all shortest paths between two nodes of the graph (i.e., the longest distance between any pair of nodes belonging to the graph).

Average Clustering Coefficient is a measure of the degree to which nodes in a graph tend to cluster together (connectivity of neighbours). It is calculated as the average of the clustering coefficients of all the nodes in the graph:

$$\bar{C} = \frac{1}{|N|}\sum_{i=1}^{|N|} C_i \tag{5}$$

where C_i is the clustering coefficient of a node n_i and is calculated as the number of existing edges between the direct neighbours of n_i divided by the total number of possible edges directly connecting all neighbours of n_i.

2.3 Bibliometric Indices

This section introduces two common bibliometric indices often used to measure the impact, in terms of popularity, of researchers, scientific publications, conferences and journals.

***h*-index** was proposed to measure both the number of publications and the number of citations per publication of a scientist. According to Hirsch (2005), a scientist has index h if h of his/her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have no more than h citations each. This index is also applied to estimate the productivity and impact of conferences.

i10-index indicates the number of publications of a scientist having at least ten citations[1].

3 WEBIST WORKFLOW - FROM RAW TO RDF DATA

3.1 Overview of the Process

This section overviews the process of data acquisition involving extraction, enrichment, preparation and consolidation to create the *WEBIST Dataset* and to

[1] http://googlescholar.blogspot.com.br/2011/

use it by the *WEBIST Analytics*. Figure 1 depicts the whole process.

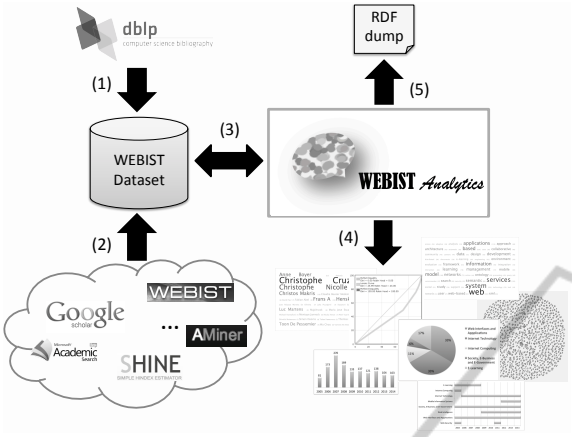


Figure 1: WEBIST workflow.

Initially, we created an interlinked open dataset, named *WEBIST Dataset*, available in RDF, following the Linked Data principles (Berners-Lee, 2006), about the 10 editions of WEBIST conference. This dataset was created by aggregating data extracted from different data sources. The initial core of the data about WEBIST was extracted from DBLP (Digital Bibliography & Library Project)[2] (Step 1). Then, the data was enriched using data crawled from different Web sources such as Google Scholar Citations[3] (Step 2).

Based on the information loaded in the *WEBIST Dataset* (Step 3), the proposed Web application provides different functionalities as exploratory search and several analysis over the data presented through different graphical visualisations (Step 4).

Moreover, through the *WEBIST Analytics* interface, the RDF dump of the *WEBIST Dataset* is available for download (Step 5). *WEBIST Dataset* creation and *WEBIST Analytics* functionalities are detailed in the next subsections.

3.2 WEBIST Dataset

Data Acquisition. Over the last ten years a huge amount of data has been generated on the Web in different formats. This also happened with WEBIST conferences, where information about the conference, such as paper acceptance or organisation committee has been published. Thus, to create a tool to seamlessly make sense of the data, we aggregated data extracted from different data sources, being aware of the possible necessity of initially preparing the data for deduplication (Elmagarmid et al., 2007) techniques.

The initial core of the data about WEBIST was extracted, in December 2014, from DBLP a digital library about computer science publications. We were not able to find an updated source of DBLP data in RDF format (containing all editions of WEBIST conference). Thus, we had to extract the data directly from the XML version of DBLP available. This XML data also contained information about the name disambiguation of the authors (different spellings of the name representing the same author in XML version of DBLP). Thus, the authors name disambiguation (Borges et al., 2011) was facilitated in this initial core. In summary, we collected information about the published papers and authors of WEBIST, reaching a total of 1,449 papers and 2,867 authors.

