Dissecting the Butterfly: Representation of Disciplines Publishing at the Web Science Conference Series

Clare J. Hooper
Culture Lab
School of Computing Science
Newcastle University, UK
clare.hooper@newcastle.ac.uk
+44 191 2464646

Nicolas Marie
Social communication/Wimmics
Alcatel-Lucent Bell Labs/INRIA
Villarceaux/Sophia-Antipolis,FR
nicolas.marie@alcatel-lucent.com
+33 1 3077 7184

Evangelos Kalampokis
Information Systems Lab
University of Macedonia
Thessaloniki, Greece
ekal@uom.gr
+30 2310 891576

ABSTRACT

Web Science is an interdisciplinary arena. Motivated by the unforeseen scale and impact of the web, it addresses web-related research questions in a holistic manner, incorporating perspectives from a broad set of disciplines. There has been ongoing discussion about which disciplines are more or less present in the community, and about defining Web Science itself: there is, however, a dearth of empirical work in this area.

This research note presents an early analysis of the presence of different disciplines in the Web Science community. To gain insight into this area, we applied Natural Language Processing and topic extraction to Web Science papers from 2009 to 2011. We compare the results to two current representations of Web Science: the 'Web Science butterfly' diagram and the Web Science Subject Categorization. We discuss the benefits of such an exploratory analysis, our early results, and steps for producing more robust results.

Categories and Subject Descriptors

K.4.m [Computers and Society]: Miscellaneous

General Terms

Human Factors, Measurement, Theory

Keywords

Web Science, community analysis, bibliometrics, disciplines

1. INTRODUCTION

There has been ongoing discussion about the representation of various disciplines within the Web Science community. Forming a stable, diverse community is no small task: members of the Web Science Trust have worked to try and ensure that the community is balanced with a rich variety of well represented disciplines, and not dominated by one field such as Computer Science.

Figure 1 shows the 'Web Science butterfly' diagram, which was used early on in the life of Web Science to convey the vision [8]. Nowadays it is sometimes used to describe the community, yet there is no evidence that the butterfly is an accurate depiction.

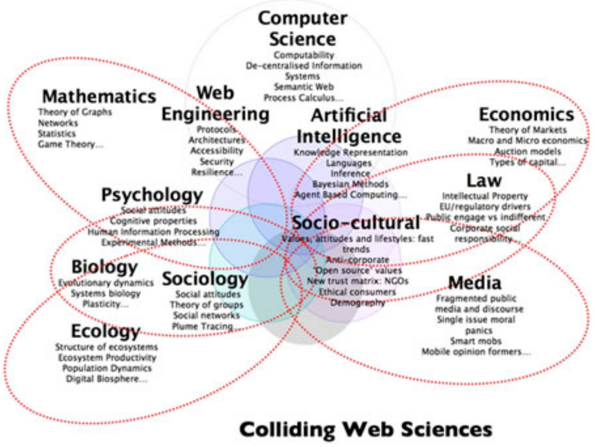

Figure 1 The Web Science 'butterfly' [8]

We can also consider the tree-based classification of Web Science subjects [9]. Like the butterfly, it lets us see subjects that are deemed to be relevant, but it reflects a vision and structure rather than providing information on these subjects' prevalence or the composition of the community.

We are unaware of work that empirically examines disciplinary representation in Web Science. This paper describes our initial efforts in this area: we took a corpus of papers from the first three Web Science conferences, used Natural Language Processing to extract topics from these, and conducted a network analysis of the resultant materials. This helped us see which disciplines were represented in the published papers, letting us 'take the temperature' of the Web Science community.

Such an analysis offers various benefits:

1. The Web Science butterfly is used to explain Web Science. By making it clearer and more accurate, we can communicate better as a community and reach out to other communities with whom we would like to engage.
2. We can ground community dialogue about diversity and disciplinary representation with data, seeing which

WebSc12, June 22-24, 2012, Evanston, Illinois, United-States.

disciplines are more or less represented, and which disciplines appear to be absent.

3. We can identify problems that we should be addressing regarding disciplinary representation, see what types of research are missing, and see what kinds of collaborations we might wish to encourage.

The field of bibliometrics is relevant to our questions, including work from co-citation analysis [2] [10], to examination of multiple conference series [5], to geospatial visualisations of collaboration [7]. Little prior work analyses the disciplinarity of conferences, although it is of note that Web Science students at the University of Southampton produced an illustration of their own disciplines (based on supervisor disciplines) in March 2011[1]. We conducted an analysis on past Web Science papers. Section 2 describes the method and results, and is followed by a discussion.

