Evaluating systematic transactional data enrichment and reuse

Jim Hahn
jimhahn@illinois.edu
University of Illinois at Urbana-Champaign
Urbana, Illinois, USA

ABSTRACT

A library account-based recommender system was developed using machine learning processing over transactional data of 383,828 check-outs sourced from a large multi-unit research library. The machine learning process utilized the FP-growth algorithm [13] over the subject metadata associated with physical items that were checked-out together in the library. The purpose of this paper is to evaluate the results of systematic transactional data reuse in machine learning. The analysis herein contains a large-scale network visualization of 180,441 subject association rules and corresponding node metrics.

CCS CONCEPTS

• **Information systems** → **Digital libraries and archives**; **Recommender systems**; *Data mining*; *Association rules*; *Personalization*; • **Computing methodologies** → *Network science*.

KEYWORDS

data reuse, machine learning

ACM Reference Format:
Jim Hahn. 2019. Evaluating systematic transactional data enrichment and reuse. In *Artificial Intelligence for Data Discovery and Reuse 2019 (AIDR '19), May 13–15, 2019, Pittsburgh, PA, USA*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3359115.3359116

1 INTRODUCTION

This paper details a network analysis of the seed data for a library account-based recommender system that was built by way of systematic transactional data reuse employing machine learning techniques. Network science metrics were computed over the network of topics in order to better understand the nature of subject metadata that the resulting recommender system was comprised. According to Börner, "The study of networks aims to increase our understanding of what entities interact with each other in what ways. Data sets are represented as nodes and edges. Nodes might denote authors, institutions, companies, and countries or words, papers, patents, or funding awards. Edges represent social, scholarly, financial, or other inter linkages" [2]. The research question in the present study is to explore, with network analysis, the nature of topic associations that drive the machine learning based recommender system. Informetric studies, with the analysis of the scholarly record, particularly the concern with citation link analysis, is a related companion field to network science [6, 24]. The preliminary offline machine learning workflows were undertaken in WEKA [12] and made use of an FP-growth algorithm [13] for seeding the recommender data for library user accounts. With these topic metadata clusters a rule set for the recommender system was developed. The prototype recommender study began in October 2016 with seed data of 33,060 consequent subject association rules from initial machine learning processes [11]. These clusters form the basis for the prototype library account-based recommender incorporated into the library mobile app. In the current version, updated over time with ongoing data collection the system contains 383,828 transactions, which after data mining association rules with FP-growth, resulted in 180,441 association rules.

2 BACKGROUND

A previous study on mobile account-based recommender systems detailed the processing and middleware development steps taken to develop such a system [11]. While the case study was descriptive of the machine learning process, it did not undertake the systematic evaluation of the topic outputs of the machine learning processes. Related work also investigated several methods to reuse library circulation data and corresponding topic associations to improve relevancy rankings for search algorithms [9]. In light of systematic bias in the real world, objectively studying the outputs of machine learning has become increasingly important. Prior studies in the ethics of machine learning have shown that without attention to algorithmic bias, that machine learning systems will contain bias inherent in the original training source [17, 18]. Modern algorithms have also exhibited the problematic nature of reinforcing systematic bias intensifying many of the pressing social concerns of our era; such as poverty, racism, and the erosion of democracy [8, 19, 20]. While it is beyond the scope of this work to propose an alternative subject classification scheme for the universe of knowledge, this work attempts to undertake objective measures of the recommender systems by way of network science metrics; while at the same time underscoring here the inherent biases of the underlying subject data attached to books in the library.

