IFLA WLIC 2013
SINGAPORE

Submitted on: May 31, 2013

Connecting library content using data mining and text analytics on structured and unstructured data

Chee Kiam Lim
Technology and Innovation, National Library Board, Singapore.
E-mail address: chee_kiam_lim@nlb.gov.sg

Balakumar Chinnasamy
Technology and Innovation, National Library Board, Singapore.
E-mail address: balakumar_chinnasamy@nlb.gov.sg

Abstract:

With so much data available, how can busy users find the right bits of information? The venerable search does a great job but is it sufficient?

A typical search at a popular Internet search engine returns thousands of results for a user to sieve through. This tedious process continues after finding a relevant article to find the next one and the next one until the user's information needs are satisfied (or they give up).

Instead of having users repeat the tedious search and sieve process, we should push relevant information packages to them. And to do this, we must connect our content.

The advances in Big Data technologies present significant opportunities for us to connect the huge and growing amount of information resources in our repositories.

By leveraging data mining and text analytics techniques, and Big Data technologies, the National Library Board (NLB) of Singapore has connected our structured and unstructured content. This has allowed us to provide comprehensive, relevant and trusted information to our users.

Keywords: data mining, text analytics, Mahout, Hadoop, National Library Board (NLB) of Singapore.

1 CONNECTING STRUCTURED DATA

The extensive network of libraries of the National Library Board (NLB) of Singapore has a collection of over a million physical titles. These titles generate over 30 million loans annually.

By using data mining techniques on past loan transactions and the bibliographic records of the books, NLB has successfully connected our titles and launched our own title recommendation service since 2009.

Our title recommendation service can be found on many of NLB's websites and portals. These include:

- "Read On" (http://readon.sg)
  - "Read On" is NLB's search-engine friendly micro-site that allows search engines to easily crawl and index NLB's collection of books

Figure 1.1: Title recommendations on NLB's "Read On" micro-site

- "Library in Your Pocket" (LiYP) (http://m.nlb.gov.sg)
  - LiYP is NLB's mobile library portal

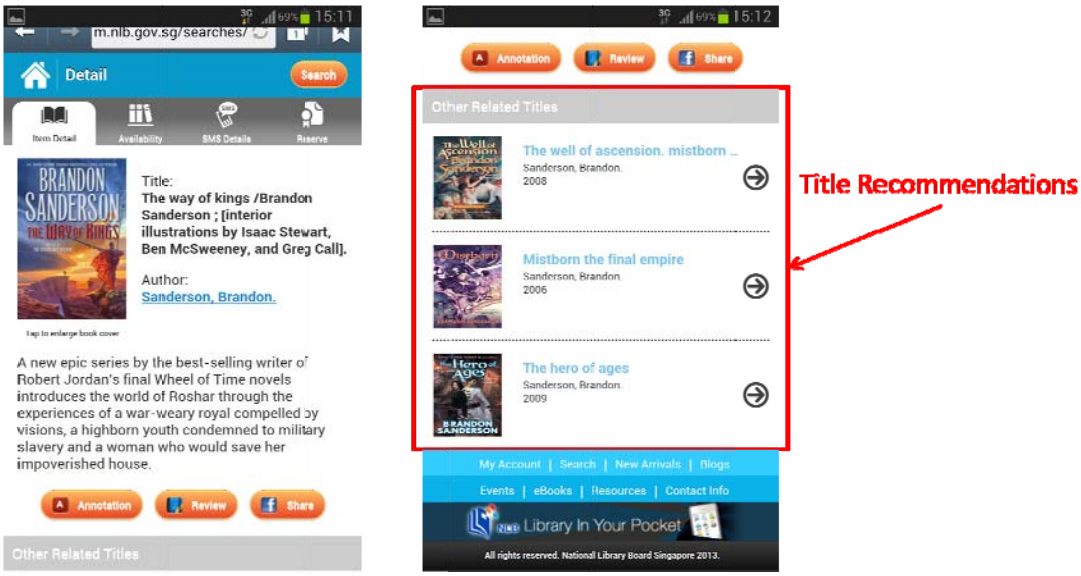

Figure 1.2: Title recommendations on NLB's LiYP mobile site

- "SearchPlus" discovery portal (http://searchplus.nlb.gov.sg)
    - "SearchPlus" is NLB's discovery portal for both physical titles and digital resources, including e-books and e-Databases

