Submitted on: 5/30/2014

Inter-connected Network of Knowledge – The NLB Journey

Siang Hock KIA
Technology & Innovation, National Library Board, Singapore
kia_siang_hock@nlb.gov.sg

Yi Chin LIAU
Technology & Innovation, National Library Board, Singapore
liau_yi_chin@nlb.gov.sg

Ian ONG
Technology & Innovation, National Library Board, Singapore
ian_ong@nlb.gov.sg

Abstract:

Libraries and archives collect and provide access to digital resources of significant national, cultural and heritage values. Their collections are painstakingly curated, described and made accessible.

NLB adopted a holistic digital strategy that involves the digitisation of uniquely Singapore content, findability enhancement via search engine optimisation, and multi-screen delivery. The next leap is to link the digitised content into a network to enable contextual discovery.

NLB has successfully created high quality associations amongst its digital resources. Text analytics technologies were used to automatically sieve through millions of items, and cluster related resources together. The result is a massively inter-connected network of knowledge. The users are now able to view and explore related resources within and across collections and media formats. Information is no longer viewed in isolation, but seen as a part of the larger context.

Keywords: contextual discovery, knowledge network, text analytics, digital strategy.

1 INTRODUCTION

The National Library Board of Singapore (NLB) oversees the National Library, the Public Libraries and the National Archives. The NLB's mission is to provide a trusted, accessible and globally-connected library and information service through the National Library and a comprehensive network of Public Libraries.

Through its innovative use of technology and collaboration with strategic partners, NLB ensures that library users have access to a rich array of information services and resources that are convenient, accessible and relevant.

A key function of NLB is to connect people to the country, by connecting the past to the present. As the key memory institution in Singapore, NLB collects and provides access to digital resources of significant national, cultural and heritage values. Akin to the corporate memories of enterprises, the NLB collection constitutes a major portion of the national memory of Singapore.

2 THE NLB DIGITAL STRATEGY

A seismic shift in the information seeking behaviour has taken shape. With the ever-expanding reach of the Internet, users are accustomed to quick and easy access to content, and are expecting an online experience that is rich and instant. This is a generation that is always connected, through multiple devices.

NLB saw great opportunities to ride on the confluence of technological advances and the connected lifestyle to bring relevant information to the users. Its digital strategy involves:

- digitising uniquely Singapore content
- making NLB content findable through the popular search engines
- delivering content on all devices through responsive web design

'Content is king' in the Internet era. NLB is in the privileged position to acquire and digitise uniquely Singapore content, through its statutory standing and through the extensive network of partners of content owners. Significant resources have been invested in the digitisation of valuable content, and the efforts continue unabated. More details on the NLB digital collections can be found in the next section of this paper.

Most users go to the popular search engines (e.g., Google) as their first port of call to search for information. Leveraging on this behaviour, NLB embarked on sustained search engine optimisation (SEO) efforts to ensure that its digitised resources are easily findable via the popular search engines. SEO and the user friendly content sites resulted in a wider reach and higher usage of these valuable digital resources. An oft-cited example is the increased usage of the Infopedia[1] service from 400 to 150,000 pageviews a month after the trial (Chellapandi, Chow & Tay, 2010). The average monthly pageviews in 2013 is 270,000.

[1] http://eresources.nlb.gov.sg/infopedia/

NLB has also adopted the Responsive Web Design framework since 2011, and implemented device-aware online viewing and streaming capabilities to ensure that the user experience on any device will be optimal.

3 THE NLB DIGITAL COLLECTIONS

Over the years, NLB has built up a huge collection of content to meet the diverse needs of its patrons. When the National Archives of Singapore (NAS) joined the NLB family in November 2012, it brought with it a huge and highly valuable collection of primary and unique materials on Singapore's history. Aggressive digitisation continues unabated to preserve as well as give wider access through more platforms. The digitised contents from NAS and NLB constitute a valuable repository of the memories of the nation.

Figure 1 shows the collections available for online and onsite access.

