Exploring Library Loan Data for Modelling the Reading Culture: Project LibDat

Mats Neovius[1], Kati Launis[2] and Olli Nurmi[3]

[1] Åbo Akademi University, Dept. of Computer Science, Turku, Finland
[2] University of Eastern Finland, Dept. of Literature, Joensuu, Finland
[3] VTT Technical Research Centre of Finland Ltd, Data-driven solutions, Espoo, Finland
mats.neovius@abo.fi, katijl@uef.fi, olli.nurmi@vtt.fi

Abstract. Reading is evidently a part of the cultural heritage. With respect to nourishing this, Finland is exceptional in the sense that it has a unique library system, used regularly by 80% of the population. The Finnish library system is publicly funded and free-of-charge. On this, the consortium "LibDat: Towards a More Advanced Loaning and Reading Culture and its Information Service" (2017–2021, Academy of Finland) sets out to explore the loaning and reading culture and its information service to the end that this project's results would help officials to elaborate upon Finnish public library services. The project, as well as the data-analysis, has just started, so the paper contains more hypotheses than final results. The project is part of the constantly growing field of Digital Humanities, and its most important scientific benefit is to show how large "born digital " material, new computational methods and literary-sociological research questions can be integrated into the study of contemporary literary culture. The project's collaborator, Vantaa City Library, has collected daily loan data since July 27, 2016. This loan data is objective, crisp, and big. In this position paper, the main contribution is a discussion on the limitations that the data poses and the literary questions that may be explored by computational means. For this, we describe the data structure of a loan event and outline the dimensions of how to interpret the data.

Keywords: Finnish loaning and reading habits, library loan data, sociology of literature.

1 Introduction

Reading is an inherent part of our culture. With sufficient data on the reading behaviour and with modern computational methods to analyse this data, this paper sets out to discuss the possibilities of deriving novel views on contemporary reading behaviour. Given historical data on reading behaviour, depicting the evolution of this part of the culture is possible. Such analysis opens a possibility for posing novel qualitative humanist questions, especially in the field of literary studies, and acquiring an answer by the analysis of data. Interesting questions to explore include whether there exists any shared generational and social patterns in the reading culture, what does the current

Finnish reading culture look like, how has it changed since the 1970s when it was characterized by uniformity [1]; who maintains the Finnish reading culture, do library services reach younger readers, do the classics of Finnish literature still attract readers as they did in the 1970s in the studies of Katarina Eskola? Answering these questions provides insights to the loaning and reading culture where the data was collected and, as a side result, data that may help library services to improve their service.

The loan data explored originates from the Vantaa City Library in Finland's metropolitan area. This systematically collected and digitally preserved data offers information about hundreds of thousands of people's loaning behaviour using Vantaa City Library services. It is peculiar in its character. In this data, an event is a loan. An event's basic field consists of the library user (as a hash over the personal information), hereinafter an agent; and the book copy's id, hereinafter an artefact. A series of such events resembles data collected by streaming services, e.g. Netflix and Spotify. However, where streaming services are typically interested only in recommending the next artefact to the agent, library services frequently look for the trend with a long-term and not purely financial motivation. A characteristic of the data is that loan data do not capture the most significant indirect metadata: whether an artefact was consumed to its full extent or not. Other characteristics include repeated loaning of an artefact; yet the agent may only (possibly) have consumed this once. Pragmatic reasons may be that the agent was unable to return it by the due date, and web-renewal is easily available. At the same time, the loan data is subject to privacy with hashed identity making identification irreversible.

The contribution in this paper is threefold: (1) Firstly, we present the peculiarities of the loan data we explore. This loan data is acquired through the project LibDat, whose goal is in enabling the asking of a set of novel questions on the data from a literary research perspective. Unfortunately, we are unable to publish the raw loan data, as it is disclosed specifically for this research. Instead, it is possible to publish the data related to the book collection that the UPCV recommender system has generated based on this data. (2) Secondly, we discuss computerised analytical methods that may be well suited to this approach. We discuss algorithms for analysing this data and suggest coarse directions in the selection. Finally (3), we discuss the literary questions that may be explored by computational means. By computational means, we mean here methods of data science / analytics to mine the data. We omit discussing scoring loan events, as this a challenge in its own right that is specific to the data type and use case. Thus, this paper is a position paper presenting ideas on how, why and for what library loan data could be used, and serves a cause within the digital humanities.