Data Enrichment. Data enrichment serves as a means to extending the initial data from additional data sources. For this, we developed a focused crawler to obtain this complementary information. In this step, information from Google Scholar Citations[4] and Google Scholar were used to obtain bibliometric indices of WEBIST authors. Specifically, the key of authors in Google Scholar Citations and the authors indices (*h*-index, i10-index and number of citations) were extracted from Google Scholar[5] and Google Scholar Citations, respectively. The crawling process used the name of the authors to perform the searches. Using this strategy, 748 authors profiles were found in Google Scholar Citations, representing 26.09% of the total WEBIST authors. Other complementary information about some publications citations was crawled from Google Scholar. We collected the number of citations for the assumed most cited papers: the candidates to be most cited papers were obtained by the topmost ranked WEBIST papers presented in SHINE (Simple H-INdex Estimator)[6], Arnetminer[7] and Microsoft Academic Search[8]. Additional information about the main research areas of each edition of WEBIST were extracted from each conference Web site[9].

Data Transformation. Another crucial step is data transformation, carried out after data acquisition involving the preparation and enrichment steps, requiring a common format for the data. For this, we followed the Linked Data principles (Berners-Lee, 2006)

[2]http://www.informatik.uni-trier.de/~ley/db/

[3]http://scholar.google.com/citations

[4]http://scholar.google.com/citations

[5]http://scholar.google.com

[6]http://shine.icomp.ufam.edu.br

[7]http://arnetminer.org/

[8]http://academic.research.microsoft.com/

[9]*2005-2011:* http://www.webist.org/WEBIST$*year*$; *2012-2014:* http://www.webist.org/?y=$*year*$

that encourage data publishers to expose their data through HTTP mechanism and to use RDF as the data description language. According to this guide lines, the publishers should name things using HTTP URIs and provide appropriate clipping of data in RDF when users follow the URIs. We used a relational-to-RDF framework (D2RQ) (Bizer and Seaborne, 2004) that dinamically transforms relational data into RDF graphs. It provides an HTML browser for relational databases as well as a SPARQL interface to query the database. This framework also provides a mapping language to define rules for transforming relational data and schema into RDF graphs.

Data Publication. The successful completion of these previous steps ensured that the dataset was available to others (both in terms of users and/or applications) that want to use it for a myriad of different purposes. The RDF dump of the *WEBIST dataset* is available for download from the *WEBIST Analytics* interface.

3.3 WEBIST Analytics Application

WEBIST Analytics, a Web-based application, was created to provide multiple perspectives of the data produced by WEBIST conferences over the 10 editions. The proposed application is composed of analytics tools, graphical visualisations and a simple search engine that assists users in finding, uncovering and making sense of the information available. *WEBIST Analytics* application can be accessed at:
http://lab.ccead.puc-rio.br/webist_analytics/.

Based on the information loaded in the *WEBIST Dataset*, the proposed Web application provides different functionalities as both exploratory search and several analyses over the data, presented through different graphical visualisations. Free text search is available over two different WEBIST graphs, the co-authorships graph (among authors) and a more complete graph composed by co-authorships and authoring relations (among authors and publications). It allows users to search and retrieve related information about WEBIST conferences, including an interactive visualisation of networks. Other exploratory search is allowed via tag cloud visualisations. In this case, the terms in the tag cloud can be selected and the associated publications retrieved, this in turn assisting users in finding papers related to each research topic. Details about the different analyses available and their results discussions are presented in Section 4.

4 ANALYSIS AND RESULTS

This section presents and discusses the results of the analysis available in *WEBIST Analytics*. Note that the outcomes obtained in this section were computed using the methods and metrics presented in Section 2.

4.1 General Analysis

An initial analysis of all WEBIST conferences was conducted with regard to its authors and publications. In this analysis we gathered 1,449 publications, which included all full papers, short papers, posters and selected papers. Figure 2 depicts the distribution of the papers over the conference editions. The number of accepted papers reaches its peak in 2007, where 270 papers were accepted to a single conference, a figure almost twice the average number of papers accepted to other editions. This peak number of publications may be an indication of the rapid increase in the popularity of WEBIST, and its reaching a certain level of maturity over the years, settling on a stable conference-size and community.

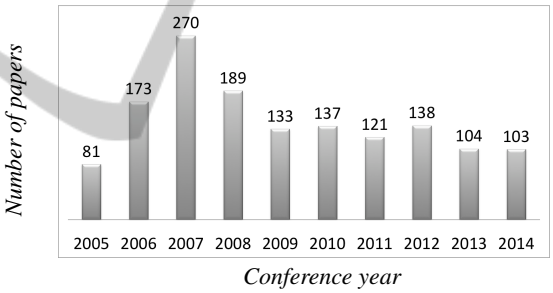


Figure 2: Number of papers published per year.

