Book Recommendation System using Data Mining for the University of Hong Kong Libraries

Sandhya Rajagopal & Alvin Kwan , Faculty of Education, University of Hong Kong
[sandhyaraj@gmail.com; cmkwan@hku.hk]

This paper describes the theoretical design of a Library Recommendation System, employing k-means clustering Data Mining algorithm, with subject headings of borrowed items as the basis for generating pertinent recommendations. Sample data from the University of Hong Kong Libraries (HKUL) has been used in a Quantitative approach to study the existing Library Information System, Innopac. Data Warehousing and Data Mining (k-means clustering) techniques are discussed. The primary benefit of the system is higher quality of academic research ensuing from better search results. Personalization improves individual effectiveness of learners and overall in better utilizing library resources.

Key Words: Recommender Systems, OPAC, Subject searching, Data Mining, Library Technology

Introduction

Internet search engines have significant influence over how users pursue information, with speed, convenience and ease of use identified as key preferences in information seeking behaviors. Consequently they "... expected the library catalog to function as a search engine, where one types words into a box, and gets results" (Novotny, 2004, p. 533). In some cases there is even an alarming tendency, to compromise information quality for convenience (Connaway, Dickey, & Radford, 2011). Furthermore, "Users find OPACs frustrating; they complain of finding the catalogue illogical, counter-intuitive, and intimidating" (Fast & Campbell, 2004, p. 139).

A key advantage of Library OPACs is their capability to provide precise access to well-organized information. Subject organization in libraries provides a distinct advantage over that of the Internet and subject searches have a distinctly better recall rate than keyword searches (Voorbji, 1998). Despite this tremendous benefit for researchers, a decline in the use of the subject index for searching has been identified by Larson (1991), as due to " ... user difficulties in formulating subject queries with Library of Congress Subject Headings, ..." (p. 197).

One way to enhance the effectiveness of OPACs and simultaneously capitalizing on the power of subject searches is by personalizing access services such that the onus of retrieving related items lies with library systems rather than with the patrons. Recommendation Systems, which assist in navigation of complex information spaces by making relevant suggestions to users, can be adopted well, in personalizing access services. "From a librarian's point of view, a recommendation system can be seen as a form of catalog enrichment or as a substitute for traditional subject classification" (Monnich & Spiering, 2008).

Most Recommendation Systems include, (i) a 'Recommendation component' or 'Recommendation generator', which processes data inputs and (ii) the 'Recommendation list', which is the final output from the system, (Jannach, Zanker, Felfernig, & Friedrich, 2011).

Among many techniques used to configure the ‘Recommendation component’, a particular instance of a Data Mining algorithm forms its core. It can therefore be deduced that, Data Mining is which can be defined as : “... the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules” (Berry, 2004, p. 7) is one of the most effective methods of personalizing information access in Library OPACs.

In this study appropriate data from the University Of Hong Kong Libraries (HKUL) was used to design a theoretical model of a Recommendation System, by applying *k*-means clustering Data Mining algorithm, with Subject headings forming the cluster centers. The contention is that, by capitalizing on the efficiency of Subject searches, it will be possible to maximize the output of the Recommendation System hence enhancing the effectiveness of Library OPACs.

Literature Review

Recommendation System Algorithms

Recommendation Systems are software applications, which serve as personal advisors and support decision-making processes of users by suggesting items of interest to them. The two primary types of Recommendation System algorithms identified in literature and employed in commerce are:

(i) Collaborative Filtering (CF) algorithm: In this method, the system computes and matches similarities in preferences across users based on the ratings they have assigned in the past to various items, making it a preferred algorithm on e-commerce web-sites where number of customer transactions and their social interactions on the site are high.

(ii) Content-based algorithm: These systems recommend items that a user has preferred in the past by matching them with the characteristics of the item and generate new items of potential interest to this particular user, hence making the recommendations more customized.

Since the rate of repeat utilization of the same book across Library users and interactions among them is low, for this study, content-based algorithm is deemed more suitable for the design of a Recommendation System.