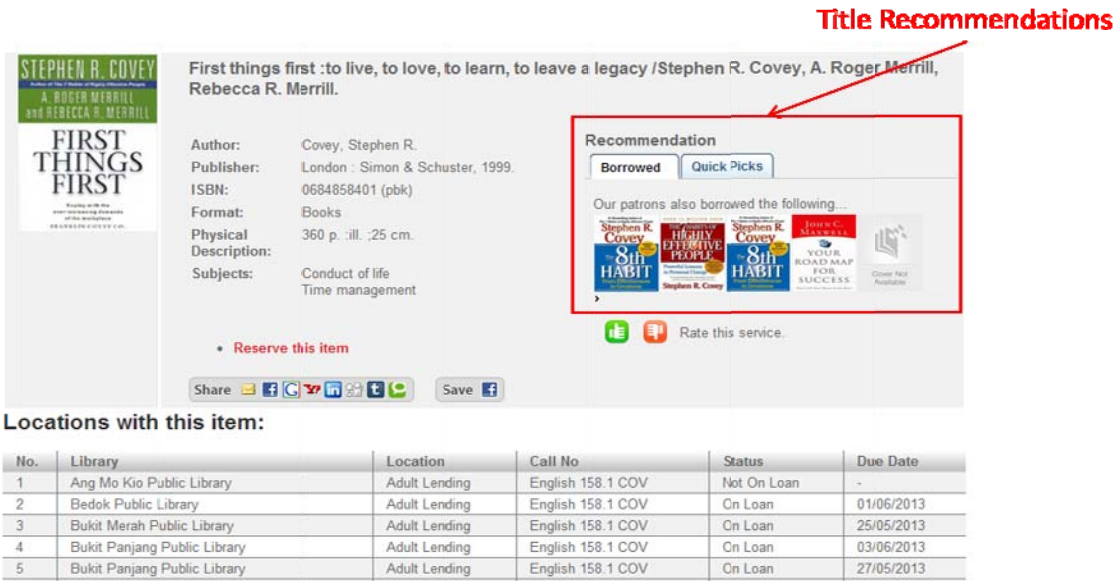

Figure 1.3: Title recommendations on NLB's "SearchPlus" discovery portal

- Online "Check Your Account" patron service on NLB's Public Library portal (http://www.pl.sg)
    - NLB's Public Library portal's "Check Your Account" service allows our patrons to check their account information such as loan items, due dates and outstanding fines

