Connecting Users to Content Beyond Keywords : The National Library of Singapore's Experience

Kia Siang Hock, Chan Ping Wah, Ngian Lek Choh

Abstract

The National Library of Singapore shares their recent developments on library services such as mobile services, search service, and digitization of local materials. This paper describes some of the experiments the innovation team has been doing to improve on the search service and to bring related content to users beyond keywords.

The experiments include data mining experiments, text analytics experiments and linked open data. Data mining experiments mines the data from the NLB's records to generating better results. Text analytics experiments use linguistics and/or statistical techniques to extract concepts and patterns from text based documents. It would be very useful to use text analytics to bring related articles to users when they conduct a search. Linked open data can be understood and used by both people and machines. With the categorisation of data based on the content created by the library, machines can create linkages between the data and the meaning attached. This makes it easier for the data to be discovered, used and reused for a variety of purposes.

Keywords: National Library of Singapore; Linked open data; Data mining; Text analytics

Kia Siang Hock: Solution Architect, National Library of Singapore
Chan Ping Wah: Assistant Chief Executive, National Library of Singapore
Ngian Lek Choh: Director-general, National Library of Singapore; E-mail: lekchoh@nlb.bov.sg

1. Introduction

The National Library Board (NLB) is a system of 40 libraries. This comprises the national library, 24 public libraries and 15 special libraries including the Parliament Library and one Polytechnic library.

In the past 17 years, since the formation of the NLB as a statutory board, the library has been focusing on bringing libraries closer to the public so that they can benefit from the library's services and collections more conveniently.

In recent years, more effort has been put into bringing library services and collections to the public through new channels such as mobile phones, and social media so as to reach users more effectively. Singapore content such as local newspapers, photographs, manuscripts and rare books have been digitised, and these have been well used by the general public, students and researchers.

In the past two years, increasing effort has been put into improving the search service, as we found that as we digitise more materials, it is not easy to index nor catalogue everything that we have collected or digitised, due to the large size of the collection items.

This paper describes some of the experiments the innovation team has been doing to improve on the search service and to bring related content to users beyond keywords.

2. Library Blueprints

Like most other libraries, the NLB started its operations using physical library cards. Borrowing and returning of books and other library materials was through a physical library counter. Users had to queue for up to 1.5 hours for library transactions on very busy days, and soon, there were many complaints from users, and use of libraries was adversely impacted.

The first blueprint of the NLB called Library 2000 focused on building more libraries to meet the needs of residents living in new housing estates, and automating library transactions using the radio frequency tagging of library materials. Ten new shopping mall libraries were set up in new housing estates and almost every library transaction was automated. These included borrowing and returning of books, renewals, payment of fines, checking of outstanding loans and registration of new members.

The second blueprint, called Library 2010, focused on digitisation of Singapore content and the building of the digital library infrastructure. This enabled the NLB to reach out to the digital library users, including born digitals who would go first for digital information before anything else.

NLB is currently implementing its third blueprint “Library 2020” which includes a project on Libraries of the Future. Library 2020 focuses on strengthening reading and information literacy skills, reinventing physical libraries to meet changing needs, bringing Singapore content out to users and building on the digital library.

3. Recent Developments

In line with the NLB vision to continually extend reach, ensure relevance and usefulness to its stakeholders, the NLB started working on improving its search capability in a more intensive manner in the past few years.

With the digitisation of newspapers, the NLB innovation team realised that it was not possible for the library to catalogue or index every newspaper article as these are in the tens of millions. The Optical Character Recognition (OCR) technique was used to enable search and find of the newspaper articles through internet search engines. This has proved to be very successful. Today, most users of the NLB newspapers discover the library’s digitised newspapers through internet search engines.

Search of the content however is not precise, and search results often included content that may not be relevant. Also, useful and related content is not brought to the users’ attention, and users are likely to miss these related useful content. This is an area we felt we could do more to help users.