Data Mining

Functionality

Data mining algorithms are used for several purposes during the exploratory process of uncovering hidden knowledge. These can be enumerated as follows:

Description, where the Data Mining tool is expected to provide a clear depiction of underlying patterns in data, enabling intuitive interpretation (Larose, 2005)

Classification categorizes new items into pre-defined groups called a training set and the Data Mining task is to construct a model based on the training set to enable unclassified data to be assigned groups automatically

Estimation is another form of classification where the target variables are numeric rather than categories

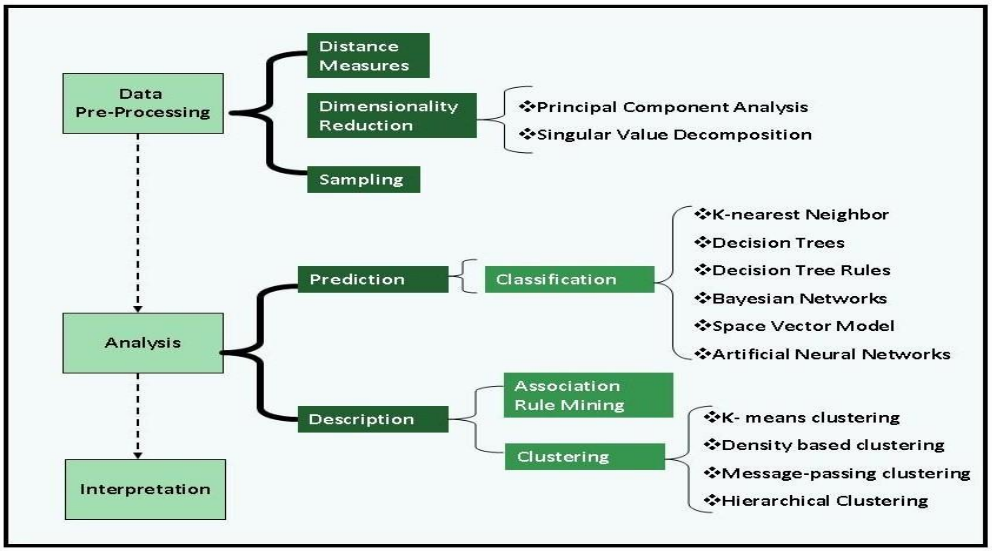
Prediction, much like estimation makes forecasts, but the results lie in the future and are not immediately verifiable

Clustering is the clumping of all items in a data set, around random or pre-determined cluster centers, based on similarities in their characteristics. The objective of clustering is to maximize similarity of items within a cluster and minimize similarity across other clusters in the data set, hence enabling the clear identification of like items

Association or affinity analysis or market-basket analysis, determines which characteristics go together. Association rule mining is, "... the process of efficiently discovering meaningful rules from the large collections of data" (Lee & Lee, 2011, p. 483).

Process

The process of Data Mining comprises of three parts, namely: Data Preprocessing, Data Analysis and Results Interpretation. The following diagram [adapted from (Amatriain, et al., 2011)] depicts the various algorithms that are considered suitable in the application to these parts.



Phases in Data Mining and associated algorithms

Of these methods, the *k*-Nearest Neighbor method, considered the de-facto method in Collaborative filtering systems, Artificial Neural Networks, considered too complicated to construct and Support Vector Machines, which is logically similar to clustering, are deemed unsuitable for this research.

In this study, *k*-means Clustering was used to design a theoretical model of a functionally feasible Recommendation System, where subject headings formed center clusters. This paper describes the process of this design.

Methodology

The four steps used in the design were:

(1) Analysis to gain an understanding of existing Library Information Systems (LIS) at HKUL by studying access reference Services, circulation and usage information

(2) Design a Data Warehouse (DWH) that will serve as the primary Information Repository for the Recommendation System. In the context of this research, a DWH can be defined as a data store which holds

(a) organization-specific entities such as patrons, circulation, sales, etc., as opposed to Operational Databases which store transactional, application-specific data

(b) integrated and consistent data, assimilated from multiple sources and which offers a comprehensive, cohesive view of institution-wide information

(c) non-volatile data whose progress over time is inherent to its structure.

