Mendeley – A Last.fm for Research?

Victor Henning
Bauhaus-University of Weimar, Germany
Mendeley Ltd., London, UK
victor.henning@mendeley.com

Jan Reichelt
University of Cologne, Germany
Mendeley Ltd., London, UK
jan.reichelt@mendeley.com

Abstract

This paper aims to explore how the principles of a well-known Web 2.0 service, the world's largest social music service "Last.fm" (www.last.fm), can be applied to research, which potential it could have in the world of research (e.g. an open and interdisciplinary database, usage-based reputation metrics, and collaborative filtering) and which challenges such a model would face in academia. A real-world application of these principles, "Mendeley" (www.mendeley.com), will be demoed at the IEEE e-Science Conference 2008.

1. Introduction

Ways of how to turn Web 2.0 applications into productive social research tools are currently being discussed at major academic conferences (e.g. European Science Open Forum 2008, Science Blogging Conference 2008, Science in the 21st Century Conference), and both multi-purpose social software, such as wikis, blogs, and social networks, and more specific services such as Twitter, Friendfeed, or CiteULike are currently being used and evaluated by a number researchers and academics. This paper aims to explore how the principles of a well-known Web 2.0 service, "Last.fm", can be applied to the domain of academic research.

2. How Last.fm works

Last.fm (www.last.fm), which bills itself as a "social music service", has managed to create the largest ontological classification (and the largest open database) of music in the world, by aggregating the musical tastes of its 20 million users and then data-mining it for similar musical genres, artists, and songs. The users form a social network that is not based on pre-existing real-world relationships; instead, Last.fm's network emerges around data that describes its users' listening behavior

and musical preferences. The data is gathered as follows: Last.fm's "Audioscrobbler" desktop software, after having been installed on a user's PC, starts tracking a user's music listening behavior. The listening data is sent to the Last.fm website, where a profile of the user's musical tastes is created. Listening statistics for each song, album, artist, and genre are aggregated and made available online. In this way, Last.fm has created the world's largest open music database, comprising over 80 million songs, accessible by everyone. The user-generated data also lays the foundation for personalization, collaborative filtering, and ontological classifications:

- Users can view timelines and statistics about their own listening behavior,
- view the most popular tracks for each of their favorite artists, and most popular artists for their favorite genre,
- receive music recommendations based on the song library already existing on their PC, and
- discover similar tracks/artists for every track/artist in the Last.fm database.

3. The model of Last.fm applied to research

Last.fm's service is based on aggregating the users' existing *music libraries*, relationships between *artists* writing *songs* in different *genres*, and the users' *music listening behavior*. Similarly, a service for academic researchers could be based on aggregating scholars' existing *research paper libraries*, relationships between *researchers* writing *papers* in different *disciplines*, and the scholars' *paper reading behavior*.

Along these lines, a "Last.fm for research" would be able to display statistics to each individual user about his personal library, to aggregate readership statistics about papers, authors, journals, and academic disciplines, and to recommend interesting articles and researchers to the user. We envision that such a tool consists of two parts: First, a desktop application which helps researchers

IEEE
computer
society

manage their academic papers and anonymously tracks their reading habits and literature usage. Second, a website where the users can discover aggregated statistics, top papers, trends and charts for each discipline, paper recommendations, and introductions to people with similar research interests. Adoption of such a service would have a number of advantages for academia at large, of which we will discuss three important ones.

The creation of an open and interdisciplinary database: Similar to Last.fm's efforts in the space of music, a tool which aggregates metadata, tags and article usage of a large number of researchers could lead to an open, interdisciplinary and ontological database of research, providing a free and invaluable source of information to every individual researcher. Working in conjunction with Open Access libraries, this would be another cornerstone in building alternatives to expensive pay-walled databases.

Usage-based reputation metrics and real-time statistics: Usage-based reputation metrics, which could be described as "Nielsen ratings for science", would alleviate many of these problems associated with traditional citation-based reputation metrics. A starting point for usage-based metrics would be to track the pervasiveness of research papers, i.e. whether they are present on the computers of a wide-ranging, distributed sample of academics. This would be a measure of the popularity or awareness that a paper – and by association, its author, publication journal, and topic – is enjoying. A second, more fine-grained usage metric would be the actual time spent reading each research paper (on screen, e.g. in Adobe Reader), and the number of repeat readings per paper. This would be a measure of the intensity with which the paper (its author, publication journal, topic) is being examined – did readers only skim through a paper, or did they peruse it in detail? Finally, these metrics could be augmented with quality ratings and tags that help differentiate mere measures of attention from explicit quality judgments.

Collaborative filtering: Usage data would also be the basis for developing paper recommendation engines based on collaborative filtering (CF) principles. It has been argued that citations already act as reading recommendations, and recommendation models based on co-citations or citation networks ("localized graph search") already exist. CF recommendations, however, have additional potential. First, they could increase the interdisciplinarity of research because they may be better in uncovering parallels between academic disciplines than citation networks. For example, when doing research on the psychology of emotion, a psychologist would typically also have read papers on emotion published in adjacent fields (philosophy, literature,

linguistics, neurophysiology). However, due to space constraints, he would mostly limit his citations to papers published in psychology journals. Picking up such patterns, CF recommendations would thus help researchers to discover literature that could be of interest to them, even though it is not found in the citation network of their existing library. Moreover, CF would enable researchers to identify people with similar research interests (based on their paper libraries) and thus foster collaboration and academic networking. Finally, CF would start generating a rich network of relationships for a paper as soon as it is published, rather than having to wait months or years to get cited.

Of course, there are a number of obstacles to be overcome before such a model can be turned into reality in the field of academic research. Arguably the biggest obstacle is that a sufficient number of participants is needed to gather reliable usage data. So how could scholars be convinced of taking part in generating research paper usage data? In our opinion, the answer is that doing so must confer some type of utility to them beyond the idea of contributing to a fuzzy "greater good of science". More specifically, the tool that does the measuring on the researchers' computers should do this only as a secondary purpose, and must have some other primary usage value. Moreover, whereas Last.fm openly displays each user's listening behavior, privacy is critical in the space of research. While it isn't much of a technical issue to hide a researcher's library and reading data, a "Last.fm for research" would have to convince potential users that it can be trusted with such sensitive data, and that no personally identifiable data would ever be made public without the researcher's explicit consent

4. Mendeley – A Last.fm for research

Mendeley is an application of Last.fm's principles to the domain of research. Mendeley Desktop, a free and cross-platform desktop application, automatically extracts metadata, full-text and cited references from research papers to minimize manual data input when setting up a local research paper database. It then enables researchers to manage, tag, full-text search, cite in Word and LaTeX, and share research papers, thus providing researchers with usage value independent of any network effects.
The companion website, Mendeley Web, can be used for backing up research papers, creating a public research profile, and connecting to like-minded researchers. Mendeley Web already displays the pervasiveness of research papers, authors, journals and tags as measured by Mendeley Desktop. Reading time and quality rating metrics as well as CF recommendation mechanisms will be implemented soon.