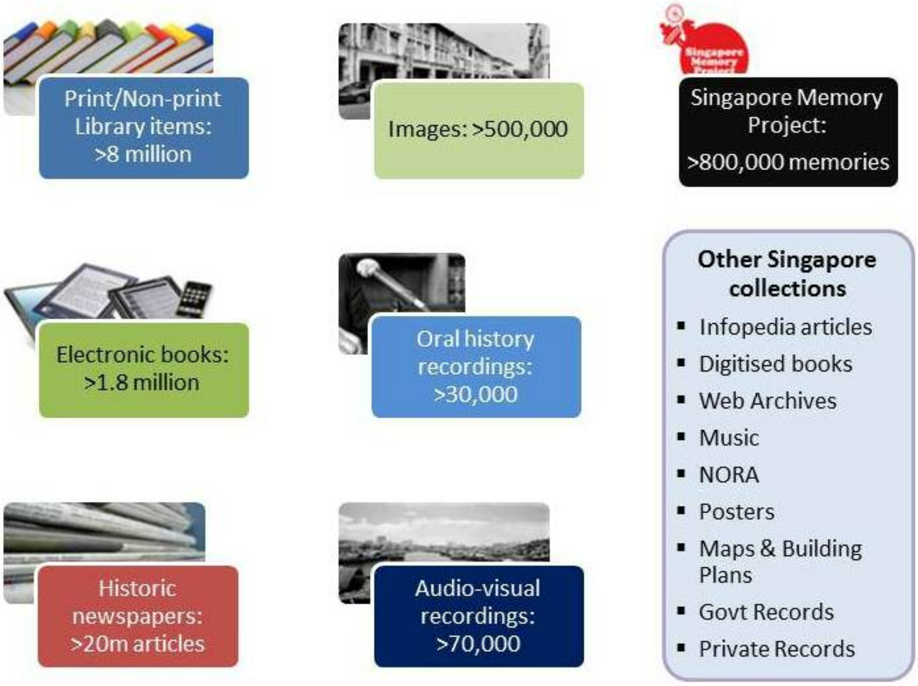
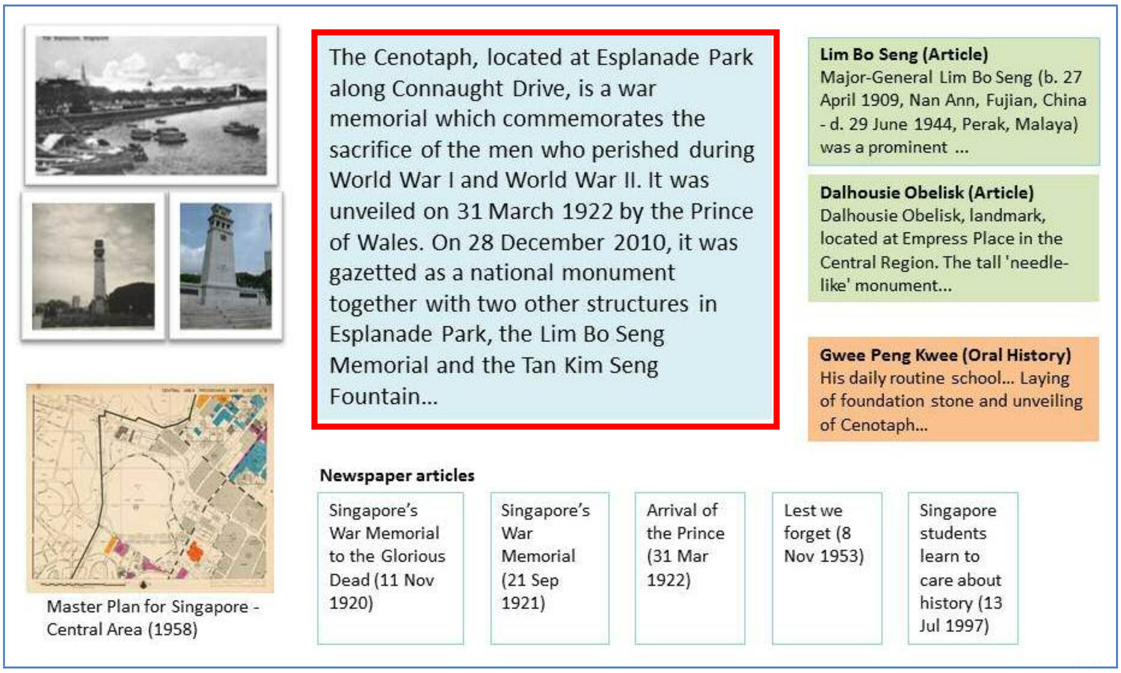

Print/Non-print Library items: >8 million
Images: >500,000
Singapore Memory Project
Singapore Memory Project: >800,000 memories
Electronic books: >1.8 million
Oral history recordings: >30,000
Historic newspapers: >20m articles
Audio-visual recordings: >70,000
Other Singapore collections
Infopedia articles
Digitised books
Web Archives
Music
NORA
Posters
Maps & Building Plans
Govt Records
Private Records

Figure 1: The NLB collections

4 CONTEXTUAL DISCOVERY

Riding on the success of the digital strategy, NLB looked for the next 'big thing'. Every year, NLB users collectively contribute to tens of millions of e-retrievals. We see every single one of these interactions as a golden opportunity to 'push' relevant content to the user. The success of the 'customers who bought this item also bought' recommendation feature at amazon.com is a clear testament of the power of pushing relevant recommendations.

However, to effectively connect people to knowledge, we need to connect knowledge to knowledge first. Figure 2 shows a subset of the related content within NLB's collections on the Cenotaph monument as described in the Cenotaph article highlighted in a red box.



The Cenotaph, located at Esplanade Park along Connaught Drive, is a war memorial which commemorates the sacrifice of the men who perished during World War I and World War II. It was unveiled on 31 March 1922 by the Prince of Wales. On 28 December 2010, it was gazetted as a national monument together with two other structures in Esplanade Park, the Lim Bo Seng Memorial and the Tan Kim Seng Fountain...

Lim Bo Seng (Article)
Major-General Lim Bo Seng (b. 27 April 1909, Nan Ann, Fujian, China - d. 29 June 1944, Perak, Malaya) was a prominent ...

Dalhousie Obelisk (Article)
Dalhousie Obelisk, landmark, located at Empress Place in the Central Region. The tall 'needle-like' monument...

Gwee Peng Kwee (Oral History)
His daily routine school... Laying of foundation stone and unveiling of Cenotaph...

Master Plan for Singapore - Central Area (1958)

Newspaper articles

- Singapore's War Memorial to the Glorious Dead (11 Nov 1920)
- Singapore's War Memorial (21 Sep 1921)
- Arrival of the Prince (31 Mar 1922)
- Lest we forget (8 Nov 1953)
- Singapore students learn to care about history (13 Jul 1997)

Figure 2: Contextual discovery through the recommendations of related content

To identify all these (and more) related resources would have required many searches to be performed with various search keywords, and sieving through hundreds if not thousands in the search result lists. This is very time-consuming. Given the current 'good enough' information seeking behaviour, it is unlikely that the typical users will be willing to spend the efforts needed to get to such a comprehensive set. It would be such a pity as they would be missing many treasures in the trove of digital resources available.

