IFLA 2014 LYON

Submitted on: 7/28/2014

News Alike – Text analytics to link related NewspaperSG articles

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

Chee Kiam LIM
Technology & Innovation, National Library Board, Singapore
lim_chee_kiam@nlb.gov.sg

Chinnasamy BALAKUMAR
Technology & Innovation, National Library Board, Singapore
balakumar_chinnasamy@nlb.gov.sg

Cally LAW
Technology & Innovation, National Library Board, Singapore
cally_law@nlb.gov.sg

Peter PAK
Technology & Innovation, National Library Board, Singapore
peter_pak@nlb.gov.sg

Abstract:

The National Library Board (NLB) of Singapore explored the use of text analytics to associate its vast and growing collections of uniquely and valuable Singapore content. The first content service to implement and benefit from text analytics was Infopedia in June 2013. We have been progressively extending the feature to other NLB content-rich services.

The NewspaperSG collection however posed seemingly insurmountable challenge due to its sheer collection size of over 20 million (and growing) published newspaper articles and the presence of OCR errors as a result of the newspaper digitisation process.

Various approaches were tried, with mixed results. We eventually adopted a 2-step process. First, automatic clustering technique was used to identify smaller clusters of articles. Text analytics were then applied to each of