Mendeley's Open Data for Science and Learning: A Reply to the DataTEL Challenge

Dr. Kris Jack*, Maya Hristakeva, Rosario García de Zúñiga

Mendeley Ltd.,
144a Clerkenwell Road,
London, EC1R 5DF, United Kingdom
E-mail: {kris.jack; maya.hristakeva; rosario}@mendeley.com
*Corresponding author

Prof. Michael Granitzer

Chair of Mediainformatics,
University of Passau,
Passau, Germany
E-mail: michael.granitzer@uni-passau.de

Abstract: In this paper, two open access data sources, provided by Mendeley Ltd., are described in detail. The first data source is a snapshot of 50,000 user libraries, enabling researchers to test collaborative filtering algorithms for scientific paper recommendation. The second data source is Mendeley's API that provides access to 50 million research articles, crowdsourced from over 1.5 million users. This API allows researchers to build third party applications on top of real-time and large scale data. By providing data sources for the study of collaborative filtering, metadata extraction, deduplication and related research, Mendeley hopes to enable researchers in these domains to better collaborate and share knowledge, ultimately encouraging good research practices and advancing the state of scientific knowledge.

Keywords: Digital Libraries; Recommender Systems; Data sets; DataTEL; Mendeley

Reference to this paper should be made as follows: Jack, K., Hristakeva, M., Garcia de Zúñiga, R. and Granitzer, M. (xxxx) 'Mendeley's Open Data for Science and Learning: A Reply to the DataTEL Challenge', *Int. J. of Technology Enhanced Learning*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Kris Jack is Mendeley's Chief Scientist and leads the development of its data mining tools.
Maya Hristakeva is the data mining engineer responsible for Mendeley's content framework.
Rosario García de Zúñiga is lead software engineer for Mendeley's API.
Michael Granitzer is Professor for Mediainformatics at the University of Passau.

1 Introduction

Recommender systems have been successfully employed in both commercial and non-commercial environments in a variety of domains since the appearance of the first papers on the subject in the mid-1990s (Resnick et al. 1994). Recommender systems normally follow either a collaborative filtering, content-based or hybrid approach to recommending relevant content to users (Candillier et al. 2009). From Amazon's early use of collaborative filtering (Linden et al. 2003), to more recent start-up company successes, such as LastFM (Haupt 2009), recommender systems have played a pivotal role. Unlike traditional search systems where users pull data, recommendation systems push data to users. The strategy of pushing data is well adapted to the use case of a user who wants to keep up-to-date with relevant changes but does not have the time to survey them all. This use case is recognisable to researchers who want to stay abreast with recent developments in their field, but find it hard to survey every advance due to the high volume of research that is published.

There is a growing body of work in the investigation of recommendation systems in science and learning. The TechLens Project (Kapoor et al. 2007), for example, takes on the challenge of building a recommendation system for a digital library. They don't only try to generate recommendations for individual researchers, but also for groups of researchers. Some research has also indicated that that articles become easier to recommend to researchers when they are tagged (Parra-Santander & Brusilovsky 2009). More recently, Wang & Blei (2011) show that a collaborative topic regression model, a hybrid approach, can outperform collaborative filtering and content-based methods alone in recommending articles based on Cite-U-Like data. These results demonstrate the utility of recommendations in this domain.

In 2010, the DataTEL Challenge was proposed in order to address the problem of a shortage of data sets in this field (Drachsler et al. 2010, Verbert et al. 2011). In the challenge, contributors were asked to put forward data sets that would serve as test data for recommendation systems in science and learning. Mendeley replied positively to the challenge.

Mendeley is a research platform that produces tools with the aims of helping users to organise their research, collaborate with colleagues and discover new knowledge (Henning & Reichelt 2008). In doing so, Mendeley warehouses and analyses a large volume of data. As of January, 2012, Mendeley's user base has grown to over 1.5 million researchers who have, in total, contributed over 150 million research articles, since being launched three years previously. Mendeley has crowdsourced the world's largest scientific research library, containing 50 million unique articles, 10 million more than Reuter's Web of Science. More generally, Mendeley aims to tackle the problem of scientific knowledge being siloed away on individual scientists' computers by bringing it together on a single,

standardised platform, where collaboration and discovery are made easier through the appropriate use of Science 2.0 technologies.

This paper focusses on the data that Mendeley is making public in an attempt to encourage good research practices in an open environment of sharing and collaboration. To these ends, Mendeley sampled a data set of 50,000 user libraries for the DataTEL Challenge. These libraries were generated primarily through users adding them to Mendeley's reference management tool, Mendeley Desktop. This is the first data set that gives access to Mendeley user libraries. Mendeley's DataTEL data set can be seen as an additional contribution to data sets generated for scientific literature, such as CiteSeer (`http://csxstatic.ist.psu.edu/about/data`). The data set has been made anonymous to protect user privacy by removing all article metadata. This makes the data set appropriate for testing collaborative filtering algorithms.

