Modeling Science using Social Media and Web Data

Xin Shuai
School of Informatics and Computing
Indiana University
Bloomington, IN, USA
xshuai@indiana.edu

ABSTRACT

The development of science is inevitably affected by social dynamics since more and more scholarly activities are happening online. With the development of Internet and Web 2.0, the evaluation, spread, and discussion of scholarly information and publications are exposed to the whole society, breaking the strict borderline between scientists and amateurs. This paper proposes to model scientific development by leveraging a large amount of diverse social media data, including Twitter, Wikipedia, and web usage data, from three aspects. First, to investigate how social impact is related to scholarly reputation; Second, to analyze how scientific messages propagate on social web; Third, to study the relationship and dynamics of scientific disciplines using web usage data.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

General Terms

Measurement, Verification

Keywords

Social Media, Bibliometrics, Web 2.0, Scholarly Impact

1. INTRODUCTION

Many scientific activities are now online. Any one can read and download articles from open-access digital libraries; Scientists can talk about their work and discuss with peers through their personal blogs; Twitter or Facebook users can post messages about science news or articles they find interesting, which can further propagate through social networks; Scholars can edit and view Wikipedia pages across different scientific disciplines. These social tools have greatly reduced the information gap between scholars and the public as well as the scholars themselves, gradually turning the development of science into a socialized process.

The influence of social dynamic on the scientific development can be summarized as: (1) the social discussions of scientific publications indicate or affect their scholarly impact; (2) scientific messages can spread much faster to a larger amount of persons than traditional citation through informal social interactions online; (3) collaborations between researchers can be greatly facilitated through social communications, based on which a new structure of science can be drawn.

In summary, the availability of large scale of social media and web data enables us to investigate the influence of social mechanism on scientific development. This paper aims to illustrate the theme in three aspects. First, what's the relation between social attention and scholarly impact, and can the former predict the latter? Second, how the scientific messages propagate on social web and what factors contribute the propagation? Third, how the social interaction pattern can reflect the interdisciplinary communication, and can we use the historical usage data to predict the emerging research areas or trending topics?

2. LITERATURE REVIEW

2.1 Traditional Bibliometrics

Bibliometrics aims to quantitatively analyze scientific activity. The basic method in bibliometrics is "citation" analysis since the creation of Institute for Scientific Information (ISI) Science Citation Index (SCI) by Eugene Garfield [7]. In academia, the action of citing others' articles has two important implications. One, scholars or articles being cited show intellectual merit; Two, scientific ideas spread from the cited articles/scholars to the citing article/scholars. Consequently, the main goals of bibliometrics are two folds. One, to measure the scholarly impact of articles, scholars, journals, institutions, etc (impact analysis), which helps policy makers to allocate research funding. Two, to investigate the relations among different disciplines and the dynamic of scientific development (relational analysis), which provides insight about the disciplinary evolution and communications, especially the emergence of new fields and fading of old fields.

The pure citation count is a natural indicator to measure the scientific impact of an article, journal, or scholar, based on which more advanced indicators have been proposed. *Impact Factor* is used to measure the impact of a journal and calculated as the average number of citations received per paper published in that journal during two preceding years [8]. Journals with high impact factor are considered scholarly influential than those with lower values. *H-index* attempts to measure both the productivity and impact of published articles of a scholar [11]. A scholar with an index of h has published h papers and each of them has been cited at least h times. *H-index* is found to be a good indicator to predict whether a scientist will win Nobel Prize. Both *Impact Factor* and *h-index* are well accepted indica-

tors of scholarly impact, which can be used to rank journals or scholars in different fields.

The scientific ideas propagate through the citation links. Cawkell et al. [6] first analyzed the network flows of citations between articles. Co-citation is a measure of how two articles or journals similar to each other based on their common references. Such measure is used to map the structure of science by clustering journals based on their co-citation or inter-citation similarities [4].

2.2 Web 1.0 Bibliometrics

2.2.1 Webometrics

Webometrics is "the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches." [2]. The emergence of webometrics opens a new direction for bibliometrics, where the evaluation and characterization of scholarly activities and outputs are largely depend on web resources and digital traces.