A rough analysis of the community can be carried out based on the number of authors of a scientific publication. The number of authors of a paper gives us a hint of the average size of the community and research groups. Across the 10 editions of WEBIST, there have been contributions from 2,867 authors, which gives an average of 2.91 authors per publication (with a standard deviation (σ) of 1.35, the maximum number of authors being 14 per paper and the minimum 1). Figure 3 shows the distribution of the average number of authors per year.

The list of topmost authors of WEBIST may reveal not only prolific authors, but possible experts and supporters for future editions of the conference. The engagement of researchers in a specific community could be initially measured with the number of papers they have had accepted in the earlier editions of the conference. The assumption is that if they had over a specific number of papers, they might be eli-

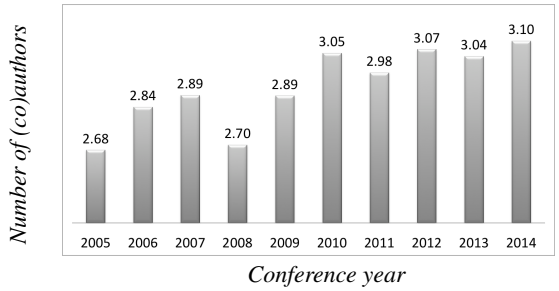


Figure 3: Average number of (co)authors per paper over the conference years.

gible to make part of the program committee. After 10 editions, a total of 29 authors had more than six papers. The most active researcher had 15 published papers and the second had 12 papers. Figure 4 shows the top authors as a tag cloud[10]. The size of the names represents how active a research is in the WEBIST conference.

Figure 4: Top authors with more than 6 papers.

Figure 5 presents the Lorenz curve[11] along with an analysis based on the Gini coefficient and the Robin Hood Index (see Section 2). The Gini coefficient resulted in 25.99% of inequality, while the Robin Hood Index was 23.06%. The results show that the Lorenz Curve is closer to the equality than to the inequality line. This is an expected result for peer-reviewed conferences, where only high quality papers are accepted for publication. Although a few authors have 6 or more papers in WEBIST editions, the Lorenz Curve and the Robin Hood Index show that no redistribution is necessary, i.e., there is no bias in accepting papers from a research group or another, but simply merit. A high Robin Hood Index would indicate a possible need for further analysis in some publications.

[10]http://tagcrowd.com

[11]http://www.peterrosenmai.com/lorenz-curve-graphing-tool-and-gini-coefficient-calculator

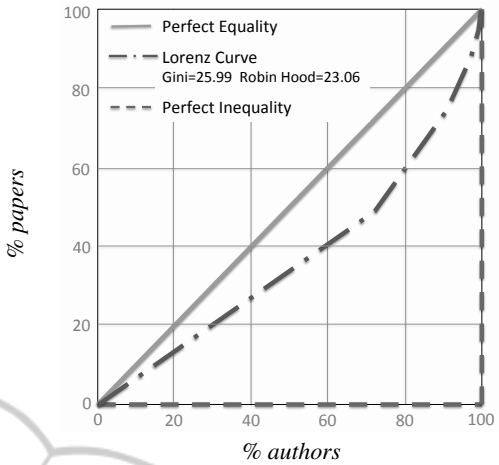


Figure 5: Lorenz curve for the number of papers per author distribution.

4.2 Co-authorships Network

Social Network Analysis (SNA) techniques were applied to obtain information about the co-authorships in the WEBIST conference (see Figure 6). The analysis was conducted over an undirected graph G (defined in Section 2), where the nodes represent the authors and the edges represent a co-authorship between researchers. A fraction of the co-authorship network is shown in Figure 6, where the size of the nodes denotes the co-authorship connectivity. The WEBIST co-authorships network is comprised of 2,867 authors and 4,235 pairs of authors (edges) having at least one co-authored paper.

Table 1 shows an analysis of the co-authorship network using SNA measures. The analysis considers all WEBIST authors in the last 10 years.

- **Average Degree** shows that the authors, on the average, have co-authored papers with 2.9 other authors.
- **Density** shows a low proportion of co-authorships in the network relative to the total number possible (situation where all authors co-authored at least one paper with all others), only 0.1%. It represents a weakly connected network. This shows an expected result in a conference network, where there are different groups of authors working in different papers. The measured modularity and the number of communities, as explained below, can reinforced this result.
- **Modularity** shows a high value representing the strength of division of the network into modules (also called groups, clusters or communities). Thus, WEBIST co-authorships network has co-authorships between the authors within the communities but none between authors in different