2 Library Loan Data

The basics of a library loan event consist of a pair: agent–artefact; mathematically $Events = \{(u, v)\}$. Thus, given an event $e \in Events$, the pair (u, v) where $u \in Agent$ and $v \in Artefact$ outline a loan event where u loaned v. The event, agent and artefact each have metadata associated with them. To mention a few, the agent's metadata include age, gender, address; metadata of an artefact include author, publication year, publisher, classification; and an event's location, time, reservation. In practice, an agent's unique

id is the library card registered to a user, where the id is the user's date of birth and social security number hashed. The reason is privacy and continuity; an agent's real identity must be anonymized and irreversible, whereas a library card may be lost or the user's name(s) may change, but date of birth and social security will remain. Moreover, a parent may loan books for the whole family, and thus one agent id may, in fact, correspond to many persons. Additional metadata for the artefact can be crawled from external sources. Such sources include e.g. reviews, authors' interviews in the mass media, book marketing by the publishers, prize laureates, literary blogs, and review platforms for the readers (e.g. LibraryThing.com and Kirjasampo.fi) or other significant input that may affect lending behaviour.

Fundamentally, an artefact is a copy of a book. Obviously, a library may own several copies of one book. In the event of a loan, the artefact is affiliated with an agent with a timestamp, i.e. a three-tuple. Some events may be chained together so that a number of timestamps may be derived from them, including the day of the loan, due date and return date. The set of data is incremental in terms of events, agents and artefacts. The temporal order of the events is also of significance, as this indicates the trend and migration from one category to another. In addition, a sign of an extraordinary event – such as a change in the political situation, economic success or anything of appeal to the reader – may affect the reader. We plan to crawl Twitter (as of its structure with # as the denominator) for this information, and start collecting some Twitter-feeds that are related to literature consumption.

There are seasonal variations in library usage and number of loans. The high seasons are the turn of the year and beginning of the summer holiday period. Also, the local socio-economic structure in the vicinity of the library is reflected in the data.

3 Methods for Exploring the Library Loan Data

The domain of making sense of data by exploring it is vast. Its roots are in the natural sciences where researchers have for decades sought to explain some phenomenon by formalizing a solution mathematically. Today, this exploration is at the intersection of machine learning, statistics and database systems. These methods can further be categorized into two main categories: descriptive and predictive analytics.

This section will concentrate on understanding the objectives and requirements from the user's perspective, and converting this view into a problem with an envisioned ensemble of methods for an analytic approach on that problem. This is normally the entry point for data mining as described in the cross-industry standard process (CRISP-DM).

3.1 Predictive and Descriptive analytics

Descriptive analytics includes methods that scrutinize data and information in order to define the current state in such a way that developments, patterns and exceptions become evident. For the analytics, statistical methods are used, such as mean, median, mode, standard deviation, variance, and frequency measurement of specific events. Descriptive models quantify relationships in data in a way that is often used to classify customers or prospects into groups and identify their relationships. This category in-

cludes also methods to probe data to confirm/reject a hypothesis, for example, analytical drill-downs into data, statistical analysis, and factor analysis. Methods for predictive analytics are applied when the domain cannot be defined due to its complexity. Frequently, the approach is then learning from the past observations, realistically e.g. weather forecasts, pollen distribution, etc.

Depending on the use case, library loan analytics may belong to either category. For a specific problem, the chosen data analytic approach is often an ensemble of the two of the aforementioned categories. This ensemble, together with careful feature selection, is often an iterative process. Lessons learned from one prototype are used to tweak the setup. Hence, this paper is a first, but very important step documenting the guiding questions for the future direction of the project. At the end, the stakeholder appreciates an expressive visual result with a well-studied user-interaction strategy for query formulation. Interesting query formulation may be at different levels of abstractions, easy transitions from one scale to another or form of aggregation to another (e.g. from neighbourhood-level to city-level). Pragmatically, these may be what "is the most trending book" given certain criteria or which "book would be recommended to me". Directions for answering the questions include, but are not limited to, methods of similarity.

3.2 Methods Determining Object (Dis)Similarity

There is an abundance of collaborative filtering methods quantifying the (dis)similarity of any pair of objects, i.e. agent-agent, agent-artefact, artefact-artefact. They are frequently based on the distance between any two points in a space; interested readers are referred to the survey by Adomavicius and Tuzhilin [2]. The vectors spanning this n-dimensional space are composed of possibly pre-processed focal elements, often called features. For a simplified example, consider a two-dimensional coordinate system with an *x* and *y* axis. In this space, the Pythagorean Theorem gives the distance of two points, i.e. the length of the hypotenuse is the distance. In *n*-dimensional spaces, this same idea is called the Euclidean distance. The distance quantifies the level of (dis)similarity between those two objects.