Thelwall [21] concluded that webometrics mainly include three aspects: link analysis, search engine and web description. Similar to JIF, Ingwersen [12] proposed Web Impact Factor (WIF), which measures the average number of external links per page to a web site. The hypothesis behind WIF is that the number of links targeting an academic web site is proportional to its research productivity and impact, at different level (i.e. university, department, individual, etc.).

2.2.2 Usage Data

The appearance of open access digital libraries makes large-amount of literatures available online. Scholars log on to web portals, view scientific articles and download them. All these usage data is recorded by web servers, constituting a totally different type of scholarly data other than citation records.

Large amount of studies investigated how usage data can be used to study science. Kurtz and Bollen [13] formally defined the usage data model; Bollen et al. [3] generated the map of science using click-stream data. In addition, the relationship between usage data and citation data has been studied, including readership and citation [14], downloads and citation [5].

2.3 Web 2.0 Bibliometrics

Although Web 1.0 enlarges the scope of scientific activities, the data is sometimes difficult to obtain and interpret. In addition, users can only read online articles but not further spread them. Recently, the boom of Web 2.0 makes scientific activities more socialized and public. Any user can post scientific messages they like, and share them with friends. Consequently, some scientific news or articles are supposed to propagate through social networks and reach out to a broader audience in a very fast speed.

Several related work has illustrated the influence of Web 2.0 on the scientific development. Shneiderman [19] forsaw the age of "Science 2.0" when the traditional scientific methods will be revolutionized by social innovation and rapid development of Web 2.0 technology. Ullrich et al. [22] analyzed the pedagogical implications of Web 2.0 on research, teaching and learning. They emphasized that the openness of Web 2.0 technology offers great facility for technology-based learning. Priem et al. proposed the concept of "all-metrics" [16] and "Scientometrics 2.0" [15] with the focus of building new metrics based on frequency of social media mentions to measure the impact of scholarly works in social web.

2.4 Limitations

As is mentioned above, the age of biobliometrics has gone from traditional citation analysis, to web usage analysis until social media analysis today. Looking back the historical studies on biobliometrics, two limitations exist:

One, most bibliometrics studies still heavily rely on citation data from Journal Citation Report (JCR) database, which has already been criticized in two aspects. One, citation data generally suffers from serious temporal lag. The process of publication is very long, from submission, peer-review to finally publishing. To receive citations takes even more time. Therefore, any conclusions drawn from citation analysis actually reflect what happened in the past, not now. Two, JCR database heavily biases towards the journals from Natural Science while largely ignoring journals from Social Science fileds. The majority of the citation records are related to journals from Natural Science, thus most conclusions are mainly about Natural Science but cannot be generalized to Social Science.

Two, the social media data based bibliometrics analysis is still at its infancy stage. Although several studies have attempt to utilize web 1.0 and 2.0 data for bibliometrics analysis, most of them only proposed the framework or conducted some descriptive analysis, without systematic and in-depth analysis. The social media data based analysis is relatively new whose rationale as a supplementary method to authoritative citation analysis needs further verification. What's the relation between social indicator and citation impact? How does the scientific information flow in social web through social connections? How different disciplines are connected to each other through users social actions? How to track the scientific innovations from early web usage signals? All these are very important questions that need systematic study and evaluations.

3. PROPOSED RESEARCH TOPICS

3.1 Collective Attention and Social Impact

Motivation.

Scholars need to decide which articles are worth reading and science foundation agencies need to decide whose works are worth funding. Citation statistics provide accurate but delayed signal about the value of publications. Can we identify those truly valuable works at an early stage by monitoring the social media data? In other words, can we leverage the collective attention from social media platforms to predict or infer the scholarly impact?

Hypothesis.

First, social media responses to newly publications are much faster than academia response in the form of citation. This is because that to mention a scholarly article or author in social media is much easier than to cite an article under peer-review process. Therefore, users' informal discussions about an article generally occur before the accumulation of its citation records. Two, social attention is positively correlated with citation counts. Although social attention and

citation statistics reflect the response from different communities, an excellent publication tends to receive universal acknowledgement. Third, those articles mentioned by social media have higher scholarly value than those are not mentioned. Given thousands of articles published each year, only a small part of them are mentioned and discussed in social media. Social media can serve as a collaborative social filtering system that is able to recommend classic papers and authors. Fourth, we may utilize the social attention to predict the further citations of articles. This is a useful applications after analyzing the relations between social attention and scholarly impact.

