Crowdsourcing Scholarly Data

Diep Thi Hoang* Jasleen Kaur Filippo Menczer

Center for Complex Networks and Systems Research
School of Informatics and Computing
Indiana University, Bloomington

ABSTRACT

One of the promises of Web Science is to leverage the wisdom of the crowds to give rise to emergent, bottom-up semantics, by making it easy for users to express relationships between arbitrary kinds of objects. Rather than starting with an ontology that determines the kinds of objects and relationships to be described and reasoned about, the idea is to give users the freedom to annotate arbitrary objects with arbitrary predicates, along with incentives for such annotations. Social tagging systems for images are one example, where the motivation can stem from the wish to organize and share one's photos or from entertaining games to guess one another's tags. Here we explore a similar approach in the domain of scholarly publications. We describe a system called Scholarometer, which provides a service to scholars by computing citation-based impact measures. This motivates users to provide disciplinary annotations for authors, which in turn can be used to compute for the first time measures that allow to compare authors' impact across disciplinary boundaries. We show how this crowdsourcing approach can lead to emergent semantic networks to study interdisciplinary annotations and trends.

Keywords

Crowdsourcing, citation analysis, scholarly data, impact measures, discipline annotations, social tagging

1. INTRODUCTION

The rapid growth of online scholarly repositories and digital libraries brings the challenge of how to organize, categorize, and retrieve the vast collections of articles contained in these repositories. Many disciplinary communities have over time developed their own classification systems to help address these issues. Examples include the ACM Computing Classification System for computer science, the Medical Subject Headings (MeSH) for the life sciences, the Physics and Astronomy Classification Scheme (PACS) for physics, and so on. Unfortunately these disciplinary categorizations make it difficult for the different communities to understand each other's literature, creating obstacles toward interdisciplinary collaboration, and leading to a more fractured scientific landscape. As a result, we see efforts to develop multidisciplinary classification schemes, the primary example being the citation indices maintained by Thomson-Reuters as part of their Journal Citation Reports (JCR) and Web of Science (WoS) commercial products. Since these indices are classifications of journals rather than articles, and they are maintained by a central authority in a top-down fashion, such approaches have serious drawbacks. One is their granularity: not all articles in a particular journal are equally well described by the categories assigned to the journal, and conversely, it is hard to identify individual articles that have a truly interdisciplinary nature. Additionally, it is very difficult for such general classification schemes to keep track of rapidly changing scientific fields: the important trends leading to emergent and dying disciplines often occur at the boundaries between established areas. Therefore a universally agreed shared vocabulary for the classification of scholarly output is unlikely achievable with such top-down efforts.

Web Science suggests ways to address the above problem. One of the promises of Web Science is to leverage the wisdom of the crowds to give rise to emergent, bottom-up semantics, by making it easy for users to express relationships between arbitrary kinds of objects. Rather than starting with an ontology that determines the kinds of objects and relationships to be described and reasoned about, the idea is to give users the freedom to annotate arbitrary objects with arbitrary predicates, along with incentives for such annotations. Here we explore such an approach in the domain of scholarly publications. If we can create appropriate incentives for scholars to annotate authors and/or articles with disciplinary labels, we can achieve several goals simultaneously. First, we get a dynamic classification that can evolve in a scalable way with the growing number of authors, articles, and specializations. Second, scholars who collaborate across disciplinary boundaries will naturally tend to use a shared vocabulary to facilitate such collaborations. Third, a flat tagging approach is more flexible and fluid compared to hierarchical classifications, making it easier to annotate contributions that do not belong to a single disciplinary branch. Fourth, while shared hierarchies or vocabularies cannot be forced, the bottom-up approach allows to track emergence of structure and consensus.

A related challenge is the evaluation of an author's scholarly output. In a quest for quantitative impact analysis, a wealth of measures based on citation data have been proposed and new ones are being formulated almost on a daily basis. Of course each measure has its strengths and weaknesses, proponents and detractors. Among the limitations of

*Contact author. Email: dihoang@indiana.edu

Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.

most citation based impact measures proposed thus far, we focus here on the challenges posed by disciplinary boundaries. Different disciplines have widely heterogenous communities with different numbers of authors, productivity, citation patterns, and cultural traditions. How do we compare a historian who writes a book after years of research, with a mathematician who publishes a long article proving a theorem, with a medical scientist who works in a large team, with an experimental physicists whose journal articles have 50 authors, or with a computer scientists who publishes almost exclusively in yearly conference proceedings?