A set of objects pairs $\{(o_1, o_2)\}$ is represented as a matrix $A = (m, n)$ where $m = |o_1|$ and $n = |o_2|$. On each row *m*, there are n references, e.g. for each user (row) there is a reference to every book (column) in the library system. Only given that the agent has loaned that book, the $(i, j) \in \mathbb{R}$. The distance of two objects is then $sim(u, v) = \frac{1}{1+dist(u,v)}$, i.e. the similarity is high when distance is small. The similarities among objects can be used to calculate a prediction of u on v by: $score_u(v) = \frac{1}{\sum_{s\in S} sim(\bar{x},\tilde{x})} * \sum_{s\in S} sim(\bar{x}, \tilde{x}) * x_{sv}$. Here S are the scored entries, i.e. such $A_{(i,j)} \neq 0$, that is any entry with a score, $\bar{x}$ is subject and $\tilde{x}$ the target and x_{sv} the score. Dually, the similarity between the artefacts is readily available by the same function merely by transposing matrix *A*. The distances do not qualify, as they do not normalize the inputs.

Normalization of the input is straightforward, $\bar{x}_s = \frac{1}{|A_{(x,j)\neq 0|}} \sum_{s\in S} x_s$, i.e. the mean for object *x* is 1 over the amount of nonempty entries times sum of the scores x_s. The standard deviation $std = \sqrt{\frac{1}{|A_{(x,j)\neq 0|}-1} \sum_{s\in S}(x_s - \bar{x}_m)^2}$ in case the score is $\frac{x_s - \bar{x}_s}{std}$. A

means to normalize the input is by using the Euclidean dot product. Given this, cosine similarity $\cos(\theta) = \frac{A*B}{\sqrt{\sum_{\alpha \in S} A_\alpha^2}*\sqrt{\sum_{\alpha \in S} B_\alpha^2}}$. If the measure is normalized by the means of respective subset, then this is equivalent to Pearson correlation coefficient.

4 Impact of Big Data Analysis for Literary Research

Library loan data provides a significant, new resource for literary scholars to understand the literary culture from a wider perspective, and thus, study literature as a sociocultural phenomenon from a new viewpoint. The data collected by Vantaa City Library gives information about each library user's age, sex, language and place of residence in Finland's metropolitan area. Likewise, it shows what sort of cultural products each library user has loaned. Hitherto, library data has rarely been used in Finnish literary studies. Now that this systematically collected and digitally preserved data is available for scholarly use, it is possible to ask what the contemporary reading culture looks like on the basis of this big data. The task of the literary scholars in the project is to formulate the exact research questions used in data-analysis and to interpret (by using the theoretical tools e.g. from the reception studies and the feminist literary studies on the act of reading) the results mined from the library data by computational data-driven methods. As such, the project is part of the second wave of digital humanities, which is not quantitative, but more like "qualitative, interpretive, experiential, emotive, generative in character" [3].

The LibDat-consortium thus serves as a model for crafting old humanist questions and new technology into something unprecedented in terms of methodology. How does the data shed light not only on the literary culture, but also on the methods of literary-sociological research? The project shows the full research potential of library loans data for literary studies, and at the same time it aims at integrating novel computational methods (see above) into humanist disciplines.

In literary studies, reception theory [4] emphasizes reader's interpretation in making meaning from a literary text at a certain historical moment. In earlier studies of Finnish readership and reading culture, methods such as interviews and queries have been widely used [1, 5, 6, 7]. However, the big library data used in LibDat -project is a different, significant resource for understanding literary culture from a wider perspective: the library loan data is daily, big and objective, unlike e.g. interviews or queries.

Our literary-sociological method or way to analyse the Finnish reading culture approaches the practice called 'distant reading' (the opposite of 'close reading'), created by literary scholar Franco Moretti [8] (see also [9]). It means a data-centric approach to novels; it means viewing literature as data, a system of using computers to analyse novels as raw data, searching and finding patterns and rules behind literature – or, in this case, the rules behind the reading culture.

It is also possible to problematize widely accepted views of the uniformity of Finnish reading culture and Finns as particularly realistic readers. On the basis of our preliminary analysis of the library data, we assume that the situation has changed and that today it is middle-aged female readers who maintain literary culture in Finland and that the national classics (such as Väinö Linna) no longer attract readers as much as they did in the 1970s. On the grounds of the preliminary analysis (based on one sampling

with 1,556 library users) (approximately) 67% of the loaners were women, 96% of the them loaners used Finnish language and 57% of them were aged 25–64. Over 85% of the all loans were books (and 62% of the books were fiction). One of the notable changes is the digitalisation of the reading culture, even though traditional, printed books are still much more popular than e-books.