Methodology.

Two types of data are used. One is social media data, including Twitter (the largest micro-blogging) and Wikipedia (the largest online encyclopedia); the other is academic papers, including arXiv (the pre-print online publication repository) and ACM (papers published in major journals and conferences in computer science filed). The frequency of mentions about scientific articles in Twitter or Wikipedia is used to quantify the social impact, while the citation count or citation network pagerank is used to quantify scholarly impact.

First, to analyze the temporal characteristic of social attention, including the life time and burst. Second, to calculate the Pearson's correlation or Spearman's rank correlation between social attention and scholarly impact. Third, to utilize linear regression model to predict the citation count using social attention.

3.2 Scientific Information Propagation in Social Web

Motivation.

Today scientific information not only flows with citation links but also propagate on social web via social links. Many studies have investigated the propagation of general information "meme" in social media but few studies ever focused on the scientific "meme". If the first research topic mentioned in Section 3.1 aims to examine the relation between social attention and scholarly impact, here the problem is how social attention grows in social web given the underlying social network topology. Some typical applications include: what types of scholarly publications will arouse much attention in social media? can we identify some influential nodes or useful features to better advertise our scientific works in social media?

Hypothesis.

First, the spread of scientific messages is narrow-ranged and fades out very fast, comparing to general breaking-news topics. "Word-of-mouth" effect tells us that some meme can penetrate through social networks in a very fast speed and reach to a large number of nodes like a virus. However, scientific memes are not supposed to act like general meme, because only those users with scientific interests tend to spread it. Second, the distribution of social attentions among different scientific memes follows power-law. In other words, only very few scientific memes will receive extensive social attentions while most of scientific memes even won't spread at all. Third, a series of features may be extracted to identify those hot scientific memes before its propagation. Such features may include the topical property, topological structure or others. Fourth, some probabilistic propagation models can be proposed to predict the popularity of scientific meme.

Methodology.

Twitter platform is used to study the propagation of scientific memes because it provides "retweeting" functionality to track memes and convenient API to extract the metadata of memes. To collect the data, tweets containing URLs linking to news/publications/blogs from five reputed (open-access) journals (Nature, Science, PNAS, ArXiv and PLoS) are extracted for over one year, using Twitter Search API.

First, to characterize the information cascades of scientific memes propagation. The information cascade is the propagation network of memes induced by retweeting links. The user who first post about some meme is considered the seed node, or the initiator. Any other non-seed users who are directly or indirectly influenced by the seed node and eventually "retweet" the meme are considered "adopters" of the meme. An edge is constructed if the meme propagates from node a to node b. Several important properties of information cascade are size (how many nodes finally adopt the meme), range (how far the meme can spread from the seed node), speed (how fast the meme spreads). Moreover, to identify some frequent structural patterns of cascades (i.e. motif) is also interesting, which may be related to the virality of meme.

Second, to propose propagation models to predict the characteristics and distribution of information cascades. Two types of propagation models are commonly used. One type is analytical model, where the speed of propagation is computed by differential equations. Two classical epidemiological models (SIS and SIR) belong to this type of model and were proposed Anderson and May [1]. The other type is agent based model, where micro-scopic probabilistic rules between individuals are defiend, based on which macro-scopic phenomenon emerges. Two examples include Linear Threshold Model [10] and Independent Cascade Model [9]. The major difference between analytical model and agent-based model is that the latter considers the underlying social network structure upon which propagation occurs while the former not. Both types of models will be applied to simulate the propagation of scientific memes, whose parameters need to be trained on empirical data.

Third, to analyze which factors contribute to the propagation. The above modeling of propagation explains how scientific memes spread but not answer why memes spread like that. Some important features that might affect the propagation include external factor (news media report), internal factor (viral effect through social connections), peer-pressure (peer group influence), strength of social ties (strong tie or weak tie), meme topic (politics or mathematics), interest match (the content similarity between meme and spreading users' profile), etc.

3.3 Disciplinary Communication Characterization through Web Usage Data

Motivation.