One way to account for the diverse citation patterns in different areas is by looking for universal regularities. Radicchi *et al.* [18] have discovered that citations follow a universal distribution across disciplines when rescaled by appropriate discipline-specific statistical quantities. Based on this, they have proposed a universal impact measure that would enable to compare authors in different disciplines in spite of different citation patterns. However, implementing such an approach requires the availability of a citation database equipped with a universal classification system. The Web Science approach described above for addressing the problem of a scholarly classification system can be leveraged to achieve this goal as well in combination with citation data. If we can create appropriate incentives for users to share citation data about the authors they annotate, the two goals — a citation database and a universal classification system — are met simultaneously. This can give us access to citation and publication data for individual disciplines, making it possible to compute the universal impact measure proposed by Radicchi *et al.*

What we have described is an instance of crowdsourcing, i.e., harnessing knowledge from a community via Web platforms in order to solve practical problems. As incentives for the knowledge provided, users may receive cash or some other reward. For instance, in a game setting, users may perform work in exchange for a chance to be entertained [21]. In our application of crowdsourcing to scholarly annotations, users have access to citation data, which they can obtain by querying services (such as Google Scholar, CiteSeer, Scopus, and Web of Science), and which they may freely share with the public. Furthermore, users who query about a particular author are in a position to annotate the author in question with appropriate disciplinary tags. We want scholars to share these two pieces of information. We propose a framework for collecting such information from users in exchange for a citation analysis service. The idea is to provide a social client interface to an existing Web source of scholarly data, allowing users to perform academic impact analysis based on author queries.

In our social approach to scholarly citation analysis, the crowdsourced information forms the very basis for the service provided. Note that some of this information comes directly from the users (the discipline tags), while other information is obtained indirectly as a side effect of user queries (the citation data). Citation data may be public or proprietary, based on how it is collected. For example, if it is obtained by crawling and parsing publications that are openly available on authors' homepages, it is clearly public. On the other hand, if its source is a commercial publisher, such as a subscription-based digital library, then citation data may be proprietary. Here we assume that the citation data is from a public source and that once users have obtained citation data from some service, they are free to share this information publicly. By using a social client interface, users can obtain citation data from a public source and then share it with other users.

Outline and Contributions

In this paper we introduce Scholarometer, a crowdsourcing tool we developed for scholarly services. After some background on related research, in the remainder of the paper we make the following contributions:

- We describe the architecture, user interface, data model, and heuristics used in the design of the Scholarometer system. (§ 3)
- To date, since the first release of Scholarometer, we have collected reliable information about 4,211 authors in 428 disciplines. Based on this data, we can create universal or disciplinary rankings of authors, according to various citation analyses. We discuss the differences in the ranked lists of top authors obtained by various impact measures. We also report on some statistics about citation patterns across disciplines. (§ 4.1)
- By leveraging the socially collected discipline statistics, we implement for the first time the universal h index [18]. We also study the convergence of relative bibliometric indicators used in the computation of the universal h index. We show that these statistics are pretty stable, suggesting that the universal h index can be a reliable indicator for comparing the scholarly impact of individual authors in different disciplines. (§ 4.2)
- As an illustration of other potential applications of crowdsourced scholarly data, we report on our first attempt to map interdisciplinary collaborations. The resulting network, in turn, suggests that the crowdsourcing framework yields a meaningful classification scheme for authors and their disciplinary interactions. (§ 4.3)

2. BACKGROUND

Many popular reference management tools can extract bibliographic information from online repositories and digital libraries. BibDesk has advanced features to search online resources and access digital libraries, such as PubMed [2]. Connotea allows users to import articles using Digital Object Identifiers (DOI) [6]. Zotero can capture bibliographic information from Web pages and import items by identifiers such as ISBN, DOI, or PubMed ID [8]. These and many other bibliographic management tools are compared in the Wikipedia [5].

The idea of tagging scholarly work is also not new. Tools like BibSonomy [3], Connotea [6], and CiteULike [4] allow users to freely tag articles that are shared online. Our system is slightly different in that we ask users to tag authors rather than articles. Furthermore, a tag in Scholarometer is supposed to be a scientific discipline.

As pointed out by Alonso *et al.* [10], there are three multidisciplinary citation databases that are increasingly being used for scientific evaluation purposes: Web of Science, Scopus, and Google Scholar. Among them, only Google Scholar is freely available online. Additionally Google Scholar claims

to cover articles, theses, books, abstracts, court opinions and other scholarly literature from all areas of research [7]. Therefore we choose Google Scholar as a source for bibliographic and citation data.

Google does not provide an API for Google Scholar, supposedly because of agreements with publishers. A Web service that crawls and parses Google Scholar results to extract and/or store information on a server would be in violation of Google's directives, expressed via the Robots Exclusion Protocol. There have been several attempts to get around this limitation. ScHolar index [19] uses configurable proxies to get its server-side scripts to pass Google's IP address checks. Citations-gadget [12] is a Google Gadget, so its Ajax requests originate from Google and comply with the same-origin policy. Publish or Perish [14] is a desktop application, therefore it acts as a client rather than a server, so that requests do not come from a single server IP address and cannot be blocked.

Bibliometrics is the use of statistical methods in the analysis of scholarly data to reveal patterns of authorship, publication, and use. It includes citation analysis, which is used to explore and measure the impact of a research field, of one or more researchers, or of a particular paper. There are many measures to calculate the impact of authors. Hirsch's original h index [15] is defined as the maximum number of articles h such that each has received at least h citations. Egghe's g index [11] gives more weight to publications with many citations; it is the highest number g of papers that together receive g^2 or more citations. Schreiber's h_m index [20] and Hirsch's $\hbar$ [16] are attempts to apportion citations fairly for papers with multiple authors. Finally, Radicchi *et al.*'s universal h-index h_f [18] allows to quantitatively compare the impact of authors in different disciplines, with different citation patterns. Pudokvin and Garfield [17] have proposed universal impact measures based on percentiles. New citation-based impact measures are being introduced all the time. Each has its own advantages and disadvantages (see, e.g., Adler *et al.* [9] for a critique). Furthermore, their values depend on the citation database used as a source. Scholarometer incorporates several of the above measures, discussed in § 4.2.