Mendeley also exposes data through an API. This API primarily provides programmatic access to Mendeley's research catalogue. It also provides access to a set of groups that link articles and users together, to form a public source of data from which recommendations can be made. Unlike the snapshot of 50,000 user libraries, the data provided through the API is current and updated in real-time.

In section 2, the tools that Mendeley provides are described in some detail in order to put the data sources in context. Mendeley contributes to the advancement of scientific recommender systems by providing a data set for the study of collaborative filtering (section 3) and also opening up its real-time data stores through an API (section 4). A brief description of supplementary data sets that Mendeley has and will provide is then made (section 5) before drawing conclusions (section 6).

2 What Is Mendeley?

2.1 Introduction

Mendeley provides digital tools with the intent of helping users in performing research activities. In this section, two of Mendeley's most popular tools are described. These tools provide the core interfaces for users to add data to Mendeley, so understanding them helps to contextualise the data sets being provided. These tools are Mendeley Desktop and Mendeley Web. A selection of the data gathered by Mendeley Desktop is put forward for the DataTEL challenge while the data offered by Mendeley Web is accessible through Mendeley's Developer API.

2.2 Mendeley Desktop

Mendeley Desktop is primarily a reference management application. Its user base stands at around 1.5 million users, as of January, 2012. It also provides support for annotating research papers and collaborating with other researchers. These functionalities are available across platforms and with cloud synchronisation support.

2.2.1 PDF and Reference Management

Mendeley Desktop helps users to organise their research papers into personal libraries (i.e. collections of papers). Users can add papers to their libraries by dragging and dropping a file into their library space (e.g. pdf), importing a file from disk, importing references from a structured format (e.g. bibtex) or adding a new entry manually. To help with the laborious task of entering the reference metadata for a research article (e.g. title, authors, year), Mendeley Desktop automatically extracts it out of the pdf using Support Vector Machines (Granitzer, Hristakeva, Knight & Jack 2012). Users can organise their articles into folder structures, following the computer file system analogy, and tag them for future reference and retrieval. Mendeley Desktop provides retrieval methods by querying over article metadata and full text. The user can open an article and view it within a built in pdf viewer. This action triggers an event indicating to the user, through a graphic next to the article, that it has been opened, thus helping users to keep track of which files they have already read in their library. Similarly, users can choose to *star* an article by clicking on a star to the left of it. This makes them easy to distinguish from non-starred articles in a collection.

2.2.2 Annotating Research Papers

Mendeley Desktop replaces the need to annotate hard copies of articles with pens and highlighters by providing digitalised annotation tools (Figure 1). Once an article is open within Mendeley's pdf viewer, users can annotate it in four different ways. First, there is a digitalised highlighter that can highlight either text or rectangular areas of the paper. On first application of the highlighter, the text is covered with a light yellow background, making it stand out on the page, just as highlighted text does on paper. Further applications of the highlighter on the same text deepens the colour of the highlight, allowing users to apply different meaning to different depth of highlighted text (e.g. the deeper the colour, the more important the text). Second, users can add notes to a paper using digitalised sticky notes. That is, the user can select a sticky note and add it to any location of the pdf document. The sticky note has a free text entry field allowing users to enter a description. The time and date that the note is added is recorded and displayed to the user for future reference. Third, users can add notes to a paper by entering them into a free text entry box. Users can enter multiple comments on a paper, without being restricted by the physical space that hardcopy margins impose. Finally, users can also tag their papers by entering tags into the article's metadata. Unlike the title and year of publication of an article, however, tags are a personalised form of metadata that can change from user to user.

2.2.3 Collaborative Features

In addition to reference management, Mendeley Desktop also supports collaborative research functionalities. For example, it is common place for researchers to exchange materials while investigating a topic (e.g. notes on a paper). Mendeley Desktop has built in e-mail functionality allowing users to send articles to their Mendeley contacts. Users can also create Mendeley Groups. A Mendeley Group brings together a collection of articles and users. It supports users