As is mentioned above, the two core tasks of bibliometrics are to measure the scholarly impact and to investigate the disciplinary relations. The above two research topics (Section 3.1 and 3.2) focus on the impact analysis while

the third topic aims at relational analysis. Given a specific timestamp, the map of science represents the relations among different scientific disciplines at that time; Given a series of timestamps, the map of science is changing, with single discipline splitting into multiple disciplines or multiple disciplines merging into one new discipline. Such relational and dynamic properties about science can provide insightful information for science policy makers. Several related studies have attempted to characterize the disciplinary relations and evolution through citation network. But few studies ever investigated the same problem on social media data, which is known to provide earlier signal about science development than citation data, therefore may be used to predict the emerging interdisciplinary areas.

Hypothesis.

First, the map of science drawn from citation data and webl usage data should be somewhat different. Citation based map of science biases towards the disciplinary connection between Natural Science (biology, medicine, chemistry, physics) while ignoring the Social Science. By contrast, usage data based map is supposed to provide more balanced disciplinary connection information. Second, the evolution of disciplinary structure including disciplinary merge or split, can be spotted earlier in social media network than citation network. Citation data is generally considered a delayed indicators about scientific dynamic development so we might rely on usage data to capture early signals about emerging or fading research areas.

Methodology.

Two types of data are used to draw the map of science and analyze the evolution of science. The citation data comes from Web of Science in the form of journal-to-journal citation records. The usage data is collected by MESUR[1] project that includes the usage logs for over 10 thousand scientific journals and books.

First, to draw the map of science using both of citation and usage data by visualizing the journal-to-journal relation (citation and usage). Community detection and force-directed layout algorithm will be applied to show the hierarchy of disciplines by clustering similar journals together. The focus is to compare the differences between citation map and usage map, in terms of various quantitative network indicators (including betweenness, centrality, pagerank, etc.), and qualitative structural property.

Second, to show the longitudinal change of science of map over ten years for both citation and usage data and to analyze their temporal correlation. For fair comparison, only Natural Science journals are taken into consideration. The proposed algorithm to track the network change comes from Rosvall and Bergstrom [18], who combined the community detection and bootstrap resampling to map the disciplinary evolution in citation network and successfully spotted the emerge of neuroscience.

4. PRELIMINARY RESULTS

A small part of work regarding to proposed three research topics has been done. Figure 1 shows the dynamic of collective attention towards a set of selected arXiv preprints from Twitter and arXiv server [20]. As shown, Twitter mentions spike shortly after submission and publication, and wane quickly with very mentions after the initial burst. ArXiv downloads peak shortly afterwards but continue to exhibit significant activity many weeks later.

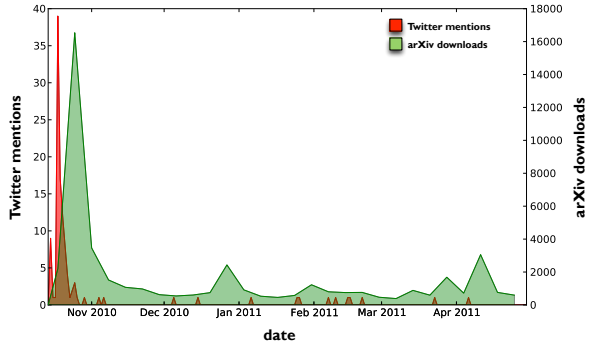


Figure 1: Response dynamics (Twitter mentions and arXiv downloads) for a selected arXiv preprint.

Figure 2 shows the retweeting graph for all arXiv preprints over ten months based on %5 public tweets. The node size is proportional to out degree and the node color and shape represent different types of fields. The structure of retweeting network mainly consists of many star-shpaed modules with the "media" type of node in the center. Some small stars are also centered with the "science" type node. The composition of the periphery of star modules are quite mixed, indicating the arXiv pre-prints have spread over different fields.