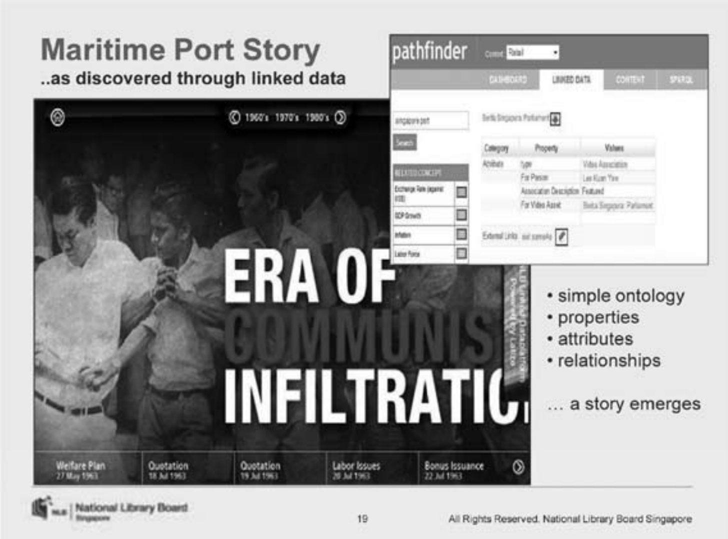
4. Bringing Related Content to Users

Today’s information world is overloaded with too much information. Users are often so busy with their lives that they do not have time to find the information that they need for their work and daily decisions. If libraries can find ways to push relevant information to users based on their information needs, users will appreciate the value of libraries more. To do this, libraries have to build a network of content that relates to each other.

5. Linked Open Data

Linked open data (LOD) is one area that the NLB is looking into actively. In early 2012, the NLB did a small trial on linked open data. Based on a few newspaper articles from Singapore's early years of independence in 1965, dates, events and quotes from founding fathers were pulled into a database. These linked open data were linked to create a story on the Singapore Maritime Port.

In this Proof of Concept or POC, with data from the newspaper articles, NLB catalogue records and external sources such as from the National Archives of Singapore, and The British Library, the NLB team created a simple ontology to categorise the linked data. From there, the story of the Singapore Maritime Port was created.



Maritime Port Story

..as discovered through linked data

Category	Property	Values
Attribute	type	Video Association
	For Person	Lee Kuan Yew
	Association Description	Featured
	For Video Asset	Berita Singapura: Parliament
External Links	owl:sameAs	

- simple ontology
- properties
- attributes
- relationships

… a story emerges

National Library Board Singapore

19

All Rights Reserved. National Library Board Singapore

NLB's linked data POC included full text mining. Names of organizations, people involved and events were extracted from newspaper articles, including the metadata. These were converted into an open data standard. The resulting database comprises terms which can be understood and used by both people and machines. With the categorisation of data based on the content created by the library, machines can create linkages between the data and the meaning attached. This makes it easier for the data to be discovered, used and reused for a variety of purposes. They can be used by librarians, end-users and the industry to create new services and useful applications, including for smartphone users.

6. Data Mining experiments

In addition to experimenting with LOD, NLB also started to look into the use of data mining and text analytics to connect content. Below we share our experimentation with these techniques.

One of the first data mining applications NLB implemented was the “Read Alike Recommendation” engine. This is similar to the Amazon’s recommendation service “Customers who bought this item also bought”.

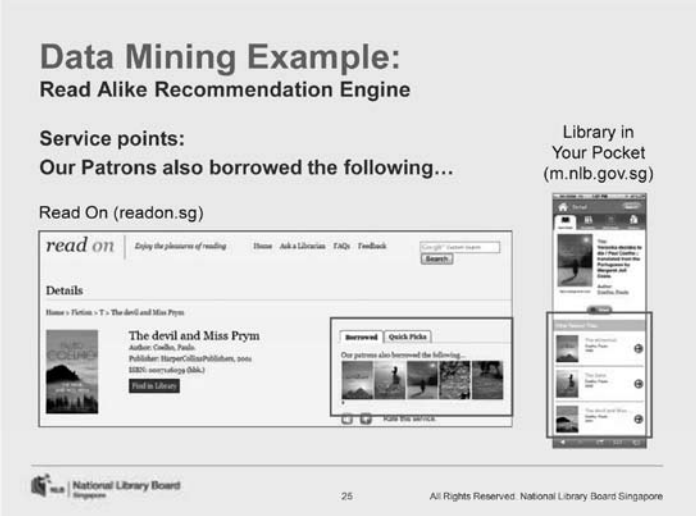
The NLB’s Read Alike Recommendation experiment is based on the “collaborative filtering” technique. It mines the larger volume of user borrowing activities from the NLB’s loan records to identify titles that are related.