2. APPROACH

We analysed papers published at the Web Science conferences from 2009 to 2011. This corpus is available online and consists of 91 papers. We conducted topic extraction with Saffron [6], an application to help understand research communities. It can use information extracted from unstructured documents with Natural Language Processing techniques. We used the topic extraction component with the following parameters: maximum topic length 5; web filter minimum 5 hits; web filter maximum 1 billion hits. We used the ACM Subject classification to build linguistic patterns for topics in the Computer Science area.

This yielded 236 tokens that Saffron identified as research topics (although it returned no result for 22 of the papers). We kept only the 96 tokens that were found in more than two papers.

We cleaned the dataset with Google Refine[2], a tool for cleaning and analysing data. We amended misspellings, removed white space, merged synonyms, and discarded topics that were irrelevant to our question of disciplinarity. (For example, topics such as ‘future work’ and ‘participation’ are in use across disciplines.) This left 77 topics. The 15 most commonly occurring topics are shown in Table I.

We used a network graph tool, NodeXL[3], to build a graph showing links between topics (Figure 2): nodes correspond to extracted topics and arcs to papers that link them. This representation let us identify ‘clusters’ of closely related topics.

[1] http://thewebscientist.co.uk/index.php/2011/03/01/webscienceinpractice/ All URLs accessed February 2011.

[2] http://code.google.com/p/google-refine/

[3] http://nodexl.codeplex.com/

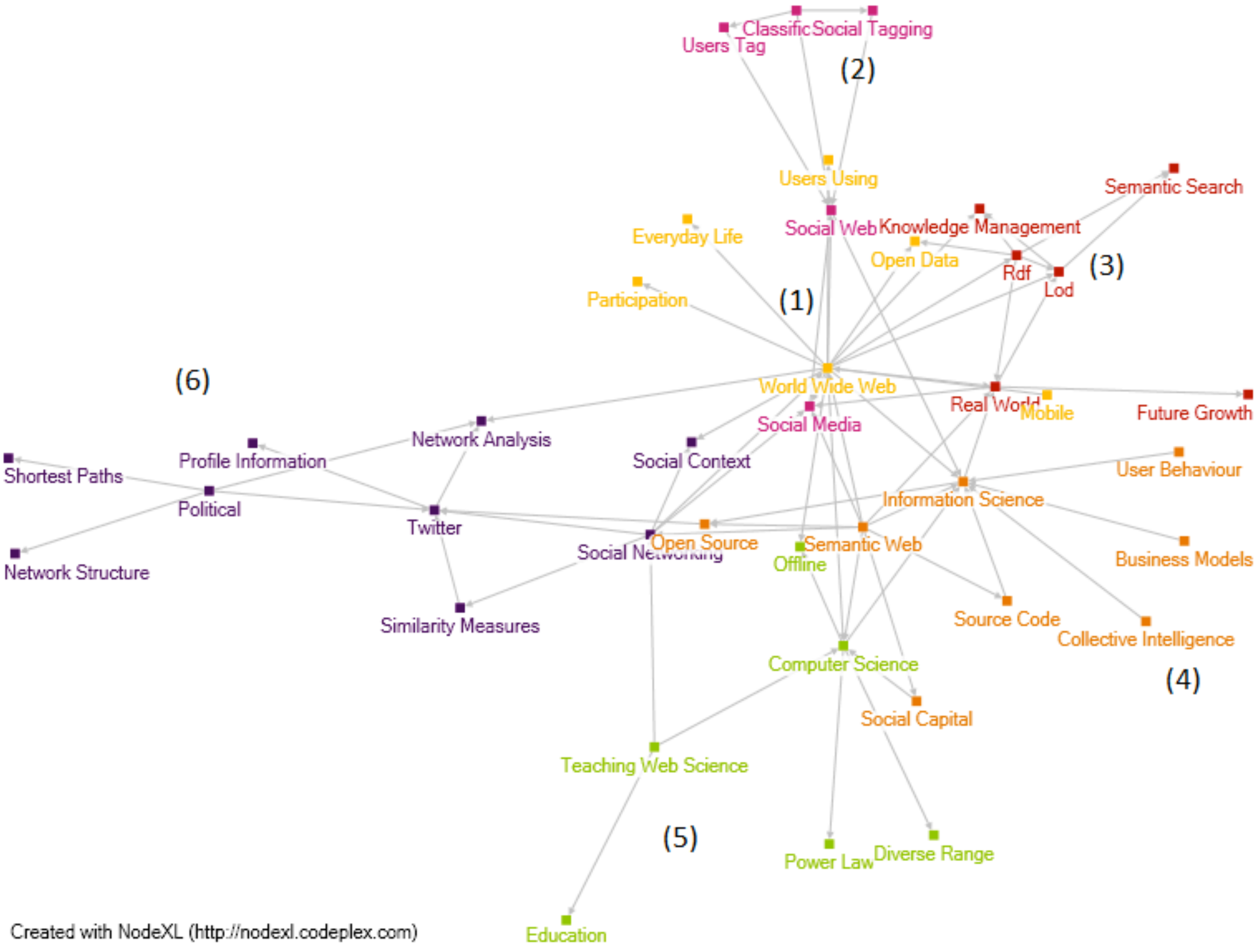

Figure 2 Web Science topics as linked by papers.