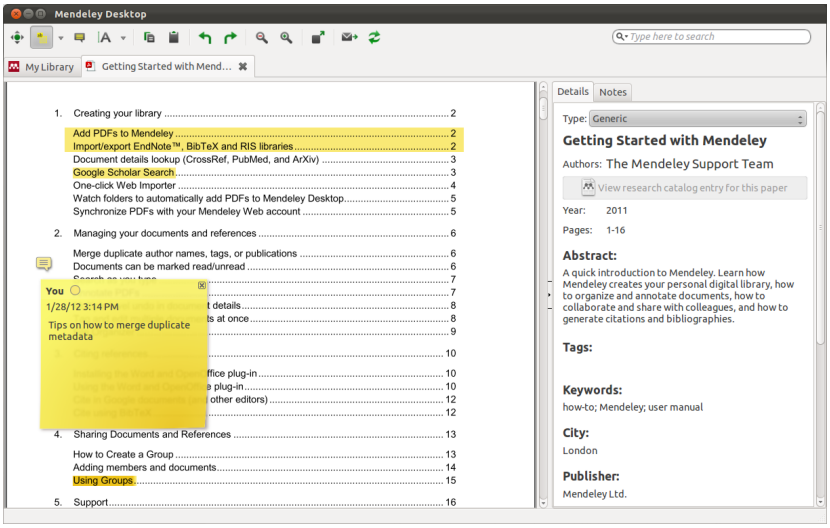

Figure 1 Screenshot of Mendeley Desktop with pdf open and being annotated with highlights and sticky notes. The editable metadata for the article appears to the right.

by providing a forum for discussing articles and commenting on the activity in the group through real-time notification feeds. Groups have been used for a variety of use cases including, amongst others, bringing together articles on a similar topic for focussed discussion, collaborating on a new research paper with a group of colleagues and creating a reading list for a class of students. Such collaborative functions build on Mendeley's social network that connects researchers to one another.

2.2.4 Cross-platform Synchronisation

User libraries are accessible from multiple locations and across platforms. The user's library, along with their annotations, folder structure, contacts and personal profile details are all uploaded to the cloud where they are kept synchronised for the user to access them from different locations. Mendeley Desktop is also cross platform, being available on Windows, Mac and Linux distributions. Additionally, the user can access their library online through the Mendeley Web Portal, which is the subject of the next section.

2.3 Mendeley Web

Similar to Mendeley Desktop, Mendeley Web also provides users with access to their personal libraries and data. While this data is private to individual users, Mendeley Web also provides an interface to public data, such as public groups (i.e. groups that users have made publicly accessible), Mendeley's userbase (i.e. a repository of user profiles) and a research catalogue (Figure 2). This is the largest collection of research articles in the world. Mendeley crowdsources the article data entered through Mendeley Desktop, aggregates and anonymises the content and generates the research catalogue.

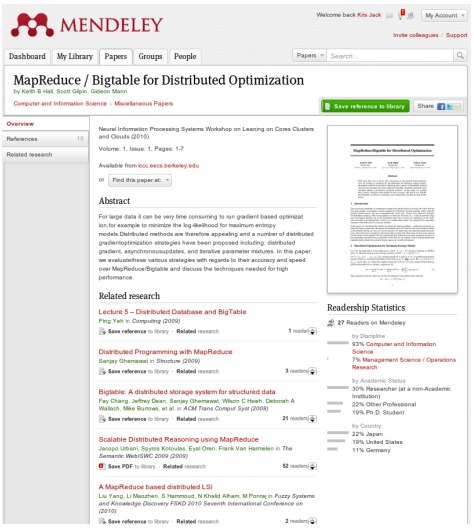

Figure 2 Screenshot of an article in Mendeley's research catalogue.

When the user adds an article to their library, Mendeley Desktop attempts to identify it and to match it with an entry in its catalogue, a near exact deduplication problem (Hammerton et al. 2012). If a match is found, then the data in the article contributes to the matching article page in the catalogue. If it is not found, then a new entry is created. All articles are described as a collection of metadata, although variations in metadata types exist across different article types (e.g. a conference paper may have a proceeding's title, whereas a book would not). The following metadata entries can be applied to almost all articles types: title; authors; year; discipline; publication venue; author supplied keywords; abstract and references. Articles in Mendeley's catalogue also have crowdsourced metadata. In particular, an article can have a set of tags that were applied to it by users. Only tags that were applied by three or more different users are shown on article pages. Statistics are calculated for articles describing its readership in the Mendeley community. For each article, the total number of Mendeley users who have it in their library is counted. Based on these counts, demographic distributions of readers are derived using academic discipline (e.g. biology, computer science), academic status (e.g. Masters Student, PhD, Professor) and country. All of these statistics are available for articles in the Mendeley catalogue. Pdf previews for articles are also generated to let users browse through the article online before deciding whether to add it to their library or not. Only the first two pages of articles are shown unless they are Open Access or Mendeley has the copyright holder's permission to show them in full. Finally, Mendeley links articles together base on their relatedness, providing related research for users to browse through, discovering new and relevant literature. In section 4, a range of applications are presented that have been built on this data, exposed by the API. Before that, the DataTEL data set is described in detail.

3 DataTEL Data Set

3.1 Overview

In order to help users test recommendation system algorithms in the domain of science, Mendeley has constructed a data set. It is comprised of a random sample of 50,000 user libraries that contained at least 20 articles. All user and article ids have been anonymised for privacy reasons. Similarly, no metadata for articles is provided for reasons of privacy. This makes the data set appropriate for testing collaborative filtering algorithms.