Currently, the NLB has a user database of over 2 million. Last year, NLB users generated around 36 million loans. With such a rich set of data, NLB is in an unique position to extract from the reading habits of Singaporean users, and to provide the most relevant recommendations in the local context.

To get into a little more details, we describe the steps here. For every title that the library owns, the team identifies all the users who had borrowed this title over a period of time. In this illustration, there were 1,070 users who borrowed this title. The team looked at all the other titles borrowed by these users over the same period of time, and make the recommendations. For the recommendations, the most popular titles were selected to make the recommendations.

With this relatively simple approach, fairly good results were generated. The recommendations are generated once a month. It runs for about 10 hours for the entire NLB collection of over 1.5 million titles. This service has been implemented in a few

NLB libraries. The service is called “Read On”.



Data Mining Example:

Read Alike Recommendation Engine

Service points:

Our Patrons also borrowed the following…

Read On (readon.sg)

Library in Your Pocket (m.nlb.gov.sg)

National Library Board Singapore

25

All Rights Reserved. National Library Board Singapore

read on
Enjoy the pleasures of reading
Home
Ask a Librarian
FAQs
Feedback
Search
Details
Home › Fiction › T › The devil and Miss Prym
The devil and Miss Prym
Author: Coelho, Paulo
Publisher: HarperCollinsPublishers, 2001
Find in Library
Borrowed
Quick Picks
Our patrons also borrowed the following...
Rate this service

Another NLB service that implemented the title recommendation service is the “Library in Your Pocket” or LiYP. LiYP is the NLB’s mobile web based application that is very popular with users who are constantly on the move.

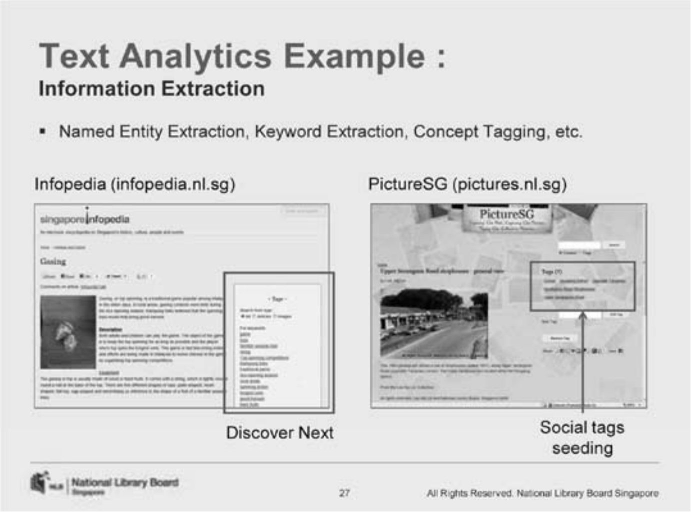
Apart from the title level recommendation, NLB also implemented the user level implementation. For this experiment, the NLB looks at the loan history of each user, and recommend titles that the user might want to borrow.

7. Text Analytics experiments

The innovation team also experimented with text analytics. Text analytics is the use of linguistics and/or statistical techniques to extract concepts and patterns from text based documents.

The first example that is described here is taken from the NLB’s popular Singapore Infopedia microsite which is a database of all things Singapore. To allow users to discover more related content within Infopedia and other NLB collections, the team extracted the most important key terms from each article. These key terms are displayed beside each article, and users can click each term to discover other related content.

Another example is what the team did with the PictureSG service. PictureSG contains digitised images of Singapore from its early days. A key feature of the service introduced is social tagging. Instead of starting the social tagging feature from scratch, the team used the same information extraction capability to extract key terms "seed" the social tags.



Text Analytics Example :

Information Extraction

- Named Entity Extraction, Keyword Extraction, Concept Tagging, etc.

Infopedia (infopedia.nl.sg)

PictureSG (pictures.nl.sg)

Discover Next

Social tags seeding

National Library Board Singapore 27 All Rights Reserved. National Library Board Singapore

A POC that has been completed in September uses text analytics to identity related content. The example that we use is on a CNN news article on 22 Sep 2012 entitled "iPhone 5: the wait is OCR". At the end of the article is a section called "We Recommend", showing other related articles.