Because of the superior numbers of middle-aged women as loaners we are at the moment analysing the gendered reading culture in contemporary Finland: Which genres and books more specifically do these women readers favour and why? Do they read domestic or foreign literature? What kind of cultural, political, and social phenomena potentially affected their book choices?

Another interesting question for the study of literary culture and reading is the historical 'horizon of expectations' (e.g. [3]), the criteria contemporary readers use to judge literary texts in Finland nowadays. How can we, for instance, explain the enormous popularity of Kjell Westö's novel Rikinkeltainen taivas (2017), with about 3 000 reservations in the HelMet-library in September 2017? On the grounds of the recommendation service and the associated loan data, we may ask, for example, whether there is an entry point for reading, a specific book that "hooks" the reader.

5 An Implementation of Data Analysis on Library Loan Data

Recommendations systems have been proposed as essential tools in assisting users to face the "information overload" problem, and they have been applied across several domains, including for recommending music, TV programmes and digital libraries to cite just a few. In September 2014, HelMet libraries (Helsinki Metropolitan Libraries) together with the Technical Research Centre of Finland (VTT) opened a novel book recommendation service recursively updating itself by means of new loan data. The approach developed is based on the user's past behaviour, items previously loaned as well as similar decisions made by other users. This information is used in order to predict items or ratings for items that the user may have an interest in.

The recommendation service applies ubiquitous personal context vectors (UPCV) and a collaborative recommendation method to predicting items that the user may have an interest in [10]. The principal idea is that each user-item interaction exchanges a set of tokens associated with both the user and the item. The update makes item data to slightly resemble user data and vice versa, leading to an increasing similarity between them. Through interactions, similarity will spread from users to items, from items to users, making it possible to inherently provide user-item, item-item, item-user and user-user recommendations, globally, across any service.

These token stacks reflect the interactions that the readers have had with the library collection, and thus carry interesting information. On the other hand, they do not carry any sensitive personal information, and can thus be exposed to suitable cluster analysis methods so as to produce valuable information on reading behaviour and its changes.

6 Conclusions and Future Work

This paper discussed the impact of big data analysis of library loan data in defining the reading culture. The task is to serve literary scholars with a platform on which to ask novel questions whose answers are served by data analytics. For this task, anonymized library data provides an objective and new data source for scholars to perform literature studies on. Presumably, the data will reveal that the reading culture has changed, and it cannot be described as uniform, rather as fragmented. This change has already been observed e.g. in media consumption due to the new digital channels and devices.

In order to be able to verify and study this kind of hypothesis, LibDat will develop Big Data analytics tools and methods. By bringing together the knowledge, algorithms, IT tools and resources, it is possible to develop a compelling research tool for the study of literature. This preliminary study reveals that some of the essential patterns in the loan data are related to the additional metadata, context and season. These issues have to be taken into consideration in order to be able to draw meaningful conclusions.

References

1. Eskola, K.: Suomalaiset kirjanlukijoina. Tammi, Helsinki (1979).
2. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. In IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734–749, (2005).
3. Berry, D.: Introduction: Understanding the Digital Humanities. Understanding Digital Humanities. Ed. by David M. Berry. UK: Palgrave Macmillan (2012).
4. Jauss, H.: Kirjallisuushistoria kirjallisuustieteen haasteena. Translated by Nevala M-L Kirjallisuudentutkimuksen menetelmiä. SKS, Helsinki (1989).
5. Eskola, K.: Ei kirjaa ilman lukijaa. Raportti kirjallisuuden julkisesta ja yksityisestä vastaanotosta. Tammi, Helsinki (1972).
6. Eskola, K.: Lukijoiden kirjallisuus Sinuhesta Sonja O:hon.Tammi, Helsinki (1990).
7. Eskola, K. ja Linko, M.: Lukijan onni. Poliitikkojen, kulttuurieliitin ja kirjastonkäyttäjien kirjallisista mieltymyksistä. Tammi, Helsinki (1986).
8. Moretti, F.: Distant Reading. Verso, London & New York (2013).
9. Elo, K.: Digitaalisen historiantutkimuksen kenttää louhimassa. Digitaalinen humanismi ja historiatieteet. Turun Historiallinen yhdistys, Turku (2016).
10. Ollikainen, V., Mensonen, A., Tavakolifard, M.: UPCV - Distributed recommendation system based on token exchange. Journal of Print and Media Technology Research, vol. 2, no 3, pp. 195–201 (2013).