2.1 Library of Congress Subject Headings

While the subject classification is flawed by contemporary critical social science standards, the Library of Congress Subject Headings (LCSH) controlled vocabulary remains a pragmatically useful classification regime. One of the drawbacks of LCSH in data mining and reuse as undertaken in this work, is the need to simplify the complex pre-coordinate LCSH system. Svenonius wrote of the LCSH

AIDR '19, May 13–15, 2019, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-7184-1/19/05…$15.00
https://doi.org/10.1145/3359115.3359116

controlled vocabulary that "...the LCSH language begins with a main heading whose purpose is to bring out the major concepts in a document, to capture its essential aboutness. This may or may not be followed by qualifying terms called subdivisions. Syntax rules specify when subdivisions can be used and in what order." [21]. The simplification of LCSH employed here necessarily causes some subject data to become less verbose. Pattern mining of the type used here may ultimately recapture some of the intended semantics. It was ultimately necessary to use a simplification of the LCSH language for ease of data mining the chosen algorithm (FP-growth) within the WEKA toolkit. WEKA requires binaries as values when creating vectors for FP-growth.

2.2 Data Reuse in Machine Learning

Data are fundamental for machine learning projects to be successful; the contemporary wave of neural network based machine learning owe their successes in part to the massively large data sets that are available to researchers[23]. Data reuse is critically important for e-science, including efforts to maximize the benefits of sponsored research projects by making data available for continued benefit after the grant funded project has completed its objectives [7]. The departure point of this machine learning project is that the transactional data of the research library can be made far more valuable for resource discovery when enriched and re-associated with its fundamental subject metadata: the assigned LCSH in the item record. In the classic integrated library system database structure, when an item circulates out of the physical collection, the system records this transaction. These transaction database tables are not designed to record associated subject metadata of subjects that circulate together. This necessitated the research approach to undertake re-association of subject metadata into the transaction record. This extended the value of two disparate data sets; item level metadata and of the transaction data. Re-association is fundamental to development of subject association rules using FP-growth. The association rules are developed using the enriched transactional data for library resource discovery by way of a supplementary recommendation paradigm. The classic paradigm of library discovery has heretofore been focused on searching known items and subject exploration with controlled vocabulary, taxonomies, or keywords, rather than incorporating data from transaction logs for item resource recommendations. Network science helped to understand more fully the results of the association patterns that were built using transactional data.

2.3 Network Science Terminology

Network Science is defined as "... concerned with the study of networks, be they biological, technological, or scholarly in character. It contrasts, compares, and integrates techniques and algorithms developed in disciplines as diverse as mathematics, statistics, physics, social network analysis, information science, and computer science. Network science is an emerging, highly interdisciplinary research area that aims to develop theoretical and practical approaches and techniques to increase our understanding of natural and man-made networks." [3]

3 NETWORK METHODS AND METRICS

The consequent association rules are stored in a production relational database server accessed through the recommender app's middleware. The association rule database is used at run-time for the machine learning based recommender system. As a result of the availability of a database with association rules, researchers further evaluated their properties using network analysis software where the edges of the graph are the premises and the nodes are topic metadata. A preliminary analysis of the network structure of the consequent topic nodes is visualized below in Figure 1 – generated using Gephi open source visualization software – a tool that is commonly used for exploring networks [1], and a cornerstone for undertaking network science research and development. Network science can be particularly valuable in assisting the understand of machine learning outputs by visualizing and analyzing graphs the distribution of the topic network. Several key network science metrics, including the average degree, diameter, and average clustering coefficient were computed by way of Gephi network plotting and network measurement software.

3.1 Network Average Degree and Network Diameter

The network average degree can be used to understand the centrality of nodes in a network [3]. Average degree is calculated as the average number of edges connected to a node. In a directed network such as the one derived here, the calculation is performed by dividing the number of edges by the vertices. Diameter, like the average degree has proven useful by researchers to help classify network types [3]. It is the shortest distance among the farthest nodes in the network.

3.2 Network Average Clustering Coefficient

Also a measurement for how a network may be classified by network science researchers [3]. This measurement can be useful in measuring the way in which nodes cluster together. Higher number of clusters tend to be associated with the regular lattice scale free network class [3].

3.3 Connected Components

The origins of study of connected components in information science seems to be drawn from the study of the topology of the web [3]. It was initially believed that much of the web was strongly connected, however, full scale studies of the early web seemed to indicate that much of the web was actually weakly connected [5]. In the context of the present study, the numeric value of weakly and strongly connected components are an indication of how closely connected parts of the graph (nodes) are to other nodes.