If we can 'push' the kind of relevant and related resources as illustrated in Figure 2 for each and every piece of digital resources in our collection, the users will have easy access to a comprehensive set of resources that provides a more complete 360 degree perspective of a topic.

Moreover, the Cenotaph article that the user is viewing (as a result of a search or browse action that the user just performed) provides a clear context of the user intent, which increases significantly the likelihood that the user will click through the recommendations, and discover many more resources in our collections. We call this 'contextual discovery', and believe strongly that it will be critical to the next generation of digital libraries.

This expands the NLB digital strategy, described in Section 2 above, by connecting the resources within and across collections, formats and languages to provide a rich contextual discovery experience.

5 USING TEXT ANALYTICS TO IDENTIFY RELATED CONTENT

The foundation to contextual discovery is the associations between the digital resources. Each association established a link between the resources, and together, a network of the resources is formed.

The value of the resources increases with the number of links, according to the 'network effect' phenomenon. With a total collection size that goes into tens of millions, it will not be cost-effective to identify the associations manually.

Text analytics technologies have been identified to be suitable and scalable to perform this task (Lim & Chinnasamy, 2013).

5.1 The Proof-of-Concept

To test the feasibility of the use of text analytics to accurately identify related content, a proof-of-concept (PoC) was conducted.

The open source machine learning software Mahout[2] from the Apache Software Foundation[3] was selected for the PoC. Newspaper articles for the year of 1989 from The Straits Times totalling over 50,000 were used. Figure 3 shows the steps involved in the text analytics processing.

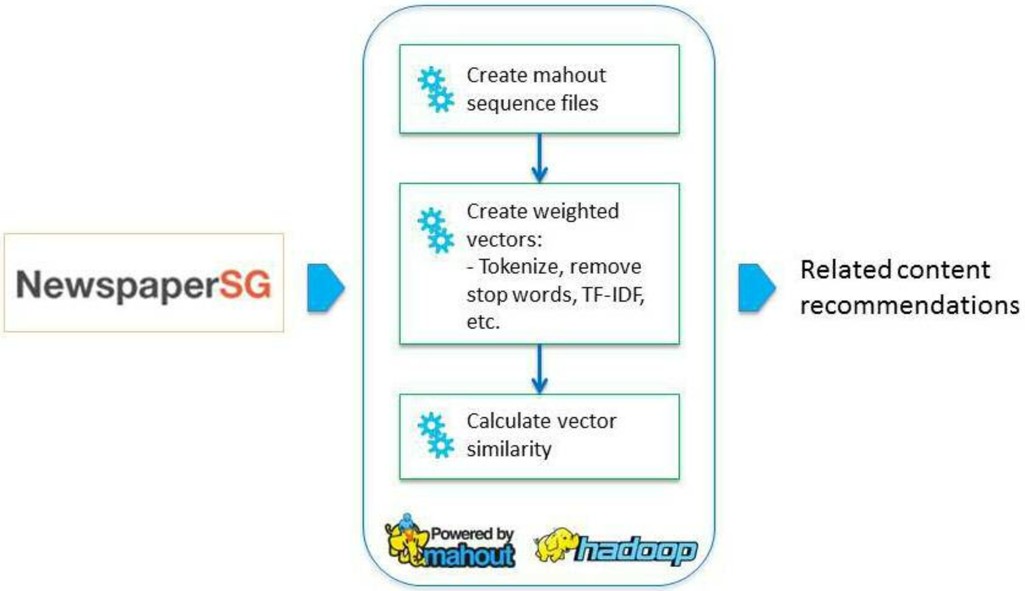
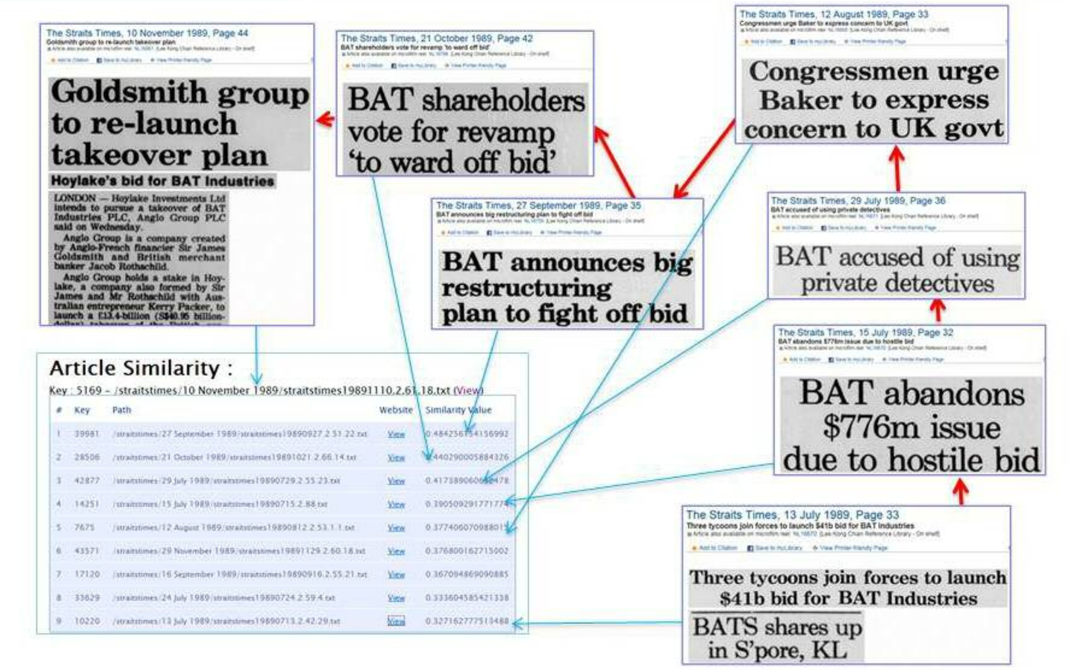