Topic	Count
Computer Science	31
Semantic Web	26
Real world	18
Social networking	15
Network Analysis	12
Social web	12
Open source	11
Information retrieval	11
Open data	10
Web users	9
Web data	9
Search results	8
Information science	8
Web search	8
Knowledge management	7

Table I The 15 most commonly occurring topics in the corpus

It is clear that the topics *World Wide Web* and *Social Media* are central and highly connected. This is unsurprising: these topics are core to Web Science. The following list presents some observations about Figure 2: list numbers correspond to numbers in the figure.

1. Technology use, its impact and users implications. (Topics: *Everyday life, Mobile, Participation, Open data.*)
2. Tagging and content classification. (*Social Web, Social Media, Users tag, Classification, Social tagging, Social Web, Social platforms*).
3. Semantic Web, including standards (*RDF*), implementation (*LOD*), application (*Knowledge Management),* and research topics (*Semantic Search*).
4. User collaboration and co-creation. (*Collective intelligence, Social capital, Open source, Source code)*
5. A mixed set including education (*Teaching Web Science, Education*), a theme absent from the Web Science butterfly.
6. A mixed set including Social Network and Analysis concepts (*Social Networking, Social Context, Twitter, Network Analysis, Shortest paths, Network structure)*. *Political* is connected to *Twitter* and *Social Network Analysis*, illustrating the application of Computer Science technique to societal topics.

We calculated the betweenness centrality of topics. Betweenness centrality measures the fraction of shortest paths going through a given node [1]. High betweenness centrality indicates that nodes play an important bridging role in a network. Here, it lets us identify topics with a high likelihood of bridging disciplines. Table II shows topics with a betweenness centrality above 100.

Topic	Betweenness centrality
World wide Web	820
Information Science	319
Computer Science	242
Social Web	219
Social Networking	196
Network Analysis	175
Political	150
Twitter	148
Semantic Web	133
Real World	104

Table II Topics with betweenness centrality > 100

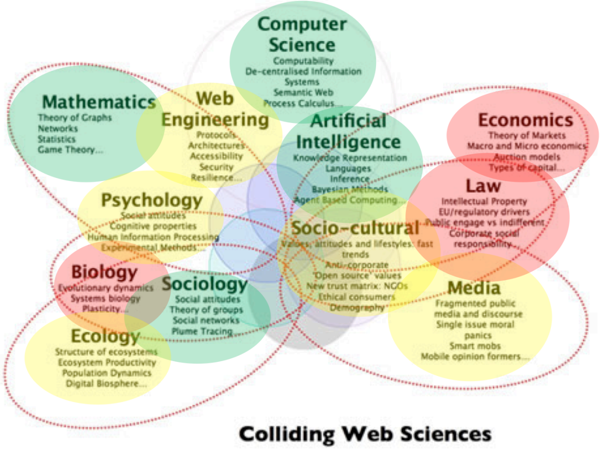

Figure 3 Butterfly heat map: green topics are well represented, yellow somewhat represented, red absent.

We mapped our findings about topic the presence and linkage to the original Web Science butterfly, providing a 'heat map' of disciplinary presence (Figure 3). N.B.: The distinction between 'well represented' and 'somewhat represented is unsophisticated, based on the authors' beliefs about the mapping of topics to disciplines. For instance, the prevalence of topics such as 'Semantic Web', 'Knowledge Management' and 'Computer Science' led us to identify Computer Science as well represented; the presence of 'Open Source' and absence of other terms from the 'Socio-Cultural' label led us to classify that domain as somewhat represented.

We also considered the results from the perspective of the Web Science subject classification [9]. Table III shows the relative presence of the Web Science categories in the topics that we identified.

3. DISCUSSION AND FUTURE WORK

It is clear that the butterfly under-represents some disciplines and omits others. For example, many topics implied the presence of Network Science, yet it is only demarcated as a sub-topic ('Networks') beneath Mathematics in the butterfly. Meanwhile, Politics and Education are both clearly present in the community, but absent from the butterfly. Some disciplines are very healthy (notably Computer Science and Sociology), while others are badly under represented (Biology, Economics, Law). Meanwhile, the following areas from the Web Science Subject Categorization are absent from the extracted topics: Economics and Business, Personal Engagement and Psychology, Philosophy, and Law.