We propose to collect scholarly data by crowdsourcing. The most popular example of crowdsourcing is Amazon's Mechanical Turk [1], a Web marketplace to coordinate the use of human intelligence for tasks that computers are unable to perform. Solutions for tasks such as choosing the best among several photographs of a store-front, writing product descriptions, or identifying performers on music CDs are distributed to a team of workers, who are then paid by the requester. Games with a purpose [21] are another class of crowdsourcing applications, based on entertainment rather than monetary payments. The best-known example is the ESP game, used to tag images. Here we describe a variation of these ideas, where annotation data is generated by users in exchange for a service, which itself is based on the data provided by the users.

3. SYSTEM IMPLEMENTATION

In this section we outline the main features of the Scholarometer system, which is under development and is available online.[1]

[1] `scholarometer.indiana.edu`

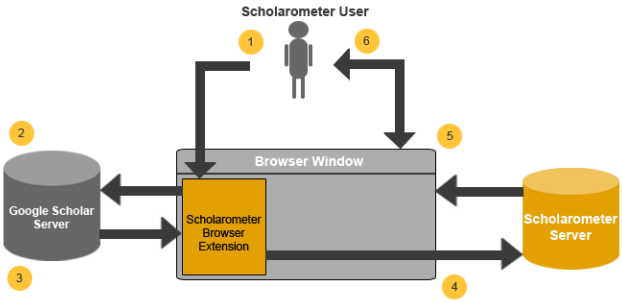

Figure 1: The Scholarometer workflow.

3.1 Architecture

The Scholarometer tool uses Google Scholar as a data source. As discussed above, Google Scholar provides freely accessible publication and citation data to users, without requiring a subscription. This leads to a caveat in the use of Scholarometer: the analysis can only be as good as the data source. Google Scholar provides excellent coverage, in many cases better than Web of Science, for example in disciplines such as computer science, which are dominated by conference proceedings; or some social sciences, which are dominated by books. Nevertheless, Google Scholar is based on automatic crawling, parsing, and indexing algorithms, and therefore its data is subject to noise, errors, and incomplete or outdated citation information. This limitation of course applies to any tool that uses the same source. The system architecture and design that we describe below are independent of the data source; they would apply equally if we were to use an alternative source, such as CiteSeer (`citeseerx.ist.psu.edu`).

As discussed above, the lack of an API to access Google Scholar makes a server-based implementation infeasible, as it would violate Google Scholar's policy about crawling result pages, extracting data (by parsing/scraping) and making such data available outside of the Google Scholar service. Indeed, server-based applications that sit between the user and Google Scholar are often disabled, as Google Scholar can detect a large number of requests coming from a particular server and block its IP address. Palliative measures such as configurable proxies do not always work, and in any case are not a desirable solution as they appear to violate policy. Due to the same origin policy, one cannot leverage Ajax technology to build such a Web service either. We excluded the gadget approach because it would render the tool completely dependent on a particular data source (Google Scholar in our case), precluding the possibility of drawing scholarly data from any other sources. We therefore turned to a client-based approach. However, we ruled out a standalone application (such as Publish or Perish) for portability reasons. A browser-based implementation is platform and system independent. These design considerations led us to a browser extension approach. The idea is that Scholarometer is just a smart extension of the browser, through which the user queries the source, annotates the results, and shares with the Scholarometer server only open citation and annotation metadata.

The architecture and workflow of Scholarometer is illustrated in Figure 1. There are six steps: (1) First, the user enters a query and discipline tags for an author into

a search form provided by the browser extension. (2) The browser extension forwards the query to Google Scholar. (3) Google Scholar returns the query results to the browser extension. (4) The browser extension then forwards the results to the Scholarometer server. This parses the results to extract citation and other metadata, which is then inserted into the database, along with annotation metadata. (5) The Scholarometer server sends to the client browser the bibliographic records and impact measures for the queried author(s). (6) Finally, the client browser renders the data in an interactive way. The user views results in a new browser tab and can perform advanced actions such as sorting, filtering, deleting, and merging records.

3.2 User Interface

The Scholarometer tool has two interfaces for communicating with users: one in the browser extension for entering queries and tags, the other in the main browser window for presenting and manipulating bibliographic data and citation analysis results. The browser extension is available in two versions[2]: one for the Firefox browser hosted at the Mozilla Firefox Add-ons site, and one for Chrome browser hosted at the Google Chrome Extensions site. The Firefox interface is illustrated in Figure 2. The query interface in the browser extension is designed to identify one or more authors and retrieve their articles. The default interface hides many advanced features and simplifies the common case of a single author uniquely identified by name. Advanced interfaces are available for multiple authors with explicit Boolean operators, for ambiguous names with controls for filtering subject areas and languages, and with additional keyword fields. We provide an autocomplete feature to make it easy for users to enter discipline tags and reuse tags from other users.