Mendeley's data set provides information on user libraries in three files. One file includes the set of articles that appear in user libraries, while the other two provide usage-based information: one of them showing which articles users have read using Mendeley Desktop; and the other showing which articles users have marked with stars using Mendeley Desktop. The structure of the three files are first described before presenting general characteristics of the data set.

3.2 User Libraries

Researchers use Mendeley Desktop and Mendeley Web to add scientific articles to their libraries. A selection of these libraries were randomly selected and entered into the data set (Table 1). The file has 50,000 user libraries that contain a total of 4,848,724 articles, 3,652,285 of them being unique. All user libraries contain at least 20 articles.

Table 1 User Library Data Schema

Element	*Description*
Schema ID	Mendeley.com user libraries
No. columns	2
Column 1	User id (string id that uniquely identifies a user)
Column 2	Article id (string id that uniquely identifies an article)

3.3 Library Readership

The second data file provides readership information for researchers and their articles (Table 2). Using Mendeley Desktop, users can open up their articles and read them. When read, the application indicates to the user that the article has been read. This file includes the readership data for the same articles presented in the first file and indicates whether the user has used Mendeley Desktop to read them or not. 1,466,489 of the articles that appear in libraries, or 30%, have been read using Mendeley Desktop.

Table 2 Library Readership

Element	*Description*
Schema ID	Mendeley.com library readership
No. columns	3
Column 1	User id (string id that uniquely identifies a user)
Column 2	Article id (string id that uniquely identifies an article)
Column 3	Read status (int that is 1 if the article has been read and 0 if it has not been read using Mendeley Desktop)

3.4 Library Starring

Researchers can also make use of Mendeley Desktop to star articles that are in their libraries. This starring information is included in the third and final file, the Library Starring table (see Table 2). In the file, 615,308 of the 4,848,724 articles library entries (13%) have been starred by users. Mendeley does not put any requirements on why users should star articles. As a result, users may star articles for different reasons, making the action semantically ambiguous.

Table 3 Library Starring

Element	*Description*
Schema ID	Mendeley.com library starring
No. columns	3
Column 1	User id (string id that uniquely identifies a user)
Column 2	Article id (string id that uniquely identifies an article)
Column 3	Star status (int that is 1 if the article has been starred and 0 if it has not been starred using Mendeley Desktop)

3.5 General Data Set Characteristics

The user libraries vary in size from 20-300 articles (see Figure 3), in which library sizes tend to be on the lower end of that range. The plot shows that around half of the libraries contain articles that have been starred or read, with the starring feature being used more often than the pdf reader. Internal experiments in Mendeley have shown that users begin to receive reasonable recommendations (e.g. 0.1 precision at 10) under 10-fold cross-validation when they have at least 20 articles in their library. Given the sizes of the user libraries, this data set is not appropriate for investigating the cold start problem as it can take some time for new users to add 20 or more articles.

A small number of articles are shared by a large number of users, but most articles do not appear in more than 10 user libraries (see Figure 4). These figures

are representative of user libraries in Mendeley's full data set making them a good sample for testing. The plots show that the data set is very sparse as it has few user item co-occurrences. This presents a challenge for collaborative filtering algorithms that typically perform better when the user-item space becomes more dense. For Mendeley's use case, it is desirable to explain to the user why they receive particular recommendations (e.g. "paper x is recommended based on papers y and z in your library"). Item-based algorithms can provide this quite easily. Unfortunately, given that there are far recommendable items than users, an item-based approach is typically not recommended. Finally, whatever algorithm is used by Mendeley, it needs to scale elegantly to data sets with tens of millions of users and items. Proposed solutions to take these requirements into account.

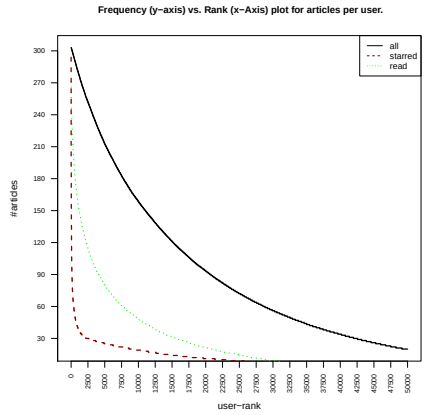

Figure 3 Frequency vs. rank plot for articles per user.

