IDENTIFYING EMERGING INFORMATION NEEDS OF LIBRARY USERS USING DATA MINING

Aditi Pawde
Shubhada Apte
Kishore Ingale
Manoj Apte
Girish K. Palshikar

Ms. Aditi Pawde
Researcher, TCS Research,
Tata Consultancy Services
Limited, Pune – 411013
(Maharashtra)
Email: pawde.aditi@tcs.com
Corresponding Author

Ms. ShubhadaApte
Manager,
Tata Consultancy Services
Limited, Pune–411038
(Maharashtra)
Email: shubhada.apte@tcs.com

Dr. Kishore Ingale
Asst. Gen Manager,
Tata Consultancy Services
Limited, Pune–411044
(Maharashtra)
Email:
kishore.ingale@tcs.comand
dumbkishor@gmail.com

Mr. Manoj Apte
Senior Scientist, TCS Research,
Tata Consultancy Services
Limited, Pune – 411013
(Maharashtra)
Email: manoj.apte@tcs.com

Mr. Girish K. Palshikar
Principal Scientist,
TCS Research, Tata
Consultancy Services Limited,
Pune – 411013 (Maharashtra)
Email: gk.palshikar@tcs.com

Library management systems in modern and large libraries capture diverse data about users, including the search queries that user make over the collection. Information needs of the users of information resources change over time, due to changing user base, changes in technology landscape and evolving business environment. To avoid user dis-satisfaction, it is crucial to accurately estimate users' information needs in a continuous and systematic manner. In this paper, the authors propose a simple data mining technique that analyzes the past user queries and accurately identifies under-provisioned and under-stocked information needs, so that the required information resources can be procured by the library, leading to improved user satisfaction. Further, the authors describe the results obtained by using the technique on a real-life dataset of actual user queries made to the LMS in a large multi-national IT company.

Keywords: Library user query analytics, Library user needs, Library procurement and provisioning, Library analytics, Data mining in libraries, Artificial Intelligence in Libraries.

INTRODUCTION

Modern computing technologies and the Internet have deeply impacted the organization and functioning of modern libraries, most of which are now equipped with sophisticated *Library ManagementSystems* (*LMS*). An LMS helps in managing large and geographically distributed libraries under a single umbrella, and support automation of many library management functions. Library users interact in different ways with an LMS to make effective use of the information resources (books, journals, magazines, reports, patents, CDs, audio-books, e-books etc.) available in the library.

While there are other touchpoints with library users, one of the most important way users interactwith LMS is through *user queries* (or just *queries*, for short), where a user searchesoverthe existing information resources in the LMS by giving an ordered sequence of single *keywords*(also called as *search terms*) or multi-word *keyphrases*, optionally connected with *logical connectives* like AND or OR. A *keyphrase* is an ordered sequence of multiple words where all words in the phrase are required in order to capture the complete meaning. A keyphrase is typically a sequence of nouns (including proper nouns and nouns denoting domain-specific terms), or a sequence of adjectives and nouns, although verbs may sometimes occur in a keyphrase, as in estimating project expenditure. In linguistic terms,keyphrases are typically noun phrases or verb phrases. Some examples of user queries are Python machine learning, Oracle cloud, Python AND data structures, Android design patterns.Here, Python and Android are keywordsand Artificial intelligence, data structures are examples of keyphrases.

Today's libraries work in a user-centric environment, where understanding of users' information needs is crucial. These needs keep changing over time, over the changes in the user base and over evolving environment, technologies and business needs. As an example, information needs of the users of IT technologies related information resources change over time; for example, information needs for COBOL reduce, and are replaced by needs related to Java, as the IT technology landscape evolves. As new technology areas – such as Internet of Things, Cloud Computing, Data Science, Machine Learning, and Artificial Intelligence – come up, the users seek information resources related to these, which may not be adequately available in the library. This lack of provision affects user satisfaction and even their work. This typically reflects in lower user satisfaction index and that in turn often leads to lower performance ratings for librarians' in appraisals.