NewspaperSG
Create mahout sequence files
Create weighted vectors:
- Tokenize, remove stop words, TF-IDF, etc.
Calculate vector similarity
Powered by mahout
hadoop
Related content recommendations

Figure 3: Text analytics PoC using Mahout

The text analytics processing generated a set of the most 'similar' newspaper articles for each of the 50,000 newspaper articles used in the PoC. The similarity between 2 articles is measured by a value from 0 to 1, with 1 being a perfect match.

Take the article 'Goldsmith group to re-launch takeover plan' in Figure 4. The top recommendations from the text analytics software are clearly about the takeover event at the British American Tobacco (BAT). When the articles are organised in a chronological order (following the red arrows), the event unfolds.

[2] http://mahout.apache.org
[3] http://www.apache.org



The Straits Times, 10 November 1989, Page 44

Goldsmith group to re-launch takeover plan

Goldsmith group to re-launch takeover plan

Hoylake's bid for BAT Industries

LONDON — Hoylake Investments Ltd intends to pursue a takeover of BAT Industries PLC, Anglo Group PLC said on Wednesday.

Anglo Group is a company created by Anglo-French financier Sir James Goldsmith and British merchant banker Jacob Rothschild.

Anglo Group holds a stake in Hoylake, a company also formed by Sir James and Mr Rothschild with Australian entrepreneur Kerry Packer, to launch a £13.4-billion (S$40.95 billion-

The Straits Times, 21 October 1989, Page 42

BAT shareholders vote for revamp 'to ward off bid'

BAT shareholders vote for revamp 'to ward off bid'

The Straits Times, 27 September 1989, Page 35

BAT announces big restructuring plan to fight off bid

BAT announces big restructuring plan to fight off bid

The Straits Times, 12 August 1989, Page 33

Congressmen urge Baker to express concern to UK govt

Congressmen urge Baker to express concern to UK govt

The Straits Times, 29 July 1989, Page 36

BAT accused of using private detectives

BAT accused of using private detectives

The Straits Times, 15 July 1989, Page 32

BAT abandons $776m issue due to hostile bid

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

Three tycoons join forces to launch $41b bid for BAT Industries

BATS shares up in S'pore, KL

Article Similarity :

Key : 5169 – /straitstimes/10 November 1989/straitstimes19891110.2.61.18.txt (View)

#	Key	Path	Website	Similarity Value
1	39981	/straitstimes/27 September 1989/straitstimes19890927.2.51.22.txt	View	0.484256754156992
2	28506	/straitstimes/21 October 1989/straitstimes19891021.2.66.14.txt	View	0.440290005884326
3	42877	/straitstimes/29 July 1989/straitstimes19890729.2.55.23.txt	View	0.417389060062478
4	14251	/straitstimes/15 July 1989/straitstimes19890715.2.88.txt	View	0.390509291771774
5	7675	/straitstimes/12 August 1989/straitstimes19890812.2.53.1.1.txt	View	0.377406070988015
6	43571	/straitstimes/29 November 1989/straitstimes19891129.2.60.18.txt	View	0.376800162715002
7	17120	/straitstimes/16 September 1989/straitstimes19890916.2.55.21.txt	View	0.367094869090885
8	33629	/straitstimes/24 July 1989/straitstimes19890724.2.59.4.txt	View	0.333604585421338
9	10220	/straitstimes/13 July 1989/straitstimes19890713.2.42.29.txt	View	0.327162777513488

Figure 4: Similar newspaper articles identified via text analytics

Algorithms implemented in Mahout, such as Term Frequency/Inverse Document Frequency (TF/IDF), a statistics that reflects the importance of a word in a corpus, and Cosine Similarity for text-based vector similarity are well-established in the domain of information retrieval.

The PoC confirmed that the use of text analytics algorithms can automatically, efficiently and accurately identify similar articles.

5.2 The Deployment

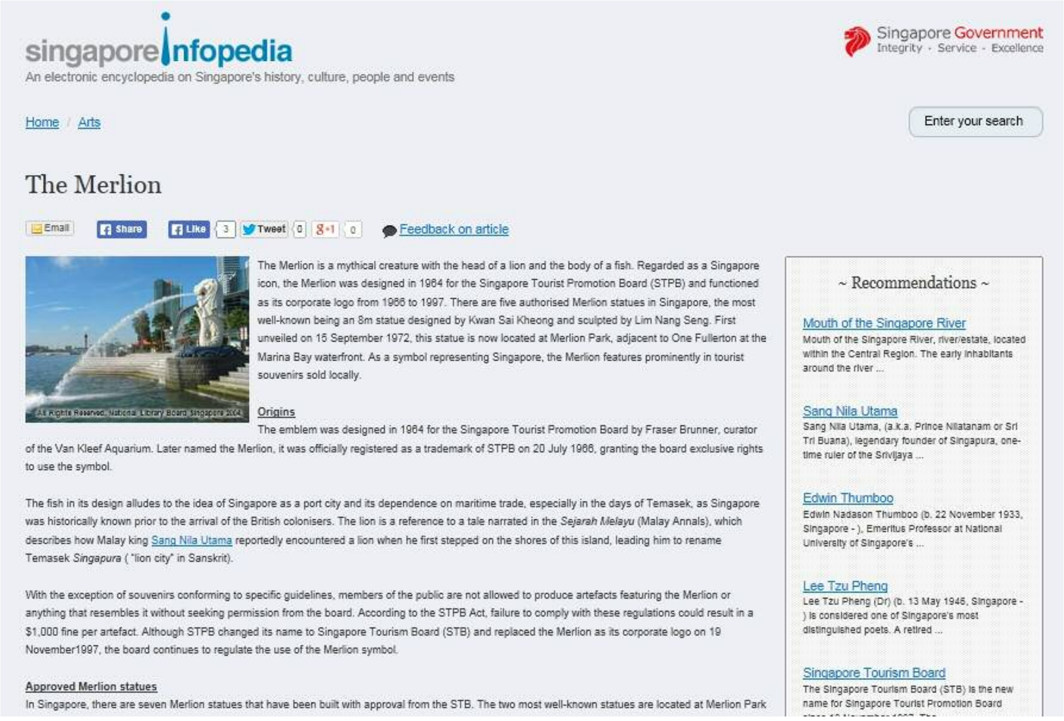
With the success of the PoC, NLB proceeded to implement the concept for its collections. We decided to start with the Infopedia collection given the smaller size of the collection (of around 1,800 articles) and the well written and fairly long text available. The text analytics processing was completed within an hour.