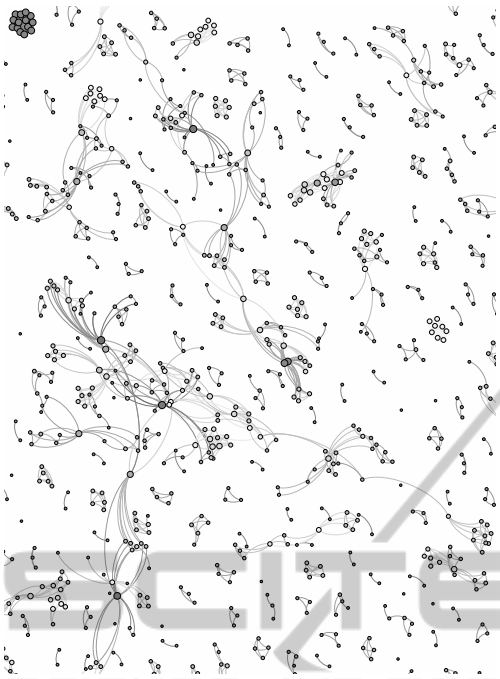


Figure 6: A fraction of the co-authorships network.

communities.

- **Number of Communities** detected based on the modularity, was 803, being exactly the same as the **Number of Connected Components**. This shows that, in the analysed network, there are isolated communities that have not co-authorships in WEBIST with the authors of the other communities.

The following analysis takes into account only the giant component of the WEBIST network:

- **Giant Coefficient** represents the percentage of authors in the Giant Component of the WEBIST co-authorships network, being approximately 1.57% (45 authors) of the total number of authors that published in all WEBIST conferences. These authors have 108 co-authorships between them (2.55% of the total possible co-authorships, i.e., if each of these authors co-authored on at least one paper with all others).
- **Diameter** represents the longest of all the shortest paths between two authors in the Giant Component, being estimated as 8. This shows that the farthest authors in the Giant Component have more than six degrees of separation, based on co-authorship in WEBIST papers. This reveals that the Giant Component probably results from a hierarchical structure, which is natural when research groups of different institutions are involved. The different research groups (subgroups) are connected by "hub" authors (probably research group leaders and/or professors) that collaborate in different research projects amongst the subgroups, while some researches (probably students) developed more specific tasks (sometimes related to only one paper).
- **Clustering Coefficient** measures the average degree to which authors in the network tend to cluster together, being approximately 93.4%. This shows that many authors belonging the Giant Component worked with other authors that also worked together in at least one paper.

Table 1: Social Networks Analysis from the WEBIST co-authorships network.

Measure	Value
Average Degree	2.954
Density	0.001
Modularity	0.995
Number of Communities	803
Number of Connected Components	803
Giant Coefficient*	0.0157
Diameter*	8
Average Clustering Coefficient*	0.934

* Estimated considering the Giant Component.

4.3 Authors Indices

In this section, we considered different bibliometric indices to analyse the profiles of WEBIST authors. As stated previously (Section 3), we identified and extracted Google Scholar Citations profiles for 26.09% of the WEBIST authors. Thus, the analysis presented in this section is related only to this portion of authors.

The bibliometric indices from WEBIST authors were firstly analysed in terms of the Average and the Standard Deviation (σ) (see results in Table 2). The bibliometric indices, obtained from Google Scholar Citations data, were separated into global indices, estimated considering all the years of the citations, and the same indices estimated, considering only the citations since 2009. On the average, the authors presented a considerable total number of citations and i10-index values greater than their *h*-index. However, the Standard Deviation was quite high, showing that the community, as expected in good conferences, is formed of both young and senior researchers.

To better understand the profile of the WEBIST authors, we performed further analyses by splitting the authors into two groups, named *A* and *B*. We assigned to Group *A* those authors who had an *overall h*-index greater than the *h*-index *since 2009* and assigned to Group *B* those authors who had a *overall h*-index equal to the *h*-index *since 2009*. This classification assumes that the authors whose *overall h*-index

Table 2: Average and standard deviation of number of citations and bibliometric indices from authors.

Measure	Average	σ
overall citations	1,634.49	4,087.46
citations *since 2009*	988.95	2,565.17
overall h-index	14.30	12.17
h-index *since 2009*	11.54	8.98
overall i10-index	28.16	54.32
i10-index *since 2009*	19.94	42.03

consisted solely of citations made after 2009 were researchers who had started their careers more recently than those whose *overall h*-index included citations from before 2009.