(3) Employ k-means clustering Data Mining algorithm to group like items around subject headings and generate most suitable recommendations

(4) Model the Recommendation System framework

Step (1): HKUL uses several modules of a commercial product called Innopac, an integrated Library Information System. Chief among these modules are:

(i) Millennium Integrated Library systems, which can be broadly divided into three functional entities – Staff functions, Patron services and Campus Computing

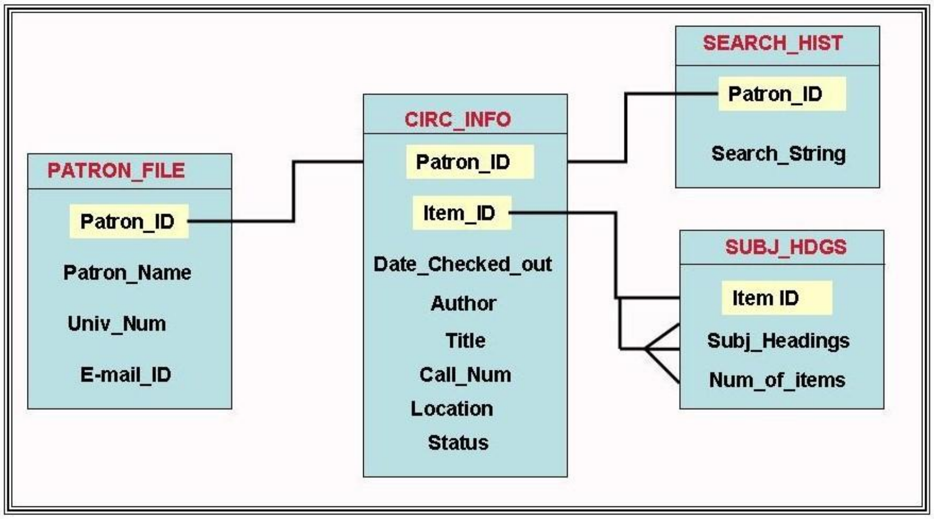
(ii) Discovery Tools, which offer several platforms for research that support a range of information needs

(iii) Resource sharing across a consortium of libraries, by enabling access to library items to patrons of any participating Library

After, an analysis of the various entities in the Integrated LIS at HKUL, the following were identified as resources and corresponding fields pertinent to this study.

Resource / Database table	**Fields**
Dragon OPAC	Author, Title, Call number, Location, Status, LC subject headings
Patron Information	Patron ID, Patron Name, University ID number, e-mail address
Circulation Information	Author, Title, Call number, Date Checked out, keyword search string, Subject Headings, Location, Status

Step (2): A simple but scalable representation of a DWH, designed based on the analysis in step (1) is shown below.



Central to this design is the Patron. The primary data stores are PATRON_FILE, containing details about the user, CIRC_INFO, containing circulation/borrowing history of the user, SEARCH_HIST, containing any search strings that may have been used by the patron during previous interactions with the OPAC and SUBJ_HDGS, containing controlled vocabulary for each item borrowed by the patron.

Since the information on search strings saved by the user and subject heading details form two separate functional entities, the DWH be structured as a conglomerate of two or more Data Marts, each catering to different operational aspects of the Recommendation System. Such a construction also allows for addition of more Data Marts, at a later time, without affecting the basic structure of the DWH.

The steps in generating information for the DWH can be enumerated as follows:

(i) Extract from Circulation Files in Innopac, those patrons whose cumulative check-outs exceed a number x, to create the PATRON_FILE. Here, x will depend on the Patron_type and the borrowing privileges programmed in the circulation module of Innopac. For e.g. at HKUL, while Under graduate students and technicians are allowed to borrow 60 items for up to 60 days, post graduates can check out 180 items and retain them for up to 120 days. Hence the value for x needs to be computed at the time of extracting patron records.

(ii) For each Patron, select all the items borrowed, with the borrower’s records sorted in reverse chronological order, so that the most recent item borrowed is on the top of the list. For each Item, extract the bibliographic information and the subject heading classifications available in the Circulation file. For each Patron extract all saved searches

(iii)For each item, extract subject headings and corresponding number of items for each heading.