This relative presence (and absence) of disciplines helps us

Topic	Presence
A. General	N/A
B. Web History and Methodology	2
C. Web Technologies	2
D. Web Analysis	2
E. Web Society	2
F. Teaching the Web	1
E.1 Economics and Business	0
E.2 Social Engagement and Social Science	2
E.3 Personal Engagement and Psychology	0
E.4 Philosophy	0
E.5 Law	0
E.6 Politics and Governance	1
E.99 Other in Web Society	N/A

Table III Classification heat map
0 = absent, 1 = somewhat present, 2 = very present

identify potential weaknesses in the Web Science community: if we are not addressing pertinent topics (such as the web and Economics or the legal implications of aspects of the web), we can consider reaching out to relevant communities in search of collaboration.

It should be made clear that this work only represents a first pass at this problem area. Due to time constraints and technical problems processing the papers, only 69 of the 91 papers were included in the final analysis. Furthermore, our method for deciding whether a discipline was strongly or weakly present could be much more robust: one approach might be to ask independent experts to list which disciplines they associate with the topics, and use this as a measure.

Our aim in this paper, however, is not to present a complete analysis of the Web Science community, but to demonstrate that such an analysis is possible while sharing our early results.

One problem with the idea of topic-based disciplinary analysis is that topics do not necessarily directly map to disciplines. Other measures can be gleaned from papers (category / subject descriptors; general terms; key words). Alternatively, researchers could identify: the discipline with which authors self-identify (this can be subjective); the research departments from which authors hail; or the home discipline(s) of the research methods used.

It is also of note that measuring the number of papers (or authors, or topics, or key words) from a discipline is only one measure of that discipline's presence: it does not measure impact. For example, the above methods would not acknowledge situations such as a conference where very few philosophical papers were presented, yet where one such paper [4] had a large impact.

There are further questions of interest for a richer analysis:

- What changes occur in discipline presence over time? For instance, did collocating with ACM WWW in 2010 result in more technical submissions?
- What disciplinary differences exist between posters and papers?
- What disciplines collaborate? Does this vary between posters and papers? (Consider co-authorship and co-citation.)
- What links exist between social networks and citation networks? (Perhaps #websci12 shall yield a corpus of Twitter data.)

4. CONCLUSION

Web Science has been described as a way to 'take the temperature' of the Web [3]. This paper concerns taking the temperature of the Web Science community, towards supporting good representation of different disciplines. As well as providing evidence about the diversity and health of our community, this helps us explain the nature of Web Science to outsiders.

Although the results are early and need to be refined, we hope that this work has raised awareness of the benefits of conducting an analysis of disciplinary presence in the Web Science community, demonstrated techniques by which this could be achieved, and presented early results that begin to indicate the state of the community.

5. ACKNOWLEDGEMENTS

With grateful thanks to Georgeta Bordea for her help with Saffron, and Marc Smith for his help with NodeXL.

6. REFERENCES

[1] Barthélémy M. 2004. Betweenness centrality in large complex networks, *The European Physical Journal B - Condensed Matter and Complex Systems*, 38,4, 163-168

[2] Chen, C., Carr, L. 1999. Trailblazing the literature of hypertext: author co-citation analysis (1989–1998), in *Proc. 10th ACM Conference on Hyperext and hypermédia,* 51-60.

[3] Hall, W., 2011. *Opening keynote*, Web Science Doctoral Summer School, DERI, NUI Galway, Ireland.

[4] Halpin, H., Clark, A., Wheeler M. Towards a Philosophy of the Web: Representation, Enaction, Collective Intelligence, in *Proc. of the WebSci10*

[5] Henry, N., Goodell, H., Elmqvit, N., Fekete, J. 2007. 20 Years of 4 HCI Conferences: A Visual Exploration. *International Journal of Human Computer Interaction - Reflections on Human-Computer Interaction*, 23(3), 239-285.

[6] Monaghan, F., Bordea, G., Samp, K., Buitelaar, P. 2010. Exploring Your Research: Sprinkling some Saffron on Semantic Web Dog Food, in *Semantic Web Challenge at the International Semantic Web Conference*.

[7] Nagel, T, Duval, E., Heidmann, F., Exploring a Geospatial Network of Scientific Collaboration on a Multitouch Table.

[8] Shadbolt, N., *What Is Web Science?* talk, http://Web Scienceence.org/Web Scienceence.html

[9] Vafopoulos, M. 2010. Web Science Subject Categorization (WSSC), in *Proc. of the WebSci 2010*

[10] White, H.D. 1998. Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972–1995, *Journal of the American Society for Information Science*, 49, 4, 327–355.