The interface in the main browser window is designed to facilitate the manipulation and cleaning of the results, to visualize how the impact measures are calculated, and to expose annotations from other users for the same author(s). The output screen is divided into three panels:

1. A filter panel with two modules. One module is for pruning the set of articles based on the publication year or the number of citations. The second module is for limiting the set of articles to selected name variations or co-authors.

2. The list of articles, with utilities for live searching and for alternating between a simplified and an extended view, as well as links to external resources. This panel also has remove and merge utilities to correct two common sources of noise in Google Scholar results: articles written by homonymous authors and different versions of the same paper.

3. A citation analysis panel reporting impact measures. As discussed in § 2, many impact measures have been proposed, and it is impossible to implement them all. Since a single measure can only capture some aspect of scientific evaluation, a good citation analysis tool should incorporate a set of measures that capture different features, such as highly cited publications, coauthorship, and different citation practices. To this end we have implemented h, g, h_m, and h_f. Note that this is the first implementation of the universal h_f index, which is enabled by the joint availability of annotation and citation data, as explained in detail in § 4.2. The citation analysis panel displays h_f values for each discipline tag of an author, along with percentiles. Finally, the panel shows two plots illustrating the citation distribution and publications per year. All the data in the citation analysis panel is dynamically generated and updated in response to any filter, merge or delete actions performed in the other panels.

[2] `scholarometer.indiana.edu/download.html`

3.3 Database and Heuristics

Figure 3 illustrates the main structure of the Scholarometer database. The data we collect consists of annotation metadata along with citation data necessary to compute impact measures. Note that we do not store information about articles that the source intends to be only accessible to end users, such as titles, journals, publishers, links to source documents, or any other bibliographic information. Instead we only store user-generated metadata, such as author names, discipline tags, author-discipline annotations, and hash signatures that uniquely identify articles so that they can be associated with citation information.

The data that we collect comes from users, so it is naturally noisy and subject to various issues that make it necessary to perform some preventive data cleaning. The first challenge is that author names are often ambiguous. When the user queries for an author from the browser extension, our system first uses a heuristic to check for ambiguity in the name. We extract the name variations from Google Scholar, and sort them by number of citations. Typical author names have two or three variations (e.g., with and without a middle initial). Therefore we look at the percentage of the total citations that are accounted for by the top three name variations. If this is high (say above 90%) then we assume the name is not ambiguous, as any further variations only account for a small fraction of the citations and therefore do not have a large effect on impact measures. On the other hand, if the top three name variations account for a low fraction of the citations, we assume that the name is ambiguous and do not enter any data into our database.

A second issue is the arbitrary nature of discipline annotations. On one hand, free tags can be noisy, ambiguous, or duplicated (e.g., "human-computer interaction," "human computer interaction," and "hci"). On the other hand, using a controlled vocabulary, such as the JCR categories, does not allow for new/emerging disciplines to be easily captured and tracked. Therefore we attempt to strike a balance between the two extremes of completely uncontrolled and completely controlled tag vocabularies. We pre-populated the database with JCR categories — composed of Science Citation Index Expanded, Social Sciences Citation Index, and Arts & Humanities Citation Index. The user is required to select at least one of these predefined disciplines along with any

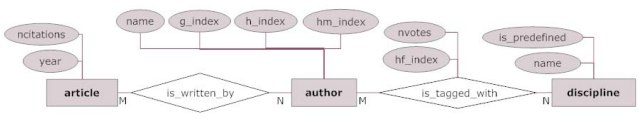

Figure 3: Simplified sketch of the Scholarometer data model.