Thus, it becomes crucial to accurately estimate these users' information needs in a continuous and systematic manner. In particular, library managers need to identify the upcoming and buildingup information needs which are under-provisioned or under-stocked. An *under-provisioned information need* is one which cannot be adequately satisfied by existing information resources in the library. Thus, if a user query *Q*does not return any existing resource fully matching the information need expressed *Q*then one can say that the query *Q*contains an under-provisioned information need. Alternatively, if too many user queries are made about a few information resources, then this may lead to many users being denied access to them, leading to *under-stocked information need*. Once the under-provisioned and under-stocked information needs are accurately estimated, the required information resources can be procured by the library, leading to improved user satisfaction.

RELATED WORKS

Data mining techniques such as clustering, classification, regression and association rule mining are widely used to improve quality,

services, collection and usage behavior of Library Management Services (Siguenza-Guzman et al., 2015). Mainly, association rule mining has been used in multiple occasions for identifying the library user's (i.e., reader's) personal choices of the books. Knowledge of such preferences has then been used for various downstream purposes. Li and Chen (2008) use association rule mining of the books circulation and user needs in library automation system. These rules help the library to optimize resources in terms of purchase of number of books and copies thereof and improve the layout of the library. This knowledge in terms of the discovered rules is also used for personalized recommendations to the readers and for identifying potential needs of the readers and for planning for their needs accordingly. Xia and Liu (2018) attempt to grasp the needs and interests of the library users using association rule mining. This information is used in decision making, suggestions on book purchasing, subject construction and collection distribution.

In order to keep up with user interests, libraries rely on basic level of recommendation services for users. Linz et al., (2017) has designed a book recommendation system with the user based and book based collaborative filtering technology. The work analyzes the associations between the readers' borrowing preferences and books and then derives the reader's borrowing rules. These rules are used for building a recommendation function for the readers.Kovacevic et al., (2010) proposed recommending library services to users using *k*-means clustering on user profiles. A *clustering* technique groups users having similar (dissimilar) search behavior into the same (different) "bucket" (also called a *cluster*). Once the clusters are identified, then a rule is generated that describes each particular cluster. Generally, users falling in one "bucket" tend to prefer similar library services. Naïve Bayes classifier maps any new search request to one of the clusters and recommends library service accordingly. However, if book collection is not managed properly, users end up borrowing undesirable books and the recommendation systems may fail. Luo (2017) has modified the well-known DBSCAN clustering algorithm to cluster borrowed books and books in the library. The results of this technique can be used to improve structural and hierarchical distribution of library resources.

Academic libraries are required to prove their impact on overall growth of students and faculty in order to keep up with an institution's goals. Renaud et al., (2015) have proposed a model based on data mining techniques that performs user behavior and book usage analysis. The data is first consolidated in a single repository. Then various data like the data of users (department wise), book checkout data, book circulation data is used for user behavior and book usage analysis. These are then correlated with the students' Grade Point Average (GPA) to analyze the relation between library usage and students' academic performance. Similarly, data from various sources such as patron data, circulation data, book data and student data is consolidated in Silwattananusarn and Kulkanjanapiban, (2020) study. Association mining and clustering is applied on this data.Interesting behavior patterns and

relation between categories of books borrowed and Cumulative Grade Point Average (CGPA) are observed in the data. The mined patterns help librarians to take measures such as setting budget, improve use services, provide recommendations, designing of book shelves etc.Another study improved traditional association rule mining using Bayesian algorithm (Xu, 2020) .This algorithm is validated using simulated data of 1000 students. The study shows that this new improved algorithm provides better recommendations for students based on their major when compared with recommendations by traditional association rule mining and collaborative filtering. Objectively, results of this algorithm has higher accuracy, recall rate and F-score.

The state-of-the-art literature is focused on understanding user behavior patterns and providing recommendations. There is no study that focuses making library systems future ready considering user information needs based on their search pattern. The recommendation systems recommend available resources to users. However, they fail to analyze need of any new resources that might be of interest by users.