With the digitisation of newspapers, users can access over 15 million articles from 30 newspapers in English, Malay and Chinese from 1831 to 2009 from anywhere 24x7. The team felt that it would be very useful to to use text analytics to bring related newspaper articles to users when they conduct a search.

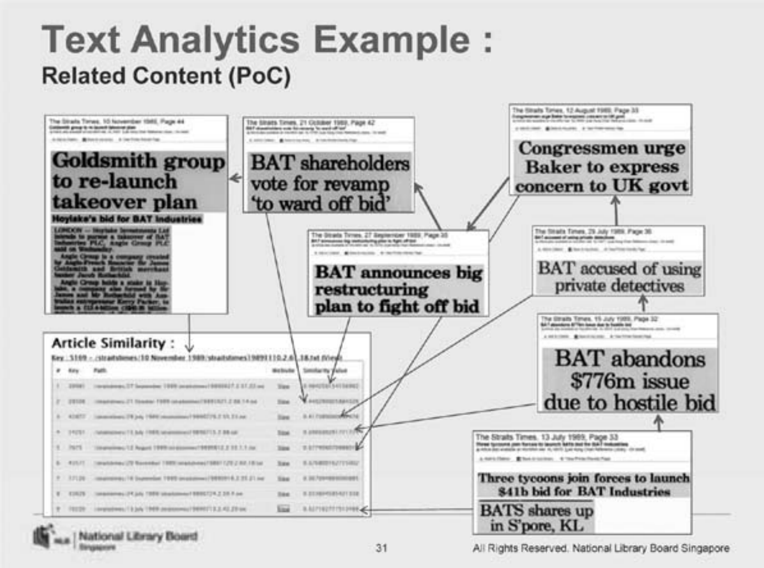
8. Text Analytics POC on Newspaper Articles

For the POC, the team extracted 85,000 articles from one year's worth of newspaper articles from Singapore's major daily "The Straits Times". These articles were pushed through the mahout software which is an open source software that does machine learning and data mining. It is available from the Apache Software Foundation.

Through mahout's standard common line interface, the team imported all the OCR-ed text into the mahout sequence files, an internal respresentation. The sequence files are then analysed and processed into weighted vectors. Here, tokenisation and stop word processing are done, and the term frequency and j verse document frequency were used to create the weighted vectors.

Similarly, a similarity algorithm is applied to the weighted vectors, and the top similar articles are generated for each article. The POC was done using the out of the box features with no tweaking. This was done using a standard developer notebook, and the processing took less than 5 minutes to complete.

The results of the POC were good. The example shown here is the article "Goldsmith group to re-launch takeover plan". The list of related articles sorted by similarity is shown on the left bottom corner of the results screen. They are indeed related to the same event!



Text Analytics Example :

Related Content (PoC)

The Straits Times, 10 November 1989, Page 44
Goldsmith group to re-launch takeover plan
Hoylake's bid for BAT Industries
The Straits Times, 21 October 1989, Page 42
BAT shareholders vote for revamp 'to ward off bid'
The Straits Times, 12 August 1989, Page 33
Congressmen urge Baker to express concern to UK govt
The Straits Times, 27 September 1989, Page 35
BAT announces big restructuring plan to fight off bid
The Straits Times, 29 July 1989, Page 36
BAT accused of using private detectives
The Straits Times, 15 July 1989, Page 32
BAT abandons $776m issue due to hostile bid
The Straits Times, 13 July 1989, Page 33
Three tycoons join forces to launch $41b bid for BAT Industries
BATS shares up in S'pore, KL
Article Similarity :
#
Key
Path
Website
Similarity Value
National Library Board Singapore
31
All Rights Reserved. National Library Board Singapore

If the articles are arranged in chronological order, it gives a good run down of the whole takeover saga.

Using this experience, the team felt that it would be useful to incorporate a section on recommendations on the newspaper website to bring related articles to users when they search for any topic.

9. Conclusion

The NLB will continue to explore and experiment with data mining, text analytics and Linked Open Data. Additional areas to explore will include how the content in the four official languages used in Singapore can be brought across languages in search results. The team is also keen to explore with concept extraction, to give recommendation based on concepts. We expect this journey to be rather bumpy, however, the results can be very interesting and exciting! If we can help reduce the time taken by users to find the information that they need more speedily, all the effort would be worth it.