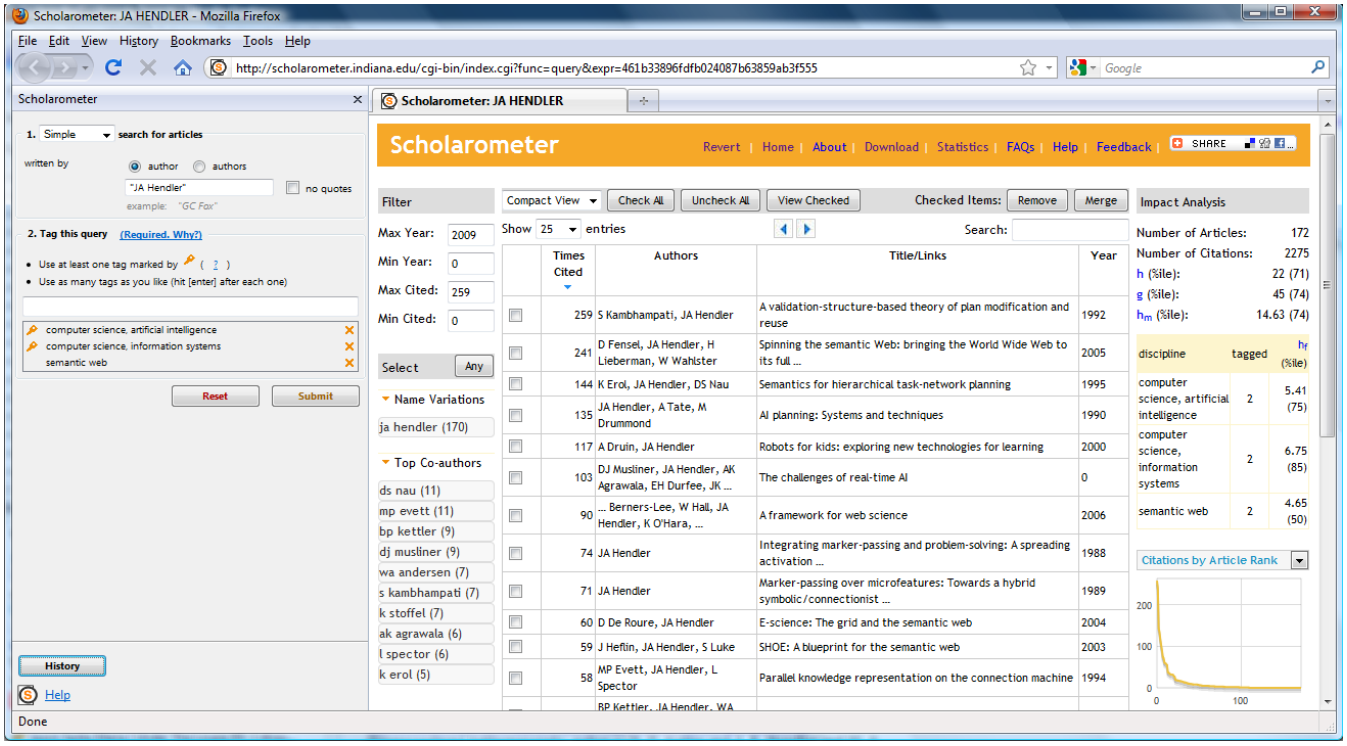

Figure 2: Illustration of the Scholarometer interfaces.

other free annotations, either tags from previous users or completely new tags.

Finally, users may be more or less precise (narrow) in their annotations, and as a result an author may be associated with noisy tags. Our crowdsourcing approach provides us with a natural solution to this problem. We view each query as a vote for the tag annotations of the queried author. For example, a query that tags Einstein with "physics" and "philosophy" generates a vote for *(Einstein, "physics")* and a vote for *(Einstein, "philosophy")*. We then use the number of votes together with the number of tags to determine heuristically which tags are reliable for each author. A tag is deemed reliable for an author with n tags if it has more than $S_{max} = -\log(1/n)$ votes. The intuition for this heuristic is that the more tags an author has, the greater the possible confusion (maximum entropy S_{max}), and therefore the greater the number of votes necessary for a tag to decrease the noise.

4. ANALYSIS OF DATA

4.1 General Statistics

The Scholarometer system was first released in November 2009. At the time of this writing, the Scholarometer database has collected information about 318,134 articles by 4,418 authors in 506 disciplines. There are 9,467 annotations, or tag-author pairs. Once we apply the heuristics described in § 3.3, we reduce these numbers to 4,211 reliable authors with 7,123 reliable annotations into 428 reliable disciplines. Naturally this folksonomy grows and evolves daily as the Scholarometer handles new queries.

Various statistics for authors and disciplines are available on the Scholarometer Web site.[3] The annotation data enables us to derive rankings for authors — both universal and disciplinary — based on impact measures. Table 1 shows the universal rankings of top authors by h, g, h_m, and h_f respectively. We can see that compared to the h index, the g index favors authors such as Giddens and Chomsky, with books that have received very high numbers (thousands) of citations. The h_m index favors authors with many top publications that are single-authored; Giddens and Chomsky are again good examples, as well as Marx, Gould, and Einstein. The universal h_f index brings to the top some authors whose citations are not as numerous in absolute terms, but who are leaders in their respective fields — chemistry Nobel prize winner Pauling and computer scientist Towsley are good examples.

Table 1: Top authors according to various impact measures (based on values as of March 25, 2010).

	h	g	h_m	h_f
1	Freud	Freud	Freud	Bourdieu
2	Bourdieu	Giddens	Bourdieu	Chomsky
3	Witten	Chomsky	Witten	May
4	Kandel	Bourdieu	Chomsky	Freud
5	Piaget	Piaget	Giddens	Caspi
6	Robbins	Shleifer	Marx	Kandel
7	Snyder	Williamson	Piaget	Pauling
8	May	Marx	May	Towsley
9	Lefkowitz	Kuhn	Gould	Lefkowitz
10	Chomsky	Barro	Einstein	Finkelstein

[3] `scholarometer.indiana.edu/statistics.html`

Table 2: Top authors tagged with "computer science, information systems" according to various impact measures (based on values as of March 25, 2010).

	h	g	h_m	h_f
1	G-Molina	G-Molina	G-Molina	Towsley
2	Towsley	Davenport	Towsley	G-Molina
3	Chakrabarti	Towsley	V.D. Aalst	Smith
4	Dey	Berners-Lee	Garfield	Dey
5	Jha	Jha	Chakrabarti	Fagin
6	Watson	Chakrabarti	Davenport	V.D. Aalst
7	V.D. Aalst	Dey	Watson	Jha
8	Liu	Bellare	Liu	Davis
9	Smith	Perrig	Davis	Staab
10	Fagin	Watson	Harrison	Brusilovsky

Table 2 shows an example ranking of top authors in a particular discipline ("computer science, information systems") by the same impact measures. Once again we observe that the g index ranks higher authors of books and other very highly cited publications, such as Davenport and Berners-Lee. Garfield has many top cited single-authored articles and as a result is highly ranked by h_m. Finally, in the ranking by h_f we see that some authors with high h are replaced by other well-known information scientists (Staab and Brusilovsky). Upon closer inspection we note that two of the replaced names, Watson and Liu, are actually ambiguous, referring to information scientists as well as other authors. Thus the universal h_f index has helped remove some noise from the rankings.