PROBLEM FORMULATION

In this paper, a historical database of user queries in the geographically distributed libraries of a large multi-national IT company is used to estimate under-provisioned information needs; under-stocked information needs can be discovered in a similar manner. The key assumption in the proposed approach is that *past user queries are a reasonable indicator of emerging information needs*. Any user query that returns less than or equal to k_0 information resources is represented as an under-provisioned information need (in this paper, the authors use $k_0 = 5$).

The key challenge is to map such under-provisioned user queries to topics and sub-topics which can be used for procurement and classification. Two types of systems are available to *classify* an information resource (such as a book) i.e., to associate appropriate categories to it.

1. A single category from a particular cataloging system commonly used in a library can be used (Joudrey et al., 2015); e.g., Dewey Decimal (DDC) and Library of Congress (LC). The cataloging systems like DDC or LC help in organizing library collections into subject-related categories, so that information resources on the same topic are put together on the shelf. In the proposed approach, it is possible to map a user query to one or more categories in the DDC or LC system.
2. Many subjects have domain-specific subject hierarchies, such as ACM Computing Classification System (Coulter, 1997) , Mathematics Subject Classification (Rusin, 1999), Physics Subject Headings (https://physh.aps.org) and Medical Subject Headings [Nelson, 2009].Such subject hierarchies work well to assign appropriate keywords to research paper within that domain. Hence, it is possible to map a user query to one or more subject headings in an appropriate subject

hierarchy. One issue here is that it may not be always known which subject hierarchy to use for a particular user query; this needs to be figured out automatically.This is because libraries contain information resources from many different disciplines.

In either scenario, techniques known as *ontology matching*can be used that maps a user query to one or more category labels in a given hierarchy. See (Shvaiko and Euzenat, 2011)for a review of the many algorithms that have been designed for ontology matching. Also, Python implementations of many ontology matching algorithms are available, which we can easily use.A list of ontology matching tools is also available.

A different approach is proposed here than mapping user queries to either DDC or LC or to domain-specific subject headings. As stated earlier, the goal of the system is to help library managers understand the demand patterns for under-provisioned user queries. Once these patterns are understood, the library managers can then identify appropriate information resources (e.g., books) that can be procured to satisfy these queries. Once the exact information resources are identified, suitable classification code can be extracted from the meta-data associated with that resource.

Thus, the hypothesis is that *identifying a set of (topic, sub-topic) from a set of under-provisioned user-queries is sufficient for the purposes of emerging information needs*. Broadly speaking, an acceptable topic T should refer to a wider field within a subject and an acceptable sub-topic U_T for T should be a sub-field of study within T. As an example, suppose we identify the (topic, sub-topic)pair (SAP, hybris) from a set of under-provisioned queries. Then the library managers can look for suitable information resources to meet this information need. They may identify books such as (Singh, 2019), from which they can get the DDC or LC classification details for that resource.

DATASET DESCRIPTION

In this paper, the database used is of user queries of the geographically distributed libraries of a large multi-national IT company. There were 72 libraries distributed across the large cities in India and they catered to the overall user-base of approximately 300,000 software professionals and researchers in 2017 (current user-base is about 4,20,000). In 2017, the libraries had a total of approximately 1,20,000 books. About 45,000 users borrowed at least one book in June 2017, and approximately 3,20,000 books were issued in June 2017.

The users made a total of 52,392 queries to this LMS in June 2017. As mentioned earlier, any user query that returns £k_0 information resources is treated as an under-provisioned information need (in this paper, k_0 = 5). The Figure 1 shows the distribution of the number of results returned for these user queries. As seen, there are 32,075 user queries which contain under-provisioned information needs. The LMS allowed queries to include bar-code or ISBN number; such queries typically result in an exact match with an information resource; that is why a large number of queries return exactly 1 result. After removing queries containing a bar-code or an ISBN number,

it was left with 31,907 queries, out of which 12,668 (39.7%) queries were under-provisioned i.e., returned 5 or less results. Note that some of the queries return few results because reasons like presence of mis-spelt words (e.g., soluytion) in the query. Note also that an under-provisioned query may be branch-specific; e.g., a query may result in returning many resources in the library of one location (city) and very few or none in the library of another library location.