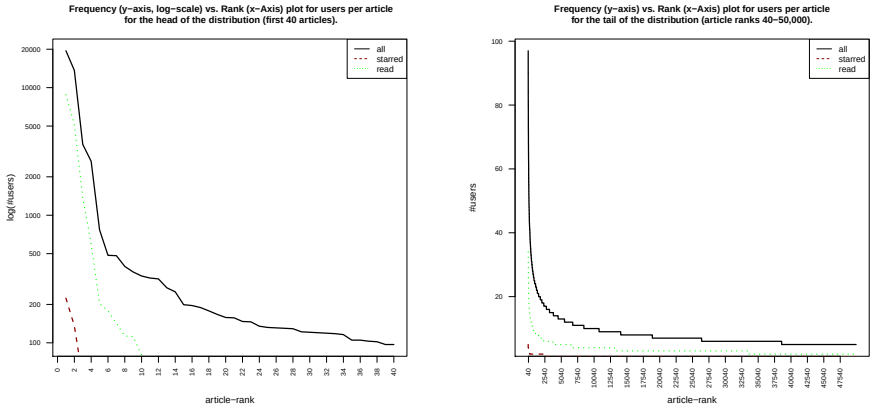

Figure 4 Frequency vs. rank plots for users per article.

3.6 The Scientific Community's Response to the Data Set

As of January, 2012, Mendeley has received over 100 requests for the DataTEL data set, granting access to over 200 researchers. The majority of correspondents identified themselves as being PhD candidates and post-doctoral researchers. The most frequent question that is posed by far concerns granting access to the metadata contents of the articles in the data set. In particular, requests for article titles, authors, tags and timestamps are regularly made. The current dataset does not include such metadata. Instead, each article is represented by a universally unique identifier that has no semantic meaning. This is a good data set for testing algorithms such as collaborative filtering but not appropriate for content-based methods. Testing content-based methods motivates the majority of these requests. While Mendeley recognises that by not providing metadata content, it limits the usefulness of the data, there is no intention to do so for fear of breaching user privacy. Netflix, for example, had to revoke their data set after successful attempts were made to decrypt private data (Narayanan & Shmatikov 2008). Mendeley does not want to take this risk and finds that the existing data set is already useful for its purposes. Respecting user privacy is of primary importance to Mendeley. If Mendeley were to include the metadata in the data set then users may be identified, which is not the aim of this exercise. While Mendeley has no intention of releasing such data under these circumstances, it is willing to contribute rich sets of metadata both through the Mendeley Developer API and for challenges that are based on publicly accessible data (see section 5). Also, it is possible that Mendeley will provide the functionality to allow users to make the contents of their libraries public and contribute to the creation of future data sets.

3.7 Obtaining the Data

Mendeley's data set is available for download from the Mendeley Developer Portal (http://dev.mendeley.com/). To obtain it, interested parties should write to datachallenge@mendeley.com and include the following information: names of the individuals who will be working with the dataset; institutional affiliation; contact address; and contact phone number.

Mendeley may contact developers if changes are required to be made to the data set. Mendeley's data set is being provided for non-commercial scientific use only. All interested parties must agree to a creative commons license for non-commercial use of the data. Authors must cite this paper when making use of the data set. Any queries regarding the data set should be sent to datachallenge@mendeley.com.

4 Mendeley's Developer Portal

Mendeley's Developer Portal hosts an API that provides developers with access to tools that can retrieve and manipulate their personal data plus aggregated Mendeley content, such as the Mendeley research catalogue. In this section, the API is described, with examples of the methods that it exposes, before describing a selection of applications that have been built on the platform. Note that the

article ids provided through the API do not correspond to the ids in the DataTEL data set.

4.1 The API

In April, 2010, a call for proposals was publicly released to challenge developers to make use of Mendeley's API (`http://www.mendeley.com/blog/press-release/announcing-mendeley-open-api/`). It provides access to the crowd-sourced and aggregated data in Mendeley's research catalogue, through public and private methods. The public methods grant access to data that can be seen by browsing Mendeley web (see Table 4), whereas the private methods grant access to user-specific private data that require authentication (see Table 5). On registering an application, developers receive a consumer key and secret that enables them to authenticate their requests through OAuth, an open protocol to allow secure API authorisation (`http://oauth.net/`). Private methods are currently not rate limited. Public methods, however, are limited to 150 requests per hour, with the exception of 5,000 calls to search methods per hour. All data is licensed under a Creative Commons Attribution 3.0 Unported License for non-commercial derivations.

Table 4 Public Methods

Method Type	*Examples*
Top Catalogue Statistics	top authors by discipline
	top papers by discipline
	top publications
	most frequently applied tags per discipline
Search Tools	results from a catalogue search query
	papers written by a given author
Related Articles	return articles that are related
Article Metadata	title
	authors
	publication year
Groups Manipulation Tools	details
	articles
	users
	create new public group

Table 5 Private Methods

Method Type	*Examples*
Top User Library Statistics	top publication outlets
	top authors
	top tags
Library Manipulation Tools	create an article
	retrieve article metadata
	upload a file
	download a file
	retrieve library contents
Group Manipulation Tools	create a group
	delete a group
	retrieve articles
	retrieve users
Folder Manipulation Tools	retrieve folder details
	create a folder
	delete a folder
	add an article to a folder
	delete an article from a folder
Social Network	retrieve contacts
	add a contact
Profiles Tools	retrieve contacts for a user
	add a contact
	retrieve profile details