4.2 Universal H Index

The *universal h index*, which we refer to as h_f, was proposed by Radicchi *et al.* [18]. For each discipline tag and year, we maintain statistics about the average number n_0 of papers written by authors in that discipline and in that year, and about the average number c_0 of citations to papers written in that discipline and in that year. When we receive a query about an author in a certain discipline, we update these statistics. Additionally, following Radicchi *et al.*, we rescale the citations of each paper by c_0 (for the discipline of the author and the year of the paper) and we rescale the rank of each paper by n_0 (again for the given discipline and year). The universal h_f value for the author is the maximum rescaled rank h_f such that each of the top h_f articles have at least h_f rescaled citations each. Since the discipline/year statistics depend on the annotations we collect from queries, they are subject to noise and may take a while to converge. Once the statistics are reliable, one will be able to compare the impact of authors in different disciplines. Note that an author tagged with several disciplines will have multiple h_f values, one per discipline. Since different disciplines have different citation patterns, an author should only pay attention to h_f values in disciplines that s/he knows to be appropriate.

We have already shown in Tables 1 and 2 how h_f identifies top authors in their respective fields. To show how h_f also allows to compare the impact of authors across disciplines, let us consider an example. The two authors G.A. Parker and R. Weibel are both are highly successful in their own disciplines — biology and geography, respectively. Their impact cannot be compared based on the h index as the two disciplines have different numbers of authors, publications, and citation patterns. Indeed, Parker has $h = 73$ while Weibel has $h = 30$, suggesting that the former has a much greater impact than the latter in absolute terms. However, when we compare the two based on the universal h_f index, we find that both authors are equally successful in their respective fields, having the same $h_f = 3.6$.

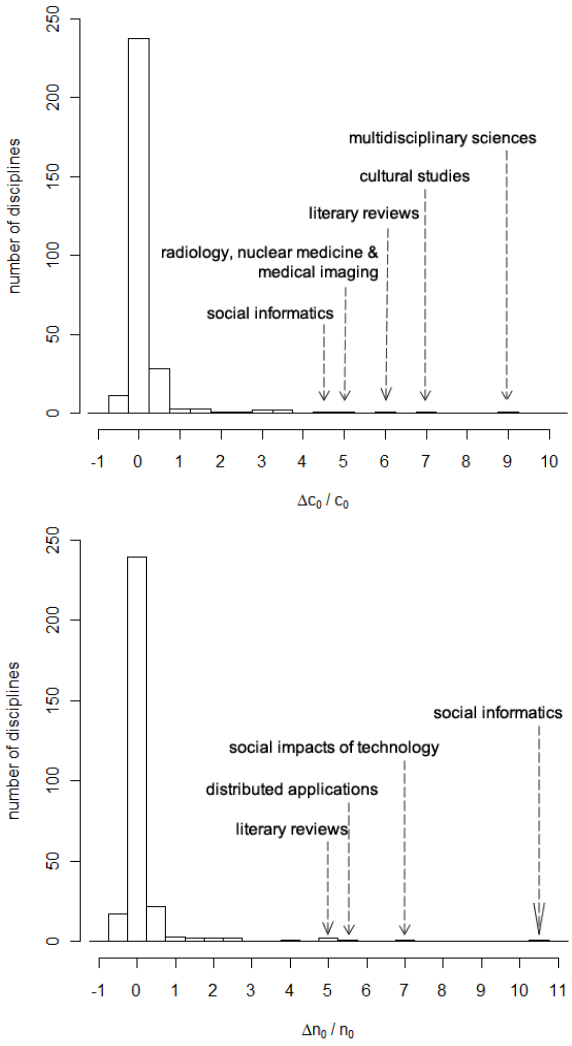

Figure 4: Distributions of the relative changes in c_0 and n_0 for all tags based on 2005 papers. Every time that c_0 is updated to a new value c_0' we can compute $\Delta c_0/c_0 = (c_0' - c_0)/c_0$ and analogously for $\Delta n_0/n_0$. The histograms refer to the latest updates for each tag.

Given the dependence of h_f on c_0 and n_0, we wish to see whether these rescaling factors are stable for all disciplines and years. As an example, we track their convergence for a particular year by plotting in Figure 4 the relative change in the values of c_0 and n_0 for all tags in 2005. We analyzed various other years finding similar results. We observe from the histograms that most of the tags have small relative changes, close to zero. This suggests that the values of c_0 and n_0 are quite stable for most of the tags. There are few outliers, labeled in Figure 4, for which the values are still noisy as we do not have sufficient data for them to converge.