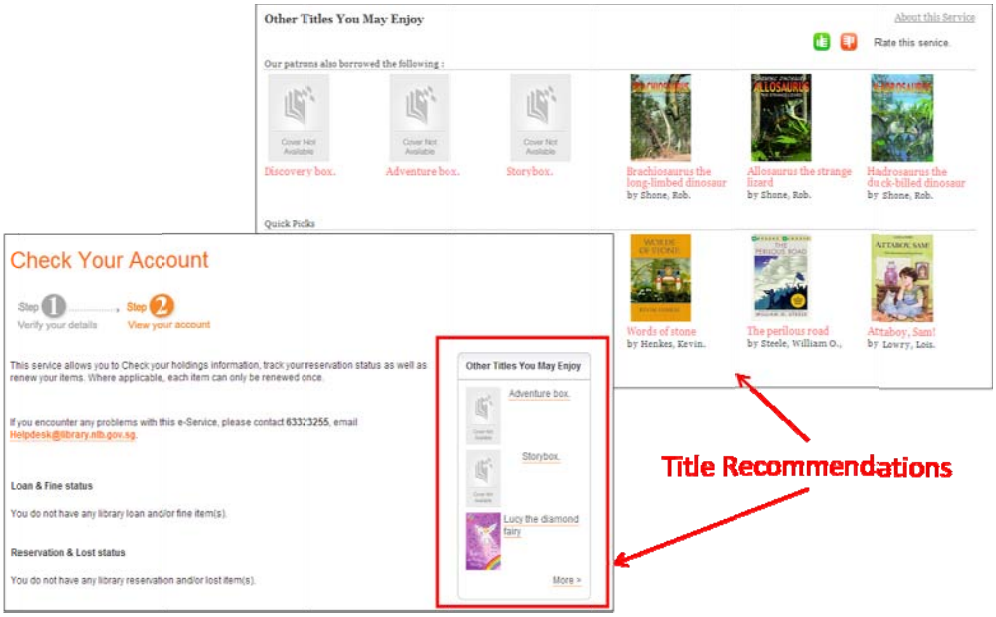

Figure 1.4: Title recommendations on NLB's Public Library Portal

The title recommendation service is also available for third party developers to use through the NLB Open Data/Web initiative (http://nlblabs.sg).

Data mining is the process to discover patterns in large data sets. The following describes the approach and process taken by NLB.

NLB's title recommendation service implementation is a hybrid recommender system that combines both collaborative and content-based filtering.

Given a title, the title recommendation service generates 2 types of recommendations: "Our patrons also borrowed the following" and "Quick Picks".

Figure 1.5: "Our patrons also borrowed the following" and "Quick Pick" recommendations

The "Our patrons also borrowed the following" category primarily relies on collaborative filtering for recommendations. Collaborative filtering mines the reading patterns within the hundreds of millions of loan records to make recommendations. It relies on the notion that we tend to obtain recommendations from people with the same interests. Only loans in the last 3 years are considered when generating the recommendations to minimize recommendations that are outdated.

In cases where collaborative filtering fails due to low or no loans, the algorithm falls back to content-based filtering using the bibliographic records.

It is therefore a robust and fine-grained recommendation system that takes into account the actual loans by patrons, and as a result, will also adjust its recommendations if there are shifts in patron reading patterns.

Figure 1.6: Collaborative and content-based filtering

"Quick Picks" recommendations rely solely on content-based filtering to come up with recommendations using the bibliographic records. However, these recommendations are made from a list of books selected by librarians instead of the whole collection.

The patron level recommendations are currently implemented on top of the title level recommendations. The patron's latest loaned titles are used to provide recommendations which are then merged into a single list.