4.2 Innovative Mendeley API Applications

Mendeley's registered API developer community numbers in the thousands. Mendeley ran a competition, in collaboration with PLoS called the Binary Battle, with the challenge of building an application using Mendeley or PLoS's APIs in order to make science more open. In this section, a selection of applications that were entered for the Binary Battle are described to demonstrate the kinds of applications that can be built using Mendeley data.

openSNP (`http://opensnp.org/`), pronounced open snip, is an application that enables large scale crowd-sourcing of genome wide association studies to create new knowledge about genes. It allows customers of direct-to-customer genetic tests to publish their test results, find others who have similar variations and find the latest primary literature on their variations using Mendeley's API.

Paper Critic (`http://www.papercritic.com/`) is an application that aims to open up critical reviews of publications. It allows users to select papers, add reviews to them, rate them and to search what other users are saying about them too. This open platform for post-publication reviewing allows users to log in

with their Mendeley accounts, search through the Mendeley catalogue and start reviewing papers.

rOpenSci (`http://ropensci.org/`) is a collaborative project that aims to exploit the power of the programming language R in order to facilitate Open Science. A package was developed named RMendeley that connected R with Mendeley's API, allowing users to search and retrieve data from Mendeley's catalogue. This application equips scientists with tools that enable them to contribute to opening up science. For example, using rOpenSci, developers could investigate the impact factor of authors and articles, like in the following two examples.

Impact factors for researchers can be used to judge the quality of the work that they produce. Reader Meter (`http://readermeter.org/`) provides a selection of impact factor results for researchers by searching through Mendeley's catalogue and deriving statistics for them such as the H_r-Index and G_r-Index based upon the readership of their works. Recent work has shown that readership counts in Mendeley's community can be correlated with impact factor measures (Kraker et al. 2012).

Similarly, Total Impact (`http://total-impact.org`) takes a collection of Mendeley articles and generates a report of their overall impact, based upon statistics gathered from several web sources, including Mendeley. For example, for a given article, Total Impact will show how many Mendeley users are reading it and the number of Mendeley groups in which it appears.

Co-authorship networks can help researchers to better understand a field by seeing who is collaborating with one another. Collabgraph (`https://collabgraph.xcend.de/`) takes a Mendeley library and visualises the connections between the authors of the articles in a graph, connecting two authors together if they have co-authored one or more papers.

Mendeley has produced a lightweight version of Mendeley Desktop that runs on mobile iOS platforms. Using the Mendeley API, an alternative android client, named Droideley (`https://market.android.com/details?id=com.droideley`), was entered into the Binary Battle. Droideley allows users to synchronise their libraries on a mobile device, search through the contents and read articles.

Articles can be related to one another in several ways. For example, an article may support the results found in another article, it may refute them, it may extend them, or it may correct them, to name just a few relationships. Kleenk (`http://kleenk.com/`) takes the articles in a user's Mendeley library and allows the user to specify the relationships between the articles. The links can also be rated and commented upon by other users.

The Mendeley API demonstrably supports a diverse set of applications. Together with the dataTEL data set, these data sources can be used to address a variety of use cases, including those that directly target education. Professors could, for example, make use of these data sources to run practical labs where students are encouraged to test the properties of different algorithms. In line with this case, Mendeley are also constructing and releasing ground truth data sets for investigating other interesting problems. These additional data sets are the subject of the next section.

5 Supplementary Data from Mendeley Case Studies

Mendeley faces a number of technical challenges in building a research platform and tools that operate on top of it. In this section, three case studies relevant to these aims are considered. First, the problem of automatically extracting metadata from pdfs is tackled, followed by the automatic deduplication of articles in the catalogue. Finally, the problem of automatically retrieving related articles is described. In each use case, Mendeley follows the scientific procedure of developing a ground truth data set that represents the desired output of a system before implementing and testing algorithms against it. These data sets are being made publicly available for researchers to test new algorithms against, emphasising the value of sharing data for solving problems.

5.1 Metadata Extraction

When a user adds a pdf article to Mendeley Desktop, the application attempts to automatically extract out its metadata contents (e.g. title, authors, year of publication). Given the wide range of styles in which papers are published, this task is particularly challenging. Previous solutions have shown that the use of support vector machines (Han et al. 2003, 2005) and conditional random fields (Councill et al. 2008) achieve good results on small scale data sets. Granitzer, Hristakeva, Knight, Jack & Kern (2012) constructed an eprints data set from the e-prints RDF Repository and tested these approaches. The results showed that a two-stage support vector machine approach produced the best results. This data set is publicly accessible from `http://team-project.tugraz.at/2011/10/17/metadata-extraction-e-prints-data-set/`. The data set contains PDF, associated metadata and preprocessed BIO annotations (see CoNLL Shared Task 2000, `http://www.cnts.ua.ac.be/conll2000/chunking/`) with fuzzy title matching.