4.3 Visualizing the Collaboration Network

One way to explore the quality of the annotations obtained through the crowdsourcing approach employed by the Scholarometer system is to map the interdisciplinary collaborations implicit in the tags. Since an author can be tagged with multiple disciplines, we can interpret such an annotation as an indicator of a link between these disciplines. For example, if many users tag many authors with both "mathematics" and "economics" tags, we can infer that these disciplines are strongly related, even though they belong to different branches of the JCR — science and social sciences, respectively. Figure 5 presents a network that visualizes the relationships between the top 100 tags in Scholarometer, based on the number of articles annotated with each tag. The nodes in the network represent disciplines. Each node's area is proportional to the number of articles in the corresponding discipline, i.e., the total number of articles by authors tagged with that discipline. It is evident from the node sizes that the majority of early Scholarometer adopters come from computing and information science disciplines. Nodes corresponding to JCR categories are colored based on the ISI citation indices: blue for science, red for social sciences, and orange for arts and humanities. User-defined disciplines are represented by gray nodes. We see a predominance of scholarly data in the sciences based on current Scholarometer usage. The presence of large gray nodes suggests a need for disciplinary labels that are not represented in the JCR classification. Edges represent interdisciplinary collaborations, as induced by author annotations. An edge connecting two disciplines has a weight proportional to the total number of articles by authors who are tagged with both disciplines. For each node we selected the 10 incident edges with the largest weights to represent the strongest collaborations. The layout of the network is obtained by Fruchterman and Reingold's force-directed algorithm [13], so that related disciplines are more likely to be near each other.

The network in Figure 5 displays several features of bottom-up semantics emergent from the crowdsourced annotations. For example, the user-defined tag "computer science" is connected with the various computing disciplines from JCR. Indeed there is a clear computing and information science cluster, as well as psychological sciences, social sciences, and engineering clusters. The plausible map of science that results from our relatively small number of annotations and our simple, automatic visualization algorithm suggests that the crowdsourcing framework yields a meaningful classification scheme for authors and their disciplinary interactions.

5. CONCLUSIONS AND FUTURE WORK

In this paper we introduced a Web Science approach to gather scholarly metadata. We presented Scholarometer, a social Web tool that leverages crowdsourced scholarly metadata with many potential applications, such as bibliographic data management, citation analysis, science mapping, and scientific trend tracking. We discussed a browser-based architecture and implementation for the Scholarometer tool, affording platform and source independence while complying with the usage policy of Google Scholar and coping with the noisy nature of the crowdsourced data. The scholarly metadata that we collect will be shared with the research community.

We outlined several citation-based impact measures that are computed by the Scholarometer tool, including the first implementation of the universal h_f index. We showed how these different measures capture various dimensions of scientific output evaluation. We also found that the statistics collected by our social tool make the novel h_f measure reliable, and capable of comparing the impact of authors across disciplinary boundaries. Finally, we found evidence that the crowdsourcing approach can yield a coherent emergent classification of scholarly output. Of course as the crowdsourced database grows, our data for each discipline will become more representative and our measures more reliable. Additional measures can be implemented as well, for instance universal ones based on percentiles [17].

We are currently working on several enhancements of the Scholarometer tool. One important functionality is to enable users to export individual or bulk bibliographic data into formats appropriate for local reference management software (e.g., BIB, RIS, etc.), or for social publication sharing systems (e.g., BibSonomy). Such a service can become an additional incentive for people to use the tool and thus provide annotation and citation data. Another functionality we are considering is to enable cross-checking of bibliographic records against local or external curated reference collections. Finally, we are working on disambiguation algorithms to better deal with the challenges of common author names.

An interactive application combining a visualization of disciplinary networks with lists of high-impact authors would be an extremely useful resource for learners as they begin to explore the scientific world. Collaboration networks could also be exploited and analyzed on the basis of co-authorship. Studies of co-authorship patterns in conjunction with citation patterns might help characterize the structure of disciplines. Moreover, we can look at trends in scientific fields and track the spikes in the popularity of certain disciplines. This will make it possible to see how new disciplines emerge and evolve over time.

6. REFERENCES

[1] Amazon Mechanical Turk. `www.mturk.com/mturk/welcome` accessed 2010.

[2] BibDesk. `bibdesk.sourceforge.net` accessed 2010.

[3] BibSonomy. `www.bibsonomy.org` accessed 2010.

[4] CiteULike. `www.citeulike.org` accessed 2010.

[5] Comparison of reference management software. `en.wikipedia.org/wiki/Comparison_of_reference_management_software` accessed 2010.

[6] Connotea: free online reference management for clinicians and scientists. `connotea.org` accessed 2010.

[7] Google Scholar. `scholar.google.com/intl/en/scholar/help.html` accessed 2010.

[8] Zotero. `zotero.org` accessed 2010.

[9] R. Adler, J. Ewing, and P. Taylor. Citation statistics. Technical report, A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS), 2008.