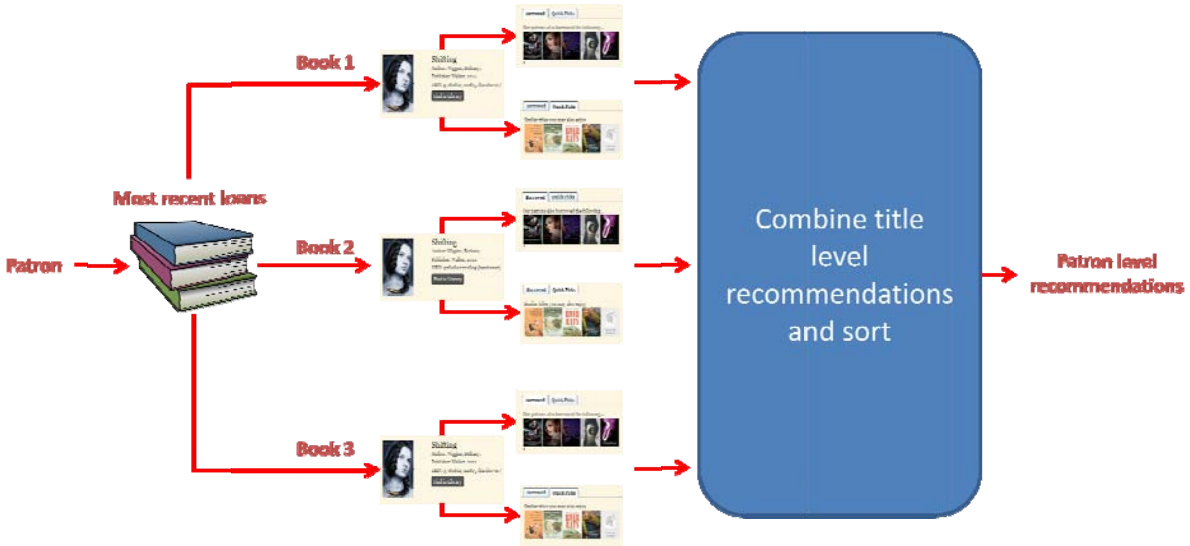

Figure 1.7: Patron level recommendations

Table 1.1 shows some statistics on recommendation coverage.

Fiction titles are highly loaned and therefore collaborative filtering generates recommendations for 59% of the total fiction titles compared to only a 28% success percentage for the less loaned non-fiction titles.

	Collaborative filtering	**Combined with simple content-based filtering**
Fiction	59%	89%
Non-Fiction	28%	53%
Overall	34%	59%

Table 1.1 NLB's recommendation coverage

The simple content-based filtering applied here uses only author and language fields. As can be seen, even combining collaborative filtering with a very straightforward content-based filtering algorithm significantly improves the recommendation coverage, raising the fiction success percentage from an acceptable 59% to an excellent 89% and non-fiction from a low 28% to an acceptable 53%.

The recommendation coverage for patron level recommendations is even higher since the recommendations are the combination of the title level recommendations based on the patron's recent loans.

2 CONNECTING UNSTRUCTURED DATA

Unstructured data makes up a huge and growing portion of the content that NLB holds. Connecting these data will allow us to make these contents a lot more discoverable.

One of our earliest attempts to connect such content involves using text analytics to perform "key phrase extraction". The extracted key phrases are then used to perform searches which throw up other related content. These connections ensure that further discovery around the key topics are just 2 clicks away.

The NLB "Infopedia" search-engine friendly micro site (http://infopedia.nl.sg) is a popular electronic encyclopedia on Singapore's history, culture, people and events. With less than 2,000 articles, it attracts over 200,000 page views a month.

It makes use of this approach alongside manual recommendations from the librarians. These manual recommendations by librarians are added as Librarian's Recommendations and provides a 1 click access to related content.

Figure 2.1: Infopedia connections through extracted key phrases and manual recommendations

However, as manual recommendation is a labour intensive and time consuming task, not all articles have Librarian Recommendations.

With most other collections many times larger than the Infopedia collection, NLB needed an alternative approach to automatically generate good recommendations for its unstructured content.

The Apache Mahout software is a scalable open source software from the Apache Foundation that implements a wide range of data mining and machine learning algorithms (http://mahout.apache.org). It supports four main use cases, namely recommendation mining, clustering, classification, and frequent item set mining.

Using text analytics in Apache Mahout to find similar articles, we successfully generated recommendations for the full Infopedia collection in less than 5 minutes.

Figure 2.2: Infopedia text analytics result sample

Using the same Infopedia article in Figure 2.1 (titled “King Edward VII College of Medicine”), Figure 2.2 shows that 3 of the manual recommendations were among the top 4 recommendations automatically generated by Mahout.

The Mahout text analytics are based on established and rigorous mathematical algorithms (such as Euclidean, Cosine, Tanimoto, Manhattan and many others for distance measurement, term frequency/inverse document frequency, etc.). The recommendations from Mahout are therefore statistically relevant.

These automatically generated recommendations allow us to connect relevant articles for discovery immediately. There is no longer a need to wait for manual recommendations before such connectivity is available.

The following figures from Infopedia in NLB’s staging environment illustrate the capability of the automatically generated recommendations to supplement manual recommendations by librarians. In cases where manual recommendations are not yet available, they enrich the content and immediately allows the discovery of other related content.