The question to be addressed is: how to identify and succinctly describe these information needs, which are hidden inside these large number of under-provisioned user queries? The answer is: to identify a set of (topic, sub-topic) pairs from these 12,668 under-provisioned queries.

Table 1: Distribution of the results returned by the LMS for user queries

Number of results found	Number of queries
0	6553
1	19505
2	2413
3	1477
4	1163
>5	20895

The summary statistics for the length of the queries (in terms of number of words in each query) are as follows: minimum:1, maximum:43, average:2.53, standard deviation:1.87, first quartile:1, second quartile (median):2, third quartile: 3. This means that on the average a query contained 2.53 words, and 75% of queries had 3 words or less. The Figure 2 shows the distribution of query lengths; as seen there is a sharp drop in the number of queries containing about 5 or 6 words, although the distribution shows a long tail.

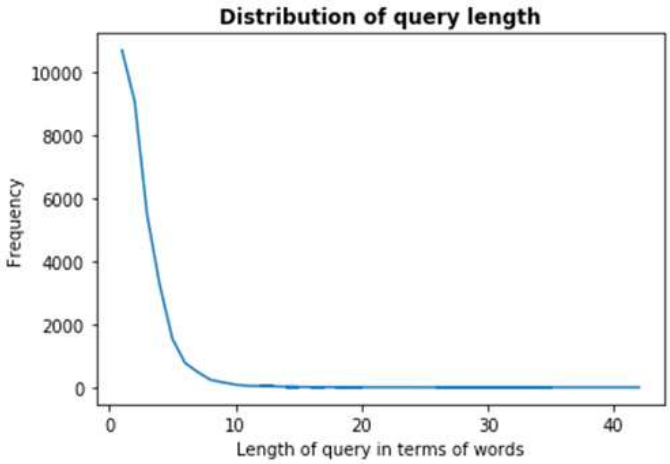

Figure 1. Distribution of the number of words in user queries.

SOLUTION

There are already 12,668 under-provisioned queries identified in our dataset. So,the first goal is to automatically discover a set of (topic, sub-topic) pairs from this set of under-provisioned user queries. Note that the set of topics and the set of sub-topics are not pre-defined and are not given; each of these pairs needs to be discovered automatically. Even the number of such pairs to be discovered is not pre-specified. Note also that even what is an acceptable topic and what is an acceptable sub-topic is also not pre-defined.

The set of (topic, sub-topic) pairs are identified as follows:There are two inputs. First, a set $Q = \{Q_1, Q_2, \ldots, Q_N\}$ of N under-provisioned user queries, and second,$P = \{P_1, P_2, \ldots, P_M\}$ a set of M well-provisioned (not under-provisioned) user queries. Each user query is just a sequence of words (search terms). In fact, the order of the words in a query is ignored and each query is treated as an unordered *bag of words*.

Given a set of queries, and a word w, the *supportcount* (or just *support*, for short) of w,

denoted *sup*(*w*), is simply the number of queries in which *w* occurs. For example, in the dataset, the word SAP occurs in 760 under-provisioned queries, and hence its support in this set of queries is *sup*(SAP) = 760.Given a support threshold s_1, a word *w* is frequent if *sup*(*w*) is greater than s. Thus, if s_1 = 300, then SAP is frequent in the set of under-provisioned queries; note that SAP is *not* frequent if s_1 is 800, for example. It is assumed that a value for $s_1$is known.

The same concept of support is easily extended for an unordered set (pair) of words *u* and *v*: *sup*(*u*, *v*) is simply the number of queries in the given set of queries in which both *u* and *v* occur in any order, and possibly other words between them. For example, in the dataset *sup*(SAP, HANA) = 26, because there are 26 under-provisioned queries in which both SAP and HANA occur. Given another support threshold value s_2, we say an unordered pair of words *u*, *v* is *frequent* if *sup*(*u*, *v*) is at least s_2 or more.