Table 3 presents the results using this classification. This table shows, for each conference year, the percentage of authors and the respective average of the *h*-index per class. The results evidence that, in all conference editions, the number of authors in Group *A* is greater than those in Group *B*. Also, the results show that, in all conference editions, the average *h*-index of authors in Group *A* is greater. Note that the average of *h*-index is 18.35 for authors in Group *A* considering all editions of WEBIST conference.

Table 3: Percentage and Average of *h*-index of scholars in groups *A* and *B*.

year	Percentage		Average of *h*-index	
	group *A*	group *B*	group *A*	group *B*
2005	84.62%	15.38%	19.77	9.50
2006	78.65%	21.35%	18.03	8.84
2007	80.59%	19.41%	18.58	7.24
2008	63.73%	36.27%	18.38	6.95
2009	67.86%	32.14%	20.39	8.15
2010	67.03%	32.97%	18.64	7.00
2011	67.06%	32.94%	20.16	5.54
2012	57.02%	42.98%	15.65	5.39
2013	53.03%	46.97%	20.74	6.52
2014	55.56%	44.44%	19.72	4.90
All	65.64%	34.36%	18.35	6.58

4.4 Topics and Conference Areas

In this section, we analyse the topics of the papers published over the 10 years of WEBIST conference and their relation to the predefined main conference areas. Firstly, Figure 7 presents, in alphabetical order, the main conference areas over the different conference editions. Some areas appear in all conference editions, such as *Society, E-Business and E-Government* and *Web Interfaces and Applications*. The third most frequent area is *Internet Technology*, which appeared from the second edition to the last one, probably as an expansion of *Internet Computing* (which appears only in the first conference edition). *Web Intelligence* and *Mobile Information Systems* appear more recently, in 2009 and 2012, respectively. *E-Learning* appears only in the first four editions of WEBIST conference. This phenomenon can be explained by the fact that the WEBIST conference, from 2009 to 2014, was held in conjunction with CSEDU (The International Conference on Computer Supported Education), a conference focused in innovative technology-based learning strategies and institutional policies on computer supported education (e-learning). *Web Security* appears only in specific editions (2005 and 2011).

Another analysis was performed over the topics covered by the papers published in WEBIST conferences. Figure 8 shows a tag cloud generated from the terms presented in the titles of the papers. This tag cloud represents the terms followed by their total frequencies in parentheses. Moreover, the term size in the graphic is proportional to their frequency. Terms as *web*, *systems*, *services*, *applications*, *model* and *information* are the most frequent. These terms are aligned with the research focuses of WEBIST conference that are technological advances and business applications of web-based information systems.

Figure 8: Top 50 terms of years 2005-2014.

For a more detailed analysis, we considered the evolution of main conference areas and terms presented in titles of WEBIST papers per conference year. Specifically, we verified what happened to the frequency of particular terms that are directly related to updates in the main conference areas.

- *e-Learning* area was eliminated in 2009. *E-learning* term was a frequent top term in titles between 2005 and 2008, but this was not true in the following years (2009-2014).
- *Web Intelligence* area was included in 2009. Terms related to topics such as information filtering and retrieval, Web mining and classification appeared in different conference years (including years prior to 2009).
- *Web Security* area appears in editions from 2005

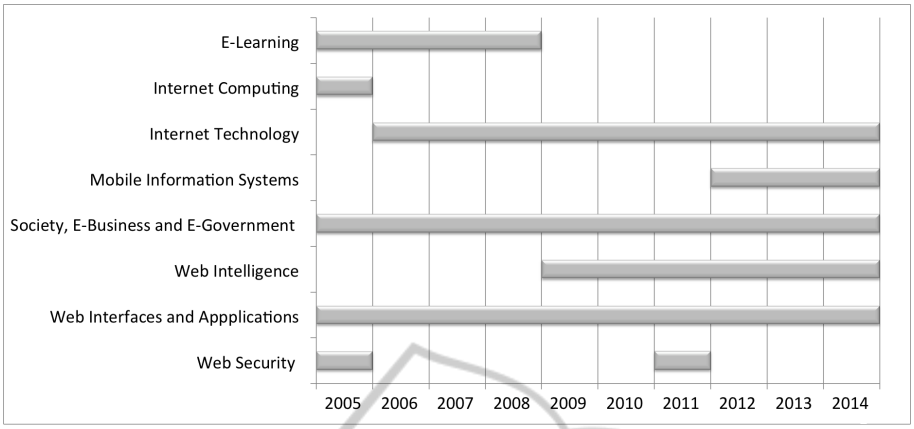


Figure 7: Main conference areas per conference year.

to 2011. The *security* term appears in the tag cloud of 2005 but not in 2011. We decided to investigate the quantity of papers published in 2011 that were directly associated with this main research area and discovered that only two short papers and one poster were published. This was probably the underlying reason which led to the deletion of this main research area in the following year.