Figure 2.3: Supplementing an Infopedia article with manual recommendations

Figure 2.4: Enriching an Infopedia article with no manual recommendations

The Singapore Memory portal's (http://singaporememory.sg) personal memory collection is the next data set that we have successfully completed this similarity processing and we are in the process of pushing it out to enhance the current content-based filtering recommendations.

Moving on from Infopedia and Singapore Memory, we looked to scale it up and successfully completed another proof-of-concept (PoC) for about 58,000 newspaper articles from our NewspaperSG collection.

NewspaperSG (http://newspapers.nl.sg) is NLB's online resource for Singapore and Malaya newspapers published between 1831 and 2009. NLB has invested significantly to digitise the newspapers at the article level, and there are now over 18 million articles in NewspaperSG.

Scaling up from 2,000 to 58,000 was relatively easy. We completed the similarity processing in around 18 hours.

The results proved to be very promising. Interesting, if we were to organise the recommended articles in a chronological order, we can discover the progression of a news item and see the story unfolds.

Figure 2.5: PoC results sample for NewspaperSG (58,000 articles)

With the successful application of text analytics on Infopedia, Singapore Memory and a year's worth of 58,000 newspaper articles, we were encouraged to take the plunge and tackle a much larger data set. We decided to pull all the articles from 2 newspapers and generate recommendations to see if we can discover additional perspectives of a news item when articles from another newspaper appear in the recommendations.

Working on the 6 million NewspaperSG articles from the 2 newspapers (The Straits Times from 1845 to 2009 and The Singapore Free Press from 1925 to 1962) gave us our first real scalability issue. The processing ran for more than a week before we ran out of disk storage.

The processing of NewspaperSG articles also surfaced another challenging issue. The newspaper articles were derived from the use of Optical Character Recognition (OCR) software during the digitisation of the historic newspaper microfilms. OCR errors are common due to the historic nature of the newspapers, particular for the older issues. These OCR errors introduced 'noise' into the data set, and also significantly increased the complexity of the computation, leading to lengthy processing and the need for huge amount of intermediate disk storage.

A reliable, scalable and distributed computing platform is required.

Apache Hadoop is a framework that supports distributed processing of large data sets across clusters of commodity computers using simple programming models (http://hadoop.apache.org). It is the de-facto open source software platform for big data processing, boasting users including Facebook, Twitter, Yahoo, LinkedIn and EBay.

Many of Apache Mahout's core algorithms are implemented on top of Apache Hadoop using the map/reduce paradigm to allow scaling to reasonably large data sets.

After much experimentation and tuning, we finally set up a full Apache Hadoop cluster running Apache Mahout in our secondary data centre. We set up a total of 13 virtual servers on 3 virtual machine hosts.

Figure 2.6: Apache Hadoop setup in NLB

For more information on our setup, do feel free to contact us.

To tackle the issue of OCR errors within the newspaper articles, we tuned the parameters for the text analytics algorithm to ignore infrequent word tokens. This has resulted in the removal of a significant portion of the OCR errors before the actual Mahout processing commenced.

With the Apache Hadoop cluster set up and applying more aggressive algorithm parameters, we were able to generate similarity recommendations for more than 150,000 articles. The aggressive change in the parameter values are instrumental in slicing away a huge chunk of the time required for the processing. From being able to process 58,000 articles in around 18 hours, we are now able to process 150,000 articles in less than 30 minutes.

However, we are still in the midst of resolving issues as we move towards our target of processing the 6 million articles.

3 NEXT STEPS

We are currently working on additional proof-of-concepts with Apache Mahout particularly in the areas of on content enrichment, clustering and classification.

In the area of content enrichment, there are lots of interesting ideas and possibilities for us to explore and conquer. We hope to enrich the content with semantic information so that content becomes connected semantically instead of just textually. We also hope to enrich the content with language translations to explore the possibilities of connecting content in different languages, given that we have four official languages in Singapore.

On the clustering front, working with user-submitted personal memories from the Singapore Memory portal, Apache Mahout clustered the memories into 43 groups (Figure 8). The key terms amongst the memories within each cluster can then be analysed to identify the key theme for the clusters.