(iv)Consolidate extracted data into one or more data marts.

Step (3): In Data Mining, clustering is the process of grouping like-elements within a data set. While Clustering as a technique is ideally suited for situations where data cannot be clearly classified either numerically or categorically, k-means clustering is particularly effective in cases where the number of clusters to be formed is known before the start of the clustering process.

In this research, the number of subject headings for each item is used to represent the number of clusters, k. In applying the technique to HKUL data, the following steps were carried out:

(i) Designate the number k for a set of Patron records. For the sample HKUL item, k was equal to 2 and record set equal to 138.

(ii) Assume k number of centers for this set [In the example, since k=2, there will be two centers and therefore two clusters. This means that at the end of the iterative clustering process, there will be two clusters with the 138 records positioned around one of the two centers which are most close in characteristics to either center. Cluster centers need to be vectorized to represent their positions in space. For k=2, say the points are represented as (1,1) and (2,1).

(iii) For each record in the set, find which of the k-centers is the nearest. Calculate the distance of the 138 records from the designated centers. Some common methods of measuring distances are : Euclidean distance, City Block distance and Minkowski distance. Next, cluster the closest items around the centers such that the distance between points around a center (i.e. within-cluster variation or WCV) is low but the distance between clusters (i.e. between-cluster variation BCV) increases.

In other words the ratio of BCV to WCV is maximized. Here, BCV is the distance between the two centroids in each iteration. Also WCV can be substituted with the Sum of Squared Errors measure (SSE), a statistical means of determining the variation within a cluster. Without going into many mathematical details , it is suffice to say that the ratio of BCV to WCV is equal to :

[Distance between centroids / SSE]. With each iterative pass, this ratio is expected to increase.

(iv) From this initial cluster centers, calculate the cluster *centroids* and re-assign the position of initial centers to the newly calculated centroid positions. Centroids are new cluster centers calculated based on the mean distance of items from the initial center.

To calculate the values of two new centroids for the two clusters, the formula suggested by (Larose, 2005), is as follows:

For n data points (a1,b1,c1), (a2,b2,c2), … (an,bn,cn), in a 3-dimensional vector space, the centroid of the points can be calculated using the formula : (Σai/n,bi/n,ci/n). Hence for the points (1,1,1), (1,2,1), (1,3,1) and (2,1,1) the centroid would lie at :

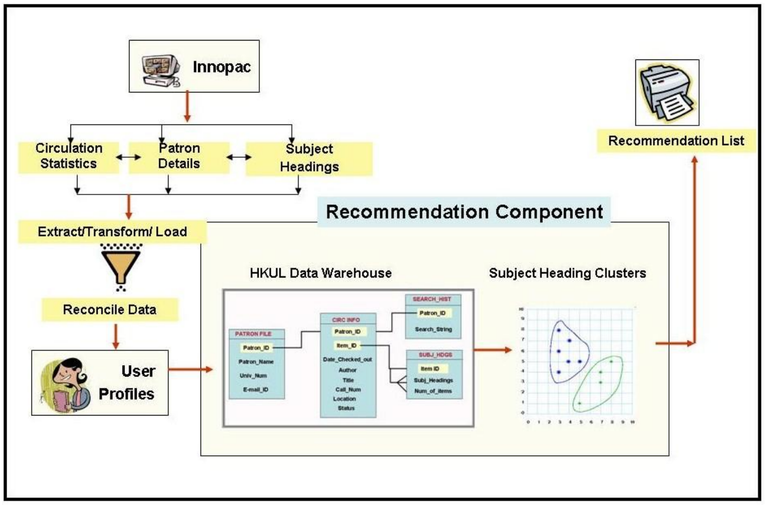
$$\left\{ \frac{1+1+1+2}{4}, \frac{1+2+3+1}{4}, \frac{1+1+1+1}{4} \right\} = (1.25, 1.75,1.00)$$

This is the mean or average values for the points and hence the name k-means clustering for the method. In this manner the centroids (or center means) can be calculated for the two clusters.