- *Mobile Information Systems* area was included in 2012. The *mobile* term appears among the top 50 terms in 2012 (previously the term already appeared in the first conference editions, but became prominent only after the inclusion of the *Mobile Information Systems* area in 2012).

We also studied the evolution of the top 50 terms in the titles over a decade of WEBIST conferences. Table 4 presents the Average and the Standard deviation (σ) of the frequency of the top 50 terms. In the first editions of the conference, for the exception of 2005, both average and σ were high, leading us to conclude that there are likely to be terms that are related to major topics, as well as marginal topics in the accepted papers. In the most recent of conference editions, the terms have a more equal distribution (greater equality frequency), showing that even whilst manifesting some peripheral change over the years, the conference found a core that is equally evolving. Average and σ, analyzed in conjunction, demonstrate that the frequency of the top 50 terms (and consequently the relative frequency of the conference topics) is becoming more homogeneous. Moreover, a high diversity (dispersion) was observed, i.e., there were many terms (topics) covered by the conference over its 10 years.

Pearson's correlation coefficient was estimated between the frequency of top 50 terms group from each conference edition (see results in Table 5). The sequence of the conference editions (underlined values in Table 5), except between 2006-2007, maintained a consistency within the group of top 50 terms: terms from one year correlated with the group of terms from the following year (Pearson's correlation coefficient is positive). Moreover, the correlation between the groups of top 50 terms from years 2008-2009 increased considerably compared with all the previous years (2005-2006; 2006-2007 and 2007-2008). This probably happened because, in this period, the main research areas were updated, with the removal of *E-learning* and the inclusion of *Web Intelligence*.

Table 4: Average and standard deviation from frequency of top 50 terms per conference edition.

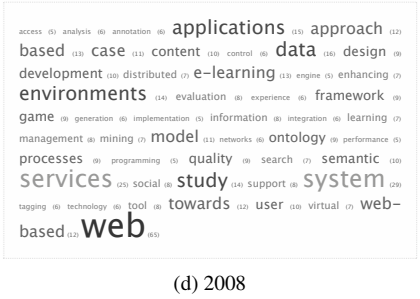
Year	Average	σ
2005	4.58	3.59
2006	9.08	6.90
2007	14.74	11.68
2008	10.50	9.12
2009	7.64	6.14
2010	7.18	5.37
2011	6.72	5.32
2012	7.12	4.53
2013	5.26	3.14
2014	5.08	3.02
All	70.30	55.66

Finally, considering the correlation between the top 50 terms of each conference edition and of all the others, an evolution on the research topics is shown. The edition of 2010 presented, on average, the highest Pearson's correlation coefficients between their top 50 terms and all the others (being positive for all cases). Moreover, recall from Figure 7 that WEBIST 2010 had as main research areas *Internet Technology*, *Society, E-Business and E-Government*, *Web Intelligence* and *Web Interfaces and Applications*, that are the only

(a) 2005

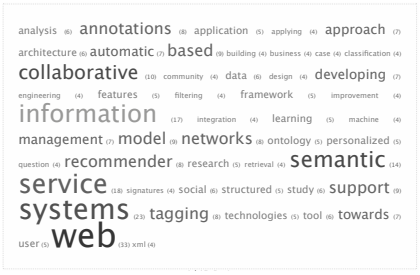
(b) 2006

(c) 2007



(d) 2008

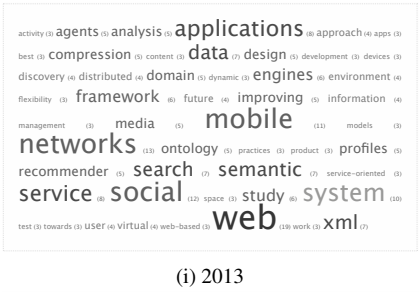
(e) 2009



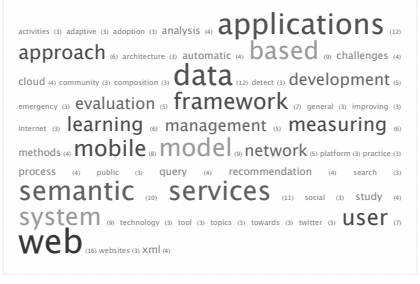
(f) 2010

(g) 2011

(h) 2012



(i) 2013



(j) 2014

Figure 9: Top 50 terms per conference year.

Table 5: Pearson's correlation between the frequency of top 50 terms from each conference edition.