4 NETWORK ANALYSIS

Figure 1 is visualized with the Force Atlas 2 settings within Gephi. The structure of the graph is comprised of 41,054 nodes and 180,441 edges. The resulting network type is comprised of massive overlapping central hubs. Node outliers were plotted along the margins and outside the central network hubs are of interest for further analysis. Network analysis was performed within Gephi by computing key

Table 1: Network Measurements

Metric	Value
Network Average Degree [1]	4.395
Network Diameter [4]	12
Average Clustering Coefficient	0.232
Number of Weakly Connected Components[22]	614
Number of Strongly Connected Components[22]	30,976

Table 2: Top 10 Network Authorities [16]

Topic Node	Authority Value
feature films	0.582934
video recordings for the hearing impaired	0.515518
man-woman relationships	0.4549
families	0.174763
murder	0.148486
comedy films	0.124917
friendship	0.109092
foreign films	0.079265
world war 1939-1945	0.076611
horror films	0.076259

descriptive metrics shown in Table 1. Plotting the topic association network revealed several motifs. For a larger example see figure 3.

5 FINDINGS

The significance of analyzing nodes within the topic graphs has led to a new understanding of recommending topics in the library. As it pertains to the particular structure of the network, we can see that central hubs and smaller networks within the network appear to be a cornerstone in information discovery within the topic based recommender system. With regard to the computing metrics of Table 1, these scores are have used in the network science literature to attempt to classify the type of network and make network to network comparisons (e.g. the World Wide Web, to that of Biological Webs); however comparing the similarities of various networks to derive findings have fallen out of favor in recent scholarship as the uniqueness of graphs has become increasingly apparent [14]. Network researchers note that, "...the origin of a network, whether it is biological, technological, or social, may not necessarily be a decisive factor for the formation of similar network structure" [15].

As an example comparison to other network typologies, certain network measurements reported in Table 1, such as the diameter of the network closely aligns with the measurements associated with lattice network types. The average clustering coefficient of the network however has some measurement commonalities with random networks, and scale free / heavy-tail networks [3]. The network average degree of the subject associations (4.395) is similarly found in each of the three network types mentioned above. Figure 2 is plotted with colors denoting network degree. It may be more fruitful to the library-specific and a collections focused analysis to examine individual node metrics. As it pertains to node characteristics within the network, Table 2 is instructive in that it

Figure 1: Directed graph network is comprised of 41,054 nodes and 180,441 edges.

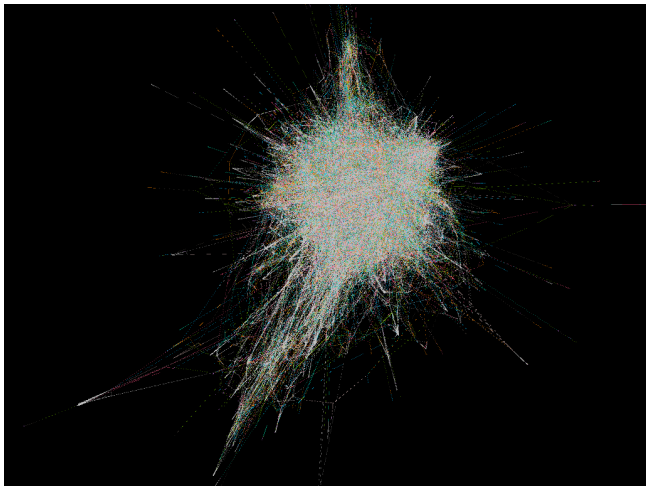

Figure 2: Color denotes network degree.

reports a previously unknown understanding of topical associations in the library. Specifically, preliminary findings of authority [16] in the network includes the emergent importance of media, feature films, video recordings, and popularly occurring associated topics that may help to tie together many disciplinary strands of inquiry. Therefore, it may be advantageous to examine in greater detail the topical relationships that can support interdisciplinary and serendipitous discovery. There is also evidence of hundreds of out-network nodes that may encompass topics in esoteric or specialized topic areas, or in topic areas in which antecedent topics