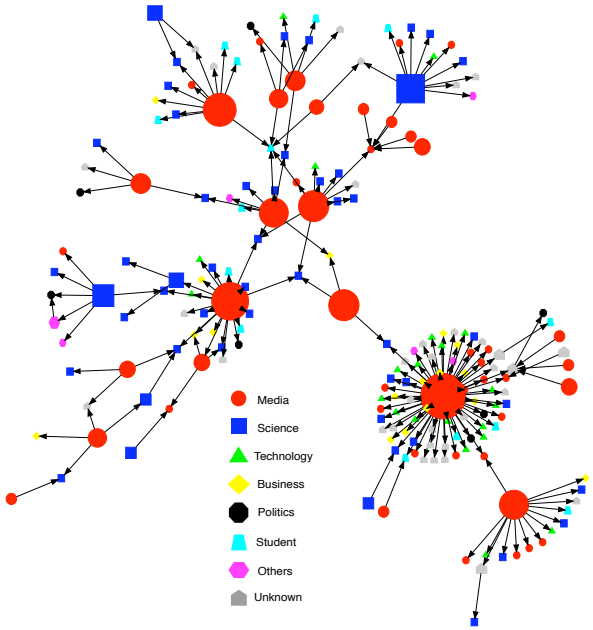
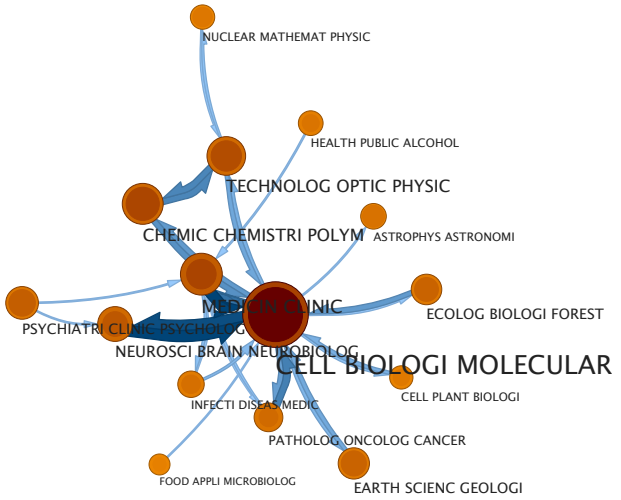


Figure 2: arXiv pre-prints retweeting network

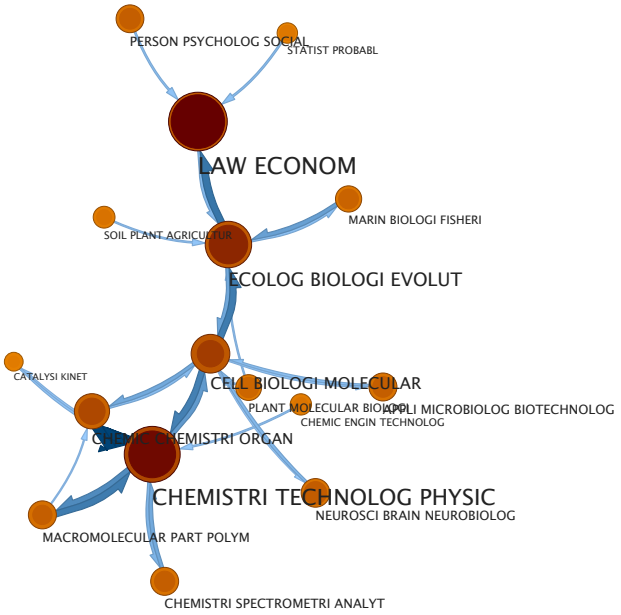
Figure 3(a) and 3(b) show the map of science drawn from journal citation data (2006) and usage data (2007), respectively, after removing most low weighted edges. Informap

[1]http://www.mesur.org/MESUR.html

community detection method [17] is applied to cluster similar journals into big modules and the labels of modules are generated by hand. As expected, citation map severely biases towards the Natural Science, while the usage map exhibits that the biology sever as a bridge to connect Natural science (i.e. chemistry and physics) and Social Science (i.e. law and economy).



(a)



(b)

Figure 3: Map of science drawn from journal (a) citation and (b) usage data.