5.2 Article Deduplication

Mendeley's research catalogue is constructed from over 150 million articles, resulting in 50 million unique entries. This means that at least 100 million of the articles are duplicates. Determining which articles are duplicates of one another is a trivial problem in small scale systems where objects are exact matches or have unique identifiers. In large scale systems where the same objects can have different values and no identifiers, this makes for a challenging problem. This is the case for Mendeley, that attempts to determine if two articles, that may differ slightly in their metadata, are duplicates of one another or different entries. This problem is similar to the one of citation matching (Lawrence et al. 1999) and systems that perform record linkage across documents. Mendeley constructed a data set to investigate this problem (Hammerton et al. 2012). This data set acts as a ground truth for testing near-duplicate deduplication problems in the field of scientific references and is publicly available from `http://team-project.tugraz.at/2012/01/13/new-de-duplication-data-set-published/`.

5.3 Related Articles

Mendeley's Related Articles service attempts to automatically find research articles that are related to one another. This is a well known problem in information retrieval and has received much attention in the literature (Gipp et al. 2009, Vellino 2009, Beel & Gipp 2010). A set of three data files are under preparation to investigate this problem. Each data set contains examples of articles that are related to one another in different ways. Unlike the DataTEL dataset, these three datasets will contain the metadata descriptions of the articles (e.g. title, authors, year), allowing researchers to test content-based algorithms.

5.3.1 Author Publications

Mendeley users can use their automatically generated profile pages as CVs to communicate their skills and experiences externally. Publication lists are of particular importance to researchers. This dataset contains a random selection of over 2,500 user publication lists that have been made public, representing over 35,000 unique documents. To reduce noise in the data set, only article lists that share at least one surname are included. As the article sets have between 10 and 20 articles in them, they are likely to contain the publication lists of professional researchers who have been involved in research for a certain number of years.

5.3.2 Groups

One of the social and collaborative features that Mendeley offers is Mendeley Groups. Researchers can create public (open for all to see and join) and private (open only through invitation) groups. They can then add articles to the groups and generate discussion around them. Some users use groups as a collaborative tool when co-authoring a new article together, some use them as reading lists for their students, while others use them to generate discussions around new topics. Regardless of the use case, Mendeley users have now created over 125,000 groups. This dataset contains a random selection of almost 4,500 public article lists that appear in groups, representing over 60,000 unique documents.

5.3.3 Venues

Articles that are published in the same venue can also be described as related to one another. To generate this dataset, the publication venues from Mendeley's articles were normalised and almost 40,000 of them were selected at random. For example, the two venues *1st International Conference of Artificial Intelligence* and the *2nd International Conference of Artificial Intelligence* were normalised to become the *nth International Conference of Artificial Intelligence*.

Through providing these data sets, Mendeley hopes to enable new research in the automatic retrieval of related research articles.

6 Conclusion

Good quality data for testing new algorithms is a valuable resource in the academic community. New fields of study often have to invest time and resources

in the necessary pursuit of developing them. Mendeley has spend the past three years crowdsourcing the world's largest research catalogue. In doing so, it has tackled a number of challenging problems from metadata extraction through to recommending scientific material. A number of rich and interesting data sources have been produced in these endeavours and are being publicly released for scientific collaboration.

Our future work in recommendation will focus on exploiting additional signals and extending the range of recommended items. With the growth of Mendeley's community, social computing paradigms like tagging become more influential for personalising recommendations. Such signals as well as signals obtained from catalouge usage might boost the recommender's accuracy. Also, recommending research articles is just the beginning. Scientific knowledge is contained in entities and relationships between entities. Harmonizing the entities that carry scientific knowledge across different papers and recommending such entities will likely remain a major challenge for many years.

Much of the world's knowledge remains siloed away on individual researchers' computers and while collaboration is recognised as a valuable practice, it remains a more difficult task than it ought to be. By providing a common platform upon which to share knowledge, using modern Science 2.0 technologies, we can create an eco-system in which research can become more accessible and more easily disseminated. In building this platform, we aim to break down both real and artificial barriers between individual researchers and scientific domains, enabling people to work in a more coherent, cohesive and scientifically rigourous manner, and ultimately opening science up to a wider audience.

Acknowledgements

This work has been partially funded by the European Commission as part of the TEAM IAPP project (grant no. 251514) within the FP7 People Programme (Marie Curie). We also thank the editors of this journal and the three anonymous reviewers for their constructive comments.

References

Beel, J. & Gipp, B. (2010), 'Link Analysis in Mind Maps : A New Approach to Determining Document Relatedness', *Mind* (January).