The recommendations were checked rigorously by the librarians on the staging Infopedia site. We went 'live' in June 2013 after the go-ahead was given (see Figure 5).

Thereafter, we extended the feature to other NLB digital collections, including PictureSG[4] and Singapore Memory Portal[5].

[4] http://eresources.nlb.gov.sg/pictures/

[5] http://www.singaporememory.sg/



singapore infopedia
An electronic encyclopedia on Singapore's history, culture, people and events
Singapore Government
Integrity · Service · Excellence
Home / Arts
Enter your search
Email
Share
Like
3
Tweet
0
g+1
0
Feedback on article
All Rights Reserved, National Library Board Singapore 2004

The Merlion

The Merlion is a mythical creature with the head of a lion and the body of a fish. Regarded as a Singapore icon, the Merlion was designed in 1964 for the Singapore Tourist Promotion Board (STPB) and functioned as its corporate logo from 1966 to 1997. There are five authorised Merlion statues in Singapore, the most well-known being an 8m statue designed by Kwan Sai Kheong and sculpted by Lim Nang Seng. First unveiled on 15 September 1972, this statue is now located at Merlion Park, adjacent to One Fullerton at the Marina Bay waterfront. As a symbol representing Singapore, the Merlion features prominently in tourist souvenirs sold locally.

Origins

The emblem was designed in 1964 for the Singapore Tourist Promotion Board by Fraser Brunner, curator of the Van Kleef Aquarium. Later named the Merlion, it was officially registered as a trademark of STPB on 20 July 1966, granting the board exclusive rights to use the symbol.

The fish in its design alludes to the idea of Singapore as a port city and its dependence on maritime trade, especially in the days of Temasek, as Singapore was historically known prior to the arrival of the British colonisers. The lion is a reference to a tale narrated in the *Sejarah Melayu* (Malay Annals), which describes how Malay king Sang Nila Utama reportedly encountered a lion when he first stepped on the shores of this island, leading him to rename Temasek *Singapura* ("lion city" in Sanskrit).

With the exception of souvenirs conforming to specific guidelines, members of the public are not allowed to produce artefacts featuring the Merlion or anything that resembles it without seeking permission from the board. According to the STPB Act, failure to comply with these regulations could result in a $1,000 fine per artefact. Although STPB changed its name to Singapore Tourism Board (STB) and replaced the Merlion as its corporate logo on 19 November1997, the board continues to regulate the use of the Merlion symbol.

Approved Merlion statues

In Singapore, there are seven Merlion statues that have been built with approval from the STB. The two most well-known statues are located at Merlion Park

~ Recommendations ~

Mouth of the Singapore River
Mouth of the Singapore River, river/estate, located within the Central Region. The early inhabitants around the river ...

Sang Nila Utama
Sang Nila Utama, (a.k.a. Prince Nilatanam or Sri Tri Buana), legendary founder of Singapura, one-time ruler of the Srivijaya ...

Edwin Thumboo
Edwin Nadason Thumboo (b. 22 November 1933, Singapore -), Emeritus Professor at National University of Singapore's ...

Lee Tzu Pheng
Lee Tzu Pheng (Dr) (b. 13 May 1946, Singapore -) is considered one of Singapore's most distinguished poets. A retired ...

Singapore Tourism Board
The Singapore Tourism Board (STB) is the new name for Singapore Tourist Promotion Board

Figure 5: Infopedia recommendations via text analytics

At the onset of selecting the text analytics software for the PoC, one of the criteria is the ability to scale to handle large data sets that go into millions. Most of the algorithms supported by Mahout have been developed to operate in an Apache Hadoop[6] cluster. Hadoop is a popular open source software framework for distributed storage and large scale processing of data sets on a cluster of commodity hardware.

As we progress to process collections that are much larger, we started to implement a Hadoop cluster. The current Hadoop cluster comprises 13 virtual servers implemented in the NLB private cloud. This allows NLB to scale and re-configure the cluster quickly depending on the resource requirement of the analytics to be performed. This is critical as analytics algorithms are generally very resource intensive.

5.3 Handling large data sets

The 13-nodes Hadoop cluster proved to be unable to process collections that go beyond hundreds of thousands items. While technically feasible, it is not practical to continue to add nodes and storage to the Hadoop cluster.