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
2005		0.211	0.219	-0.012	0.128	0.178	0.105	-0.004	0.082	0.080
2006			-0.035	0.390	0.241	0.205	0.059	0.253	0.294	0.170
2007				0.174	0.178	0.259	0.084	0.189	0.178	0.140
2008					0.341	0.289	0.088	-0.007	0.118	0.005
2009						0.036	0.206	0.250	0.203	0.013
2010							0.325	0.316	0.404	0.122
2011								0.175	0.245	-0.103
2012									0.135	0.106
2013										0.395
2014										

areas that occur in the majority of conference editions (the "core" of research areas).

4.5 Paper Citation Analyses

In this section, we performed an analysis related to the WEBIST topmost cited papers (recall for Section 3 how these topmost papers were obtained) and estimated the *h*-index for the WEBIST conference series. The *h*-index obtained was 18, indicating that there are at least 18 papers with at least 18 citations. Thus, Figure 10 presents the percentage of top 18 most cited papers per type of publication. The results show that the most cited papers are mostly full papers (more than 50%, corresponding to 10 papers).

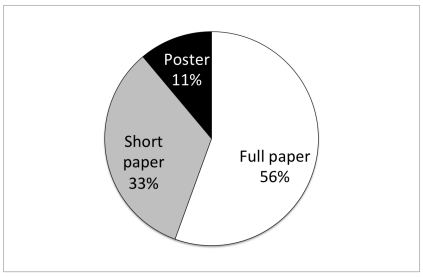


Figure 10: Top 18 most cited papers per type of publication.

Figure 11 presents the top 18 most cited papers based on the percentage per main research areas. It can be seen that the *Web Interfaces and Applications* and *Internet Technology* areas had the highest number of most cited papers in the top 18 (around 33% each). Surprisingly, *E-Learning*, which appeared only in the first four editions of WEBIST, had a higher percentage (around 17%) of the most cited papers than *Society, E-Business and E-Government* (around 6%) which appeared in all conference editions. As expected, the most recent main research areas do not have papers in the top 18 (2010 was the latest year with a paper in the top 18).

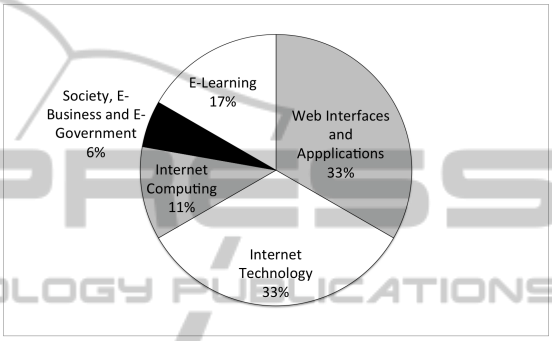


Figure 11: Top 18 most cited papers per main research area.

5 DISCUSSION AND OUTLOOK

This paper described the *WEBIST Dataset* and the *WEBIST Analytics* Web application. *WEBIST Dataset* aggregates data from different sources and follows the Linked Data principles. *WEBIST Analytics* provides different functionalities for the search, analysis and visualisation of data loaded in the *WEBIST Dataset*.

We also conducted a comprehensive analysis of 2005-2014 editions of WEBIST which showed the rapid popularity achieved by WEBIST in 2007 and its maturity along of the subsequent years, reaching a stable conference-size, community of IS experts, research topics of interest and possible supporters. Moreover, our analysis highlighted that the unbiased reviewing process of WEBIST contributed to the fast advancement of IS and the generation of knowledge for the community. The WEBIST community plays a key role in knowledge transfer and impact in IS (*h*-index =18). The *Web Interfaces and Applications* and *Internet Technology* tracks have been crucial to the development and popularity of the WEBIST conference series, as they accumulated the most cited papers. An important point to note and for future debate between WEBIST chairs is that the extinguished *E-Learning* track, which appeared only four times as a main track, obtained a proportion of top cited

papers higher than the *Society, E-Business and E-Government* track, which appeared in all conference editions. Finally, although the conference topics discussed by WEBIST authors have become more homogeneous over the last years, a higher diversity of topics/terms has also been observed.

Furthermore, the main contribution of this paper is not limited to the analysis of the WEBIST conference series, but includes the dataset and the Web application that serves as a baseline for future analysis and debate. As future work, we intend to extend the proposed workflow to analyse multiple conferences and researchers from different fields.