Candillier, L., Jack, K., Fessant, F. & Meyer, F. (2009), State-of-the-Art Recommender Systems, *in* M. Chevalier, C. Julien & C. Soule-Dupuy, eds, 'Collaborative and Social Information Retrieval and Access Techniques for Improved User Modeling', IGI Global, pp. 1–22.

Councill, I. G., Giles, C. L. & Kan, M.-Y. (2008), 'ParsCit: An open-source CRF Reference String Parsing Package'.

Drachsler, H., Bogers, T., Vuorikari, R., Verbert, K., Duval, E., Manouselis, N., Beham, G., Lindstaedt, S., Stern, H., Friedrich, M. & Wolpers, M. (2010), Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning, *in* N. Manouselis, H. Drachsler, K. Verbert & O. Santos, eds, 'Proceedings of Elsevier Procedia Computer Science', Vol. 1, pp. 2849–2858.

Gipp, B., Beel, J. & Hentschel, C. (2009), Scienstein: A research paper recommender system, *in* 'Proceedings of the International Conference on Emerging Trends in Computing (ICETiC09)', pp. 309–315.

Granitzer, M., Hristakeva, M., Knight, R. & Jack, K. (2012), A Comparison of Metadata Extraction Techniques for Crowdsourced Bibliographic Metadata Management, *in* 'Proceedings of the 27th Symposium On Applied Computing', ACM New York, NY, USA.

Granitzer, M., Hristakeva, M., Knight, R., Jack, K. & Kern, R. (2012), A Comparison of Layout based Bibliographic Metadata Extraction Techniques, *in* 'International Conference on Web Intelligence, Mining and Semantics 2012'.

Hammerton, J. A., Granitzer, M., Harvey, D., Hristakeva, M. & Jack, K. (2012), On generating large-scale ground truth datasets for the deduplication of bibliographic records, *in* 'International Conference on Web Intelligence, Mining and Semantics 2012', number January.

Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z. & Fox, E. A. (2003), 'Automatic document metadata extraction using support vector machines', *2003 Joint Conference on Digital Libraries 2003 Proceedings* pp. 37–48.

Han, H., Manavoglu, E., Zha, H., Tsioutsiouliklis, K., Giles, C. L. & Zhang, X. (2005), 'Rule-based Word Clustering for Document Metadata Extraction', *Proceedings of the 2005 ACM symposium on Applied computing SAC 05* p. 1049.

Haupt, J. (2009), 'Last.fm: PeoplePowered Online Radio', *Music Reference Services Quarterly* **12**(1-2), 23–24.

Henning, V. & Reichelt, J. (2008), 'Mendeley - A Last.fm For Research?', *2008 IEEE Fourth International Conference on eScience* pp. 327–328.

Kapoor, N., Chen, J., Butler, J. T., Fouty, G. C., Stemper, J. A., Riedl, J. & Konstan, J. A. (2007), Techlens: a researcher's desktop, *in* 'Proceedings of the 2007 ACM conference on Recommender systems', ACM, pp. 183–184.

Kraker, P., Körner, C., Jack, K. & Michael, G. (2012), Harnessing User Library Statistics for Research Evaluation and Knowledge Domain Visualization, *in* '1st International Workshop on Large Scale Network Analysis, WWW 2012'.

Lawrence, S., Giles, C. L. & Bollacker, K. D. (1999), Autonomous Citation Matching, *in* O. Etzioni, J. P. Mller & J. M. Bradshaw, eds, 'Proceedings of the 3rd International Conference on', Vol. 1, ACM Press, pp. 392–393.

Linden, G., Smith, B. & York, J. (2003), 'Amazon.com Recommendations: Item-to-Item Collaborative Filtering', *IEEE Internet Computing* **7**(1), 76–80.

Narayanan, A. & Shmatikov, V. (2008), Robust De-anonymization of Large Sparse Datasets, *in* '2008 IEEE Symposium on Security and Privacy sp 2008', Vol. 0, Ieee, pp. 111–125.

Parra-Santander, D. & Brusilovsky, P. (2009), Evaluation of Collaborative Filtering Algorithms for Recommending Articles, *in* 'Web 3.0: Merging Semantic Web and Social Web at HyperText 09', Torino, Italy, pp. 3–6.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. & Riedl, J. (1994), 'GroupLens - An Open Architecture for Collaborative Filtering of Netnews'.

Vellino, A. (2009), Recommending Journal Articles with PageRank Ratings, *in* 'Recommender Systems'.

Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M., Vuorikari, R. & Duval, E. (2011), Dataset-driven Research for Improving Recommender Systems for Learning, *in* '1st International Conference on Learning Analytics and Knowledge, LAK 11', ACM, New York, NY, USA, pp. 44–53.

Wang, C. & Blei, D. (2011), Collaborative topic modeling for recommending scientific articles, *in* 'Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 448–456.