It is in fact not necessary to compare every pair of resources as the bulk of the records in a huge data set will bear little similarity with one another. A better approach would be to break the large data set into smaller clusters of related content that can then be efficiently processed.

The good news is that Mahout also comes with the necessary clustering algorithms. With this, we went ahead to process the key English newspapers (6.7 million articles) and Chinese

[6] http://hadoop.apache.org

newspapers (2.3 million articles) in NewspaperSG[7] and NAS Archives Online[8] collections (1 million records).

Table 1 shows the top 50 terms within example clusters for the English newspaper articles. The clustering certainly worked well. Note that stemming was done during the clustering process.

Size of cluster	Top 50 terms
52,678	exhibit, art, artist, paint, museum, singapor, work, displai, galleri, open, mr, year, organis, centr, chines, cultur, held, on, nation, world, pictur, photograph, collect, hall, colour, includ, first, time, featur, sculptur, design, dai, peopl, piec, two, societi, part, visitor, fair, intern, painter, life, school, prize, old, show, road, book, trade, today
86,881	olymp, athlet, game, sport, medal, event, team, gold, championship, record, world, singapor, metr, swim, won, year, champion, win, nation, women, time, coach, asian, meet, competit, train, swimmer, two, second, race, compet, first, amateur, bronz, intern, associ, best, finish, yesterdai, silver, old, on, relai, men, set, track, medallist, mark, south, run
142,289	school, student, educ, teacher, univers, secondari, singapor, children, year, primari, studi, pupil, parent, mr, teach, colleg, cours, ministri, english, languag, chines, on, institut, examin, time, learn, princip, train, programm, work, graduat, help, nation, class, two, govern, girl, scienc, boi, first, level, malai, centr, make, dr, academ, dai, organis, scholarship, junior
125,629	polic, arrest, offic, suspect, two, men, yesterdai, man, investig, report, found, mr, gang, road, raid, on, station, crime, detain, year, night, arm, robberi, car, believ, peopl, forc, charg, singapor, robber, hous, todai, told, stolen, seiz, spokesman, old, held, four, escap, murder, reuter, chines, detect, member, street, made, drug, dai, home

Table 1: Top 50 stemmed terms of example clusters from English newspapers

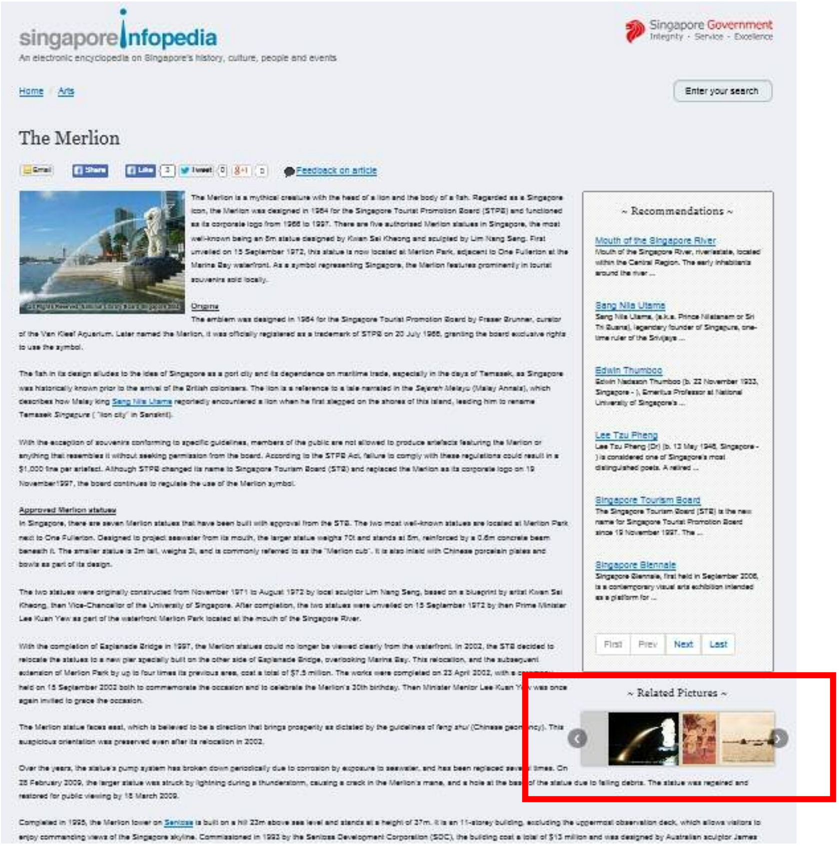
6 BENEFITS

With the use of text analytics, many high quality recommendations can be identified. Before this was implemented, associations were manually identified for some of the collections. However, due to the manual efforts required, the number of recommendations is necessarily limited. Now, the users have access to over a billion quality associations, while the librarians need not spend time identifying them manually.

While NLB is progressively introducing and enhancing the contextual discovery capability to more of its collections, the evidence so far has indicated a significantly higher usage of the digital collections after the contextual discovery implementations.

A good case in point is the cross-collection recommendations to related pictures at the Infopedia articles (red box in Figure 6).

[7] http://eresources.nlb.gov.sg/newspapers/
[8] http://www.nas.gov.sg/archivesonline


singaporeinfopedia

An electronic encyclopedia on Singapore's history, culture, people and events

Singapore Government
Integrity · Service · Excellence

Home / Arts

Enter your search

The Merlion

Feedback on article

The Merlion is a mythical creature with the head of a lion and the body of a fish. Regarded as a Singapore icon, the Merlion was designed in 1964 for the Singapore Tourist Promotion Board (STPB) and functioned as its corporate logo from 1966 to 1997. There are five authorised Merlion statues in Singapore, the most well-known being an 8m statue designed by Kwan Sai Kheong and sculpted by Lim Nang Seng. First unveiled on 15 September 1972, this statue is now located at Merlion Park, adjacent to One Fullerton at the Marina Bay waterfront. As a symbol representing Singapore, the Merlion features prominently in tourist souvenirs sold locally.