ACKNOWLEDGEMENTS

This work was partly funded by CNPq under grant 444976/2014-0, 303332/2013-1, 442338/2014-7 and 248743/2013-9, by FAPERJ under grant E-26/101.382/2014 and E-26/201.337/2014 and by CAPES under grant 1410827.

REFERENCES

Batista, M. G. R. and Loscio, B. F. (2013). OpenSBBD: Usando linked data para publicação de dados abertos sobre o SBBD. In *Brazilian Symposium on Databases - SBBD 2013, Short Papers*.

Berners-Lee, T. (2006). Linked Data. In *Design Issues*. W3C.

Bizer, C. and Seaborne, A. (2004). D2RQ - Treating Non-RDF Databases as Virtual RDF Graphs. In *Proc. 3rd International Semantic Web Conference*.

Blanchard, E. G. (2012). On the WEIRD nature of ITS/AIED conferences - A 10 year longitudinal study analyzing potential cultural biases. In *Proc. Intelligent Tutoring Systems - 11th International Conference, ITS 2012*, volume 7315 of *LNCS*, pages 280–285. Springer.

Borges, E. N., de Carvalho, M. G., Galante, R., Gonçalves, M. A., and Laender, A. H. F. (2011). An unsupervised heuristic-based approach for bibliographic metadata deduplication. *Inf. Process. Manage.*, 47(5):706–718.

Chen, C., Song, I.-Y., and Zhu, W. (2007). Trends in conceptual modeling: Citation analysis of the er conference papers (1975-2005). In *Proc. 11th International Conference on the International Society for Scientometrics and Informatrics*, pages 189–200. CSIC.

Chen, C., Zhang, J., and Vogeley, M. S. (2009). Visual analysis of scientific discoveries and knowledge diffusion. In *Proc. 12th International Conference on Scientometrics and Informetrics (ISSI 2009)*, pages 874–885.

Cheong, F. and Corbitt, B. J. (2009). A social network analysis of the co-authorship network of the pacific asia conference on information systems from 1993 to 2008. In *PACIS*, page 23. AISeL.

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Trans. on Knowl. and Data Eng.*, 19(1):1–16.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks*, 1:215–239.

Gasparini, I., Kimura, M. H., and Pimenta, M. S. (2013). Visualizando 15 anos de IHC. In *Proc. 12th Brazilian Symposium on Human Factors in Computing Systems*, IHC '13, pages 238–247. SBC.

Gini, C. W. (1912). Variability and mutability, contribution to the study of statistical distributions and relations. *Studi Economico-Giuridici della R. Universita de Cagliari*.

Henry, N., Goodell, H., Elmqvist, N., and Fekete, J.-D. (2007). 20 years of four HCI conferences: A visual exploration. *Int. J. Hum. Comput. Interaction*, 23(3):239–285.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proc. National Academy of Sciences of the United States of America*, 102(46):16569–16572.

Hoover, E. M. (1941). Interstate redistribution of population, 1850?1940. *The Journal of Economic History*, 1:199–205.

Hoser, B., Hotho, A., Jäschke, R., Schmitz, C., and Stumme, G. (2006). Semantic network analysis of ontologies. In *Proc. 3rd European Semantic Web Conference*, volume 4011, pages 514–529. Springer.

Lopes, G. R., da Silva, R., Moro, M. M., and de Oliveira, J. P. M. (2012). Scientific Collaboration in Research Networks: A Quantification Method by Using Gini Coefficient. *IJCSA*, 9(2):15–31.

Marsden, P. V. (2002). Egocentric and sociocentric measures of network centrality. *Social Networks*, 24(4):407–422.

Newman, M. E. J. (2001). Scientific collaboration networks. I. network construction and fundamental results. *Physical Review E*, 64(1):016131.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review*, E 69(026113).

Posada, J. E. G. and Baranauskas, M. C. C. (2014). A study on the last 11 years of ICEIS conference - as revealed by its words. In *Proc. 16th International Conference on Enterprise Information Systems, Volume 3*, pages 100–111. SciTePress.

Procopio Jr., P. S., Laender, A. H. F., and Moro, M. M. (2011). Análise da rede de coautoria do Simpósio Brasileiro de Bancos de Dados. In *Brazilian Symposium on Databases - SBBD Posters*.

Rodgers, J. L. and Nicewander, A. W. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1):59–66.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: methods and applications*. Cambridge University Press.

Zervas, P., Tsitmidelli, A., Sampson, D. G., Chen, N.-S., and Kinshuk (2014). Studying research collaboration patterns via co-authorship analysis in the field of TeL: The case of Educational Technology & Society Journal. *Educational Technology & Society*, pages 1–16.