5. EVALUATION

Three research topics are proposed above in Section 3.1, 3.2 and 3.3 respectively. All of them emphasize to characterize the science development using social and web usage data, compared with traditional citation data. It is a new and interesting direction in bibliometrics, but to evaluate its effectiveness is not easy, since bibliometrics itself has been criticized for lack of systematic evaluation mechanisms for long ages. For the first topic, the numeric value of correlation between social and scholarly impact can be used to quantify their relationship; while for the rest two topics, the evaluation will be mainly based on empirical knowledge and manual interpretation. The details of evaluation methods will be investigated in the future.

6. CONCLUSIONS

This paper proposes to model scientific activities using large-scale of social media data, including Twitter, Wikipedia, and web usage data. Citation data has been considered delayed and biased indicators despite of its authority. Social media and web data provides much faster and more comprehensive signals about the current scientific development. Three research topics are proposed: First, to investigate how social impact is related to scholarly reputation; Second, to analyze how scientific messages propagate on social web; Third, to study the scientific disciplinary relationship and longitudinal development using web usage data. All of them illustrate the advantages of social media and web data as a new indicator to study the science from different aspects.

7. REFERENCES

[1] R. M. Anderson and R. M. May. *Infectious Diseases of Humans Dynamics and Control*. Oxford University Press, 1992.

[2] L. Björneborn and P. Ingwersen. Toward a basic framework for webometrics. *J. Am. Soc. Inf. Sci. Technol.*, 55(14):1216–1227, Dec. 2004.

[3] J. Bollen, H. V. de Sompel, A. Hagberg, L. Bettencourt, R. Chute, M. A. Rodriguez, and L. Balakireva. Clickstream data yields high-resolution maps of science. *PLoS ONE*, 4:e4803, 2009.

[4] K. W. Boyack, K. W. Boyack, A. R. Klavans, and B. K. B. C. Mapping the backbone of science. *Scientometrics*, 64:351–374, 2005.

[5] T. Brody, S. Harnad, and L. Carr. Earlier web usage statistics as predictors of later citation impact: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 57(8):1060–1072, June 2006.

[6] A. Cawkell. Visualizing citation connections. *The Web of knowledge: a festschrift in honor of Eugene Garfield*, pages 177–194, 2001.

[7] E. GARFIELD. Citation indexes for science; a new dimension in documentation through association of ideas. *Science*, 122(3159):108–111, July 1955.

[8] E. Garfield. Citation analysis as a tool in journal evaluation - can be ranked by frequency and impact of citations for science policy studies. *SCIENCE*, 178(4060):471–479, 1972.

[9] J. Goldenberg, B. Libai, and E. Muller. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters*, pages 211–223, Aug. 2001.

[10] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[11] J. Hirsch. An index to quantify an individual's scientific research output. *Proc.Nat.Acad.Sci.*, 46:16569, 2005.

[12] P. Ingwersen. The calculation of web impact factors. 1998.

[13] M. J. Kurtz and J. Bollen. Usage bibliometrics. *ARIST*, pages 1–64, 2010.

[14] M. J. Kurtz, G. Eichhorn, A. Accomazzi, C. S. Grant, M. Demleitner, S. S. Murray, N. Martimbeau, and B. Elwell. The Bibliometric Properties of Article Readership Information. *Journal of the American Society for Information Science and Technology*, 56:111, 2005.

[15] J. Priem and B. H. Hemminger. Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, 15(7), July 2010.

[16] J. Priem, H. A. Piwowar, and B. M. Hemminger. Altmetrics in the wild: Using social media to explore scholarly impact. *ArXiv e-prints*, Mar. 2012.

[17] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, Jan. 2008.

[18] M. Rosvall and C. T. Bergstrom. Mapping change in large networks. *PLoS ONE*, 5(1):7, 2010.

[19] B. Shneiderman. COMPUTER SCIENCE: Science 2.0. *Science*, 319(5868):1349–1350, 2008.

[20] X. Shuai, A. Pepe, and J. Bollen. How the Scientific Community Reacts to Newly Submitted Preprints: Article Downloads, Twitter Mentions, and Citations. 2012.

[21] M. Thelwall. Bibliometrics to webometrics, 2007.

[22] C. Ullrich, K. Borau, H. Luo, X. Tan, L. Shen, and R. Shen. Why web 2.0 is good for learning and for research: principles and prototypes. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 705–714, New York, NY, USA, 2008. ACM.