In the first step,all words which are frequent in the set *Q* of under-provisioned queries (for the given value of s_1) are found. Let *T* denote this set of words, which are topic words. The Table 2 shows examples of some topic words found in the dataset.

Table 2: Examples of Discovered Topic Words in the Dataset

Words	Support Count in the set of under-provisioned queries
SAP	760
Data	411
Oracle	331
Java	290
Python	258
Certification	155
SQL	144
Angular	140
Web	139
Js	128
Spring	115

In the second step, a set *R* of pair of frequent unordered pairs of words (for a given value of s_2) is found, where each pair meets the following conditions: (i) the pair should contain *exactly* one word from *T*; and (ii) the other word in the pair (which is not in *T*) should *not* be frequent in *P* (which is the set of well-provisioned queries). As an example, assuming SAP is in *T*, the pair of words (SAP, ABAP) would not be

Table 3: Examples of Topics and Corresponding Sub-topic Acceptable in Set *R*

Topic Word	Example pair	Acceptable	Reason
SAP	(SAP, hana)	No	hanais in *T*
	(SAP, hybris)	Yes	hybris is not in *T* and infrequent in *P*
Data	(Data, analysis)	No	analysis is in *T* and frequent in *P*
	(Data, science)	Yes	science is not in *T* and not infrequent in *P*
Java	(Java, programming)	No	programming is in *T*
	(Java, ocjp)	Yes	ocjpis not in *T* and not infrequent in *P*

acceptable, if either (i) ABAP is also in *T* i.e., ABAP is frequent in *Q*; or (ii) ABAP is frequent in *P*.See Table 3 for some examples.

In the third step, those pairs of words are removed that are found acceptable in step (2) – which are in the set *R* –which are also present in at least k_1 queries in *P* (herek_1 = 1). That is how pairs of words in R if they also occur at least once in the set of queries *P*are removed. Table 4 shows some examples of word pairs that got removed using these criteria.

Table 4: Examples of Pairs Removed using $k_1 = 1$

Topic word	Few examples of removed pairs
SAP	(SAP, Administration), (SAP, Application), (SAP, BI)
Data	(Data, analytics), (Data, management), (Data, mining), (Data, master)
Java	(Java, programmer), (Java, ocm)

The pairs of words which remain in *R* after this pruning step are the desired (topic, sub-topic) pairs. The Table 5 shows some examples in the dataset. For example, row 1 contains the following (topic, sub-topic) pairs: (SAP, odata), (SAP, hybris), (SAP, platform), (SAP, ui5), (SAP, gateway), (SAP, grc), (SAP, ibp).

Table 5: Examples of Discovered (Topic, Sub-topic) Pairs in Our Dataset

Topic	Sub-topics
SAP	odata, hybris, platform, ui5, gateway, grc, ibp
Data	integrator, scratch, discovering, governance
Oracle	integrator, relationship, tips, 1z0071
Java	se8, 1z0809, algorithm, ocjp, driver, 1z0808
Python	language, crash
Certification	ocp, 1z0809, scrum, aws, togaf, 1z0071
SQL	implementation, maintenance, mcts
Angular	typescript
Web	selenium, driver, aspnet
Js	react
Spring	microservice

These (topic, sub-topic) pairs succinctly represent the unfulfilled information needs hidden inside the user queries. These results can now be used to search for and procure appropriate information resources that can meet these information needs. The Table 6 shows some examples of the final results obtained by the system on the dataset.The information needs shown there were manually written, but the corresponding information resources (e.g., books) to meet these needs can now be easily identified.

CONCLUSIONS AND FURTHER WORK

Information needs of the users of information resources change over time, due to

Table 6: Examples of User Queries, Associated (Topic, Sub-topic) and the Information Need

User Query	(Topic, Sub-topic)	Information need
SAP fiori implementation	(SAP, fiori)	Resources detailing implementation of SAP Fiori technology
v14 certification teradatasql	(teradata, v14)	Resources to prepare for V14 certification of Teradata SQL
oracle exam guide IZ0-071	(oracle, IZ0-071)	Examination guide for Oracle certification IZ0-071
AWS Administration - The Definitive Guide	(AWS, administration)	Study guide for AWS Administration

changing user base, changes in technology landscape and evolving business environment. To avoid user dis-satisfaction, it is crucial to accurately estimate users' information needs in a continuous and systematic manner. In this paper, the authors proposed a simple data mining technique that analyzes the past user queries to an LMS and accurately identifies under-provisioned and under-stocked information needs, so that the required information resources can be procured by the library, leading to improved user satisfaction. The authors described the results obtained by using the technique on a real-life dataset of actual user queries made to the LMS in a large multi-national IT company.