[10] S. Alonso, F. Cabrerizo, E. Herrera-Viedma, and F. Herrera. h-Index: A review focused in its variants, computation and standardization for different

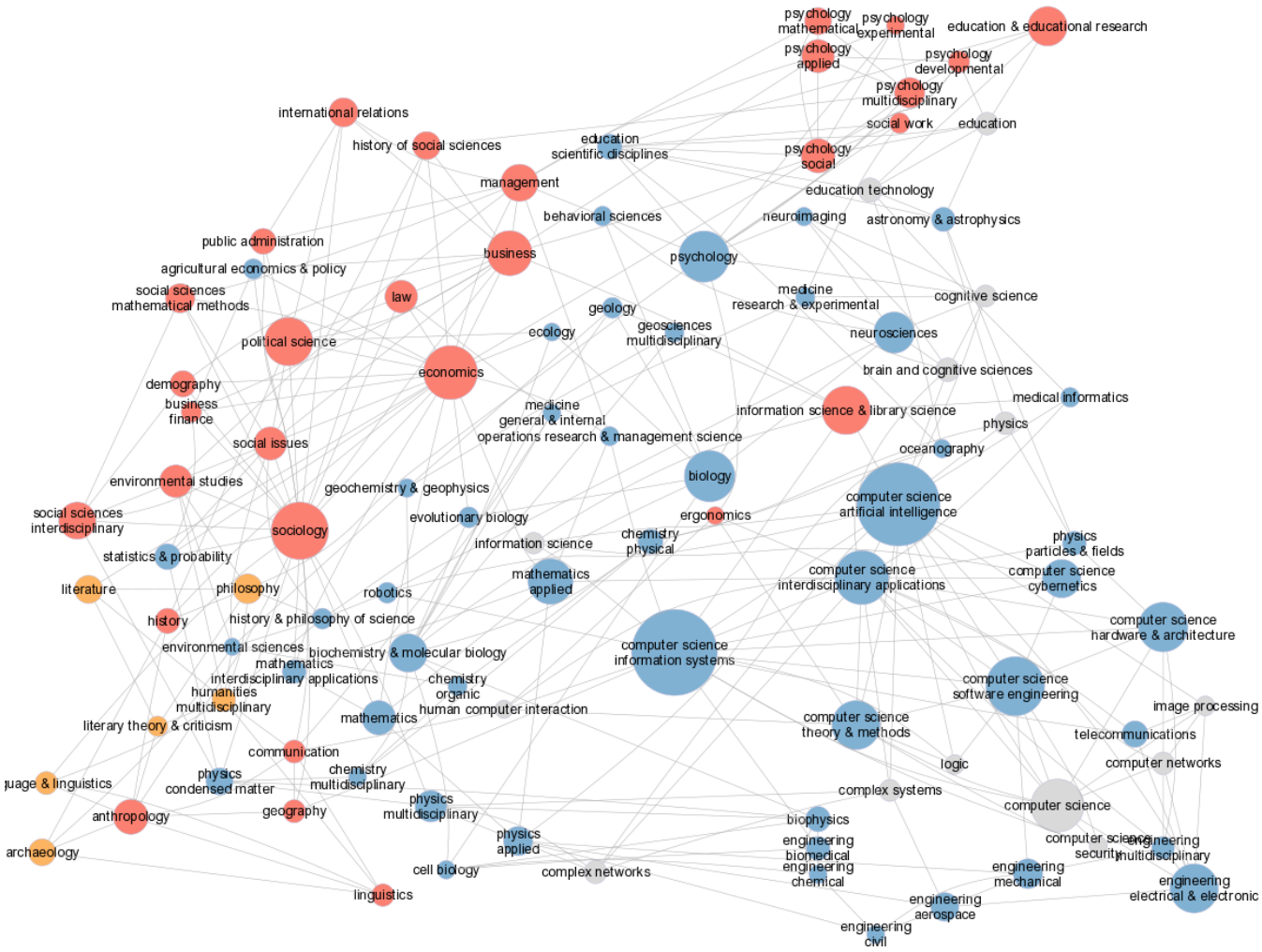

Figure 5: Collaboration network of top 100 disciplines.

scientific fields. *Journal of Informetrics*, 3(4):273–289, 2009.

[11] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[12] J. Feyereisl. Citations-gadget: A Google Scholar Universal Gadget for Scientific Publication Citation Counting. `code.google.com/p/citations-gadget` accessed 2010.

[13] T. Fruchterman and E. Reingold. Graph drawing by force-directed placement. *Software—Practice and Experience*, 21(11):1129–1164, 1991.

[14] A. Harzing. Publish or Perish, version 2.2, 2010. Available at `harzing.com/pop.htm`.

[15] J. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569, 2005.

[16] J. Hirsch. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, 2010.

[17] A. Pudovkin and E. Garfield. Percentile Rank and Author Superiority Indexes for Evaluating Individual Journal Articles and the Author's Overall Citation Performance. In *Proc. Fifth International Conference on Webometrics, Informetrics and Scientometrics (WIS)*, 2009.

[18] F. Radicchi, S. Fortunato, and C. Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268, 2008.

[19] N. Roussel. scHolar index. `interaction.lille.inria.fr/~roussel/projects/scholarindex/index.cgi` accessed 2010.

[20] M. Schreiber. To share the fame in a fair way, hm modifies h for multi-authored manuscripts. *New Journal of Physics*, 10:040201, 2008.

[21] L. Von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.