Origins

The emblem was designed in 1964 for the Singapore Tourist Promotion Board by Fraser Brunner, curator of the Van Kleef Aquarium. Later named the Merlion, it was officially registered as a trademark of STPB on 20 July 1966, granting the board exclusive rights to use the symbol.

The fish in its design alludes to the idea of Singapore as a port city and its dependence on maritime trade, especially in the days of Temasek, as Singapore was historically known prior to the arrival of the British colonisers. The lion is a reference to a tale narrated in the *Sejarah Melayu* (Malay Annals), which describes how Malay king Sang Nila Utama reportedly encountered a lion when he first stepped on the shores of this island, leading him to rename Temasek *Singapura* ("lion city" in Sanskrit).

With the exception of souvenirs conforming to specific guidelines, members of the public are not allowed to produce artefacts featuring the Merlion or anything that resembles it without seeking permission from the board. According to the STPB Act, failure to comply with these regulations could result in a $1,000 fine per artefact. Although STPB changed its name to Singapore Tourism Board (STB) and replaced the Merlion as its corporate logo on 19 November1997, the board continues to regulate the use of the Merlion symbol.

Approved Merlion statues

In Singapore, there are seven Merlion statues that have been built with approval from the STB. The two most well-known statues are located at Merlion Park next to One Fullerton. Designed to project seawater from its mouth, the larger statue weighs 70t and stands at 8m, reinforced by a 0.6m concrete beam beneath it. The smaller statue is 2m tall, weighs 3t, and is commonly referred to as the "Merlion cub". It is also inlaid with Chinese porcelain plates and bowls as part of its design.

The two statues were originally constructed from November 1971 to August 1972 by local sculptor Lim Nang Seng, based on a blueprint by artist Kwan Sai Kheong, then Vice-Chancellor of the University of Singapore. After completion, the two statues were unveiled on 15 September 1972 by then Prime Minister Lee Kuan Yew as part of the waterfront Merlion Park located at the mouth of the Singapore River.

With the completion of Esplanade Bridge in 1997, the Merlion statues could no longer be viewed clearly from the waterfront. In 2002, the STB decided to relocate the statues to a new pier specially built on the other side of Esplanade Bridge, overlooking Marina Bay. This relocation, and the subsequent extension of Merlion Park by up to four times its previous area, cost a total of $7.5 million. The works were completed on 23 April 2002, with a c held on 15 September 2002 both to commemorate the occasion and to celebrate the Merlion's 30th birthday. Then Minister Mentor Lee Kuan Y w was once again invited to grace the occasion.

The Merlion statue faces east, which is believed to be a direction that brings prosperity as dictated by the guidelines of feng shui (Chinese geom ncy). This auspicious orientation was preserved even after its relocation in 2002.

Over the years, the statue's pump system has broken down periodically due to corrosion by exposure to seawater, and has been replaced seve l times. On 28 February 2009, the larger statue was struck by lightning during a thunderstorm, causing a crack in the Merlion's mane, and a hole at the bas of the statue due to falling debris. The statue was repaired and restored for public viewing by 18 March 2009.

Completed in 1995, the Merlion tower on Sentosa is built on a hill 22m above sea level and stands at a height of 37m. It is an 11-storey building, excluding the uppermost observation deck, which allows visitors to enjoy commanding views of the Singapore skyline. Commissioned in 1993 by the Sentosa Development Corporation (SDC), the building cost a total of $13 million and was designed by Australian sculptor James

~ Recommendations ~

Mouth of the Singapore River
Mouth of the Singapore River, riverstate, located within the Central Region. The early inhabitants around the river ...

Sang Nila Utama
Sang Nila Utama, (a.k.a. Prince Nilatanam or Sri Tri Buana), legendary founder of Singapura, one-time ruler of the Srivijaya ...

Edwin Thumboo
Edwin Nadason Thumboo (b. 22 November 1933, Singapore -), Emeritus Professor at National University of Singapore's ...

Lee Tzu Pheng
Lee Tzu Pheng (Dr) (b. 13 May 1946, Singapore -) is considered one of Singapore's most distinguished poets. A retired ...

Singapore Tourism Board
The Singapore Tourism Board (STB) is the new name for Singapore Tourist Promotion Board since 19 November 1997. The ...

Singapore Biennale
Singapore Biennale, first held in September 2006, is a contemporary visual arts exhibition intended as a platform for ...

~ Related Pictures ~

Figure 6: Cross-collection recommendations

The recommendations within the PictureSG collection was implemented in September 2013, while the cross-collection recommendations from Infopedia to PictureSG was launched in November 2013. Table 2 shows some of the web statistics for PictureSG before and after the implementations.

Month	Total referrals	# from Infopedia	%	Page views	Page views per visit
Apr-13 to Aug-13	-	-	-	37,841 (average)	3.64 (average)
Sep-13	58,376	84	0.14	65,036	4.00
Oct-13	21,088	93	0.44	41,969	3.34
Nov-13	30,684	1,421	4.63	62,018	5.27
Dec-13	27,628	1,554	5.62	55,605	5.07
Jan-14	28,149	2,424	8.61	70,313	5.73
Feb-14	29,871	3,181	10.65	84,341	6.41
Mar-14	31,832	3,560	11.18	96,626	6.53

Table 2: PictureSG web statistics

Within 4 months, over 10% of the traffic came through Infopedia!! The usage and the page views per visit for PictureSG have also increased significantly.