For future work, the authors are looking at designing more complex data mining methods for identifying emerging information needs. Automatically mapping the discovered information needs to suitable DDC or LC cataloging system or a given subject hierarchy is an important future work for us. The authors can automatically search the web and identify top *k* suitable information resources (e.g., books), along with their price and other details, and estimate the total cost and time that will be required for fulfilling the discovered under-provisioned and under-stocked information needs by the system. Clustering and prioritizing the identified information needs should also help in automatically creating the above procurement plan. Estimating the impact on user satisfaction levels after a given procurement plan is fulfilled also important. The authors are exploring time-series techniques to understand and predict the temporal changes in the users' information needs.Finally, the authors need to work on devising techniques to increasing the user's engagement with the information resources and to incentivize users for improving their usage of the library remain important for the research.

REFERENCES

1. Coulter, N. (1997). ACM's computing classification system reacts changing times.*Communications of the ACM*, 40(12), 111-112.
2. Joudrey, D. N., Taylor, A. G., &Miller, D. P. (2015*). Introduction to cataloging and classification*. ABC-CLIO.
3. Kovacevic, A., Devedzic, V., &Pocajt, V. (2010). Using data mining to improve digital library services. *The Electronic library*, 28 (6), 829-843.
4. Li, J., &Chen, P. (2008). The application of association rule in library system. In 2008 IEEE International Symposium on Knowledge Acquisition and Modeling Workshop, pages 248-251. IEEE.
5. Linz, C., Muller-Stewens, G., &Zimmermann, A. (2017). *Radical business model transformation: Gaining the competitive edge in a disruptive world*. Kogan Page Publishers.
6. Luo, L. (2017). Application of data mining in library-based personalized learning. *International Journal of Emerging Technologies in Learning (iJET)*, 12(12), 127-133.
7. Nelson, S. J. (2009). *Medical terminologies that work: the exampleof mesh. In2009*

10th International Symposium on Pervasive Systems,Algorithms, and Networks, 380-384. IEEE.

8. Renaud, J., Britton, S., Wang, D., &Ogihara, M. (2015). Mining library and university data to understand library use patterns. *The Electronic Library*, 33 (3), 355-372.
9. Rusin, D. A. (n.d).*Gentle Introduction to the Mathematics Subject Classification Scheme*. Mathematical Atlas.
10. Shvaiko, P., &Euzenat, J. (2011). Ontology matching: state of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1), 158-176.
11. Siguenza-Guzman, L., Saquicela, V., Avila-Ordo~nez, E., Vandewalle, J., &Cattrysse, D. (2015). Literature review of data mining applications in academic libraries. *The Journal of Academic Librarianship*, 41(4), 499-510.
12. Silwattananusarn, T., &Kulkanjanapiban, P. (2020). Mining and analyzing patron's book-loan dataand university data to understand library use patterns. arXiv preprintarXiv:2008.03545.
13. Singh, S. K., Feurer, S.,&Ruebsam, M. (2019). SAP Hybris Commerce, Marketing, Sales, Service, and Billing with SAP, Galileo Press Inc.,OCLC Number1002285380, ISBN: 1493215388 9781493215386.
14. Xia, T., &Liu, Y. (2018). Application of improvedassociation-rules mining algorithm in the circulation of university library.International Conference on Big Data and Artificial Intelligence (ICBDAI 2018), 60-64.
15. Xu, S. (2020). Association rule model of on-demand lending recommendationfor university library. *Informatica*, 44(3), 2.

❖❖❖