After calculating the centroids, the location of the centers should be shifted to the newly calculated positions. In this process it may happen that some of the items shift from one cluster to the other.

(v) Repeat steps (iii) to (v) until either, when the centroids do not change in subsequent iterations or when there is no significant decrease in a convergence criterion, such as the *sum of squared errors* (SSE). SSE essentially measures the distance from the center of items within a cluster.

Step (4): A model of the Recommendation System based on the above analysis is presented below:



In this model, HKUL's Library Information System, Innopac, serves as the primary source of data for the Recommendation System. Through a systematic process of Data warehousing relevant data in Innopac's operational databases, are cleaned, transformed and reconciled to be stored in the specially designed data warehouse, essentially a conglomerate of smaller Data Marts. K-means algorithm is then applied, as the chosen Data Mining tool, to the data warehouse to cluster similar items around subject headings, the number of clusters and therefore the cluster centers dictated by the number of subject headings, pertaining to each patron item.

The expected output of the system is a list of like subject headings and links to recommended items under each subject heading.

Conclusions and Further Research

This paper reports analysis of all components of the Recommendation System, namely, Data Warehousing and Data Mining (using k-means clustering technique). It can be surmised from the discussions that the research questions proposed can be effectively resolved. Hence, extraction of user profiles into a DWH and the application k-means clustering to make appropriate recommendations are theoretically feasible using the proposed design.

With this preliminary study as a basis, further Research in the following areas is warranted due to its limited scope of applicability of the design, in its current form:

(a) Qualitative means

  i. to assess and establish the efficacy of subject searches in OPACs and

  ii. to ascertain the extent of need among patrons for a Recommendation System considering that this design caters to physical items in a hybrid library, whereas trends indicate a distinct preference for digital resources by them

(b) Systems Development Research process to construct, prototype and implement the design to ensure its applicability.

(c) Assess and evaluate generalizability of the design and consequent feasibility of implementing the same or a similar design in other libraries that currently use Innopac

Acknowledgements

Many thanks to Ms. Ruth Wong, the Access Services Librarian at HKUL, who has been a valuable source of critical information for the development of the project.

References

Amatriain, X., Jaimes, A., Oliver, N., & Pujol, J. M. (2011). Data Mining Methods for Recommender Systems In F. Ricci, L. Rokach & S. Bracha (Eds.), *Recommender Systems Handbook* (pp. 39-71): Springer Science+Business Media.

Berry, M. J. A. (2004). *Data Mining Techniques : For Marketing, Sales, and Customer Relationship Management* (2nd ed.). Indianapolis : Wiley.

Connaway, L. S., & Powell, R. R. (2010). *Basic Research Methods for Librarians* (5th ed.): ABC-CLIO.

Fast, K. V., & Campbell, D. G. (2004). 07. “I still like Google”: University student perceptions of searching OPACs and the web. *Proceedings of the American Society for Information Science and Technology, 41*(1), 138-146. doi: 10.1002/meet.1450410116

Larose, D. T. (2005). *Discovering Knowledge in Data : An Iintroduction to Data Mining*. Hoboken, N.J. : Wiley-Interscience.

Lee, K. C., & Lee, S. (2011). Interpreting the Web-mining Results by Cognitive Map and Association Rule Approach. *Information Processing and Management, 47*(4), 482-490. doi: 10.1016/j.ipm.2010.11.005

Monnich, M., & Spiering, m. (2008). Adding Value to the Library Catalog by Implementing a Recommendation System. *D-Lib magazine, 14*(5/6). Retrieved from http://www.dlib.org/dlib/may08/monnich/05monnich.html.

Novotny, E. (2004). 05. I Don’t Think I Click: A Protocol Analysis Study of Use of a Library Online Catalog in the Internet Age. *College & Research Libraries, 65*(6), 525-537.

Voorbji, H. K. (1998). Title keywords and Subject Descriptors: A Comparison of Subject Search Entries of Books in the Humanities and Social sciences. *Journal of Documentation 54*(4), 466-476.