This is certainly very encouraging, and we are very confident that the implementation of the contextual discovery capability across all our collections will provide a significant boost to the already high usage of the collections.

It is also interesting to note that although the average lengths of the textual information available differs significantly between the Infopedia and PictureSG collections, it did not prevent the use of text analytics to identify the associations. The average amount of textual data used for the text analytics for the Infopedia collection was 1,054 words, while that for PictureSG was 47. Similarly, the presence of OCR errors in the NewspaperSG newspaper articles has not been an issue too. There is therefore a fair amount of resilience built into the text analytics algorithms to deal with them.

The text analytics approach provides a quantitative measure of the degree of similarity between two resources. This similarity value is based on established information retrieval models and best practices. It provides us the mean to identify the best recommendations, and also the cut-off point for our recommendations.

Libraries and archives play a key role in cultivating an informed society in a knowledge intensive economy like Singapore. Contextual discovery, with the ease it provides to users to explore and deepen their understanding and insights, enables NLB to connect people to knowledge and to the nation. It makes knowledge come alive, sparks imagination and creates possibilities.

7 LESSIONS LEARNT

The implementation of text analytics to enable the contextual discovery of millions of digital resources is not a walk in the park, more like a walk in the dark.

Start with a Proof-of-concept (PoC)

NLB decided to perform a PoC to determine the feasibility of the concept. This has allowed us to start small and quickly. The PoC results also helped us to communicate the possibilities to the stakeholders.

As importantly, the PoC has enabled us to develop a better understanding of capabilities of the technologies and tools, and build up internal competency in this domain. This is especially critical in the long run given the shortage of actual hands-on knowledge in this domain locally and in the region.

Established and affordable software in text analytics are available

We selected the Mahout as the software for the PoC as it was the most established open source machine learning and data mining software that met our requirements. That the algorithms were implemented on top of the Apache Hadoop framework provided us comfort in terms of scalability.

There were good documentation, including several well written books, on Mahout, which helped us tremendously in our PoC, and subsequent deployments.

Do not simply add more Hadoop nodes when hitting performance issues

With the Hadoop platform, it is tempting to simply add more compute and storage resources whenever we hit a performance issue. After we expanded the Hadoop cluster to 13 nodes, we decided that we should stop adding more nodes, but instead look for alternative solutions. As a result, we started exploring the clustering capabilities of Mahout, conducted more PoCs, and adopted the use of clustering to handle large data sets.

Learning curve

There is a substantial skill gap in text analytics, Mahout and Hadoop internally, in Singapore and in the region. Actual implementations of projects of this nature are few and far between. Adoption of Hadoop is still at a nascent stage in Asia Pacific.

The PoCs enabled NLB to explore, experiment and pick up critical skills needed. The documentation, books, online resources and forums, and online courses at Coursera[9] are good sources that help the team with the implementation.

8 NEXT STEPS

NLB is currently extending the contextual discovery capability with the use of text analytics to all the NLB digital collections. This is expected to be completed by 2014.

At the same time, we are exploring several possibilities to bring contextual discovery to the next level:

- *Cross-institution*. With NAS coming under NLB, we now have an expanded collection of digital resources on Singapore. We will be working on recommendations across NAS and the National Library.

- *Cross-language*. There are 4 official languages in Singapore: English, Chinese, Malay and Tamil. We are considering the use of machine translation to provide cross-language recommendations.

- *Content analytics*. Describing the digital content is a very labour intensive process. There is currently a large portion of the digital content that has not been described. As a result, they remain ‘unsearcheable’ and text analytics cannot be applied to make them discoverable.

  We are exploring the use of content analytics software to extract information from the ‘raw’ content. One example is the use of voice-to-text technologies for oral history and audio visual recordings. Another possibility is to perform feature extraction on images to identify similar images.

[9] https://www.coursera.org/

9 CONCLUSION

The contextual discovery capabilities described in this paper will be a critical component of the next generation of digital libraries. This is particularly important in an era of data deluge and shorter attention span.

Memory institutions such as libraries, archives and museums play a key role in connecting the people to their country. With contextual discovery, they can now bring together the full spectrum of resources within its repositories regardless of institutions, formats and language, and present them in a way that encourages a deeper understanding of the precious memories of the nation.

It is also what knowledge seekers have been dreaming of. They can spend less time gathering the pieces, and focus on digesting and analysing the 'dossier' of relevant information to derive new insights.

The core of this rich contextual discovery experience is a massively connected network of knowledge resources. Text analytics has been shown to be a good and scalable solution to automatically identify highly accurate connections. The results so far have been very promising.

The journey has begun. Over time, with more data, better technologies and associations, the 'knowledge network' will form the bedrock of next generation knowledge management. This promotes higher level of engagement and interactivity with the knowledge network and opens up more opportunities for innovative ideas to be conceived. It will power an informed society in the new knowledge economy.

References

Sharmini Chellapandi, Chow Wun Han, Tay Chiew Boon, (2010) "The National Library of Singapore experience: harnessing technology to deliver content and broaden access", Interlending & Document Supply, Vol. 38 Iss: 1, pp.40 – 48

LIM, Chee Kiam and CHINNASAMY, Balakumar (2013) Connecting library content using data mining and text analytics on structured and unstructured data. Paper presented at: IFLA WLIC 2013 — Singapore — "Future Libraries: Infinite Possibilities" in Session 152 - Reference and Information Services.