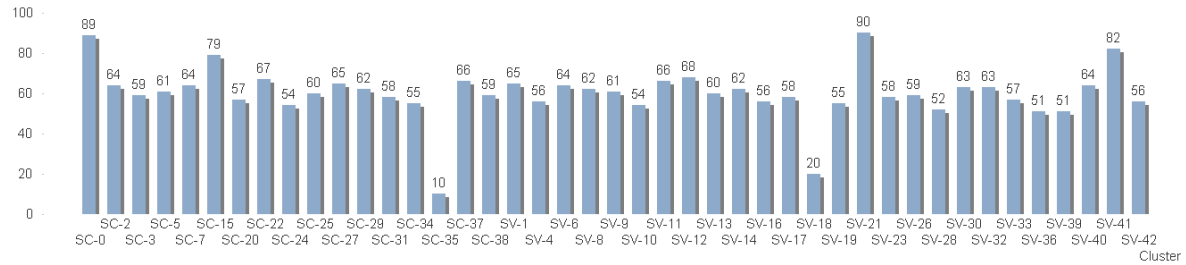

Figure 3.1: Singapore Memory Cluster PoC cluster sizes

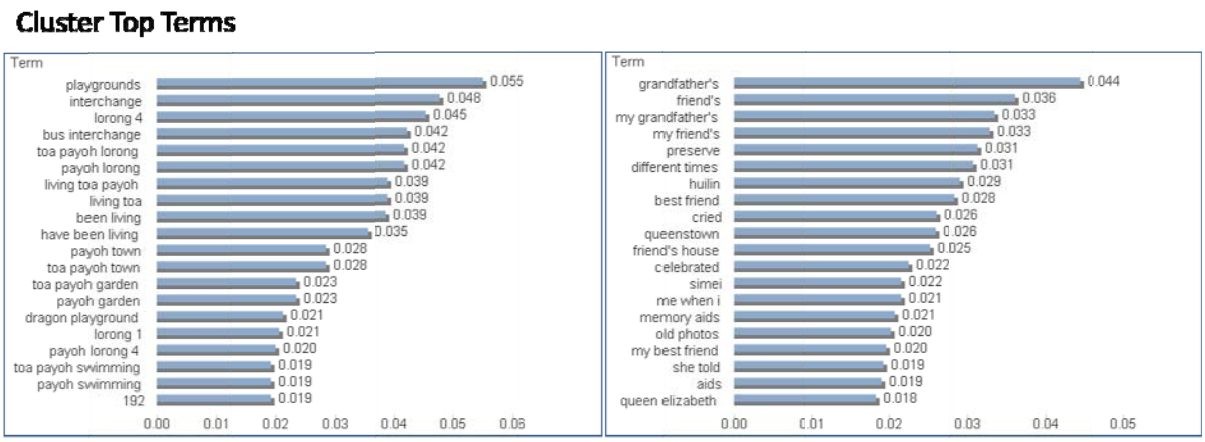

Figure 3.2: Sample of top terms in generated clusters

We are exploring visualization technologies that will allow users to navigate through the generated clusters.

Figure 3.3: Singapore Memory portal's showcase clusters

For classification, we are currently collating data for user-identified clusters. These data will form the training data that the classification algorithms use to automatically classify new user submissions into one of the identified clusters. This will allow us to drastically reduce the amount of effort required to maintain and update the current clusters that are showcased on the Singapore Memory portal.

The value of a library's collections increases with the number of connections within and across the content of the collections, a result of the well understood *network effect phenomenon*. The ability to connect content automatically is therefore essential for libraries to unearth the hidden values in their painstakingly curated collections, especially for sizeable collections that go beyond thousands of items. Data mining and text analytics technology and

software are now more matured and readily available through open source licenses. The time is ripe for libraries to leverage on them to bring out the enormous values in their collections.

There are many more possibilities that we will explore in the future to make more and better connections – connections that allow users to discover comprehensive, relevant and trusted information.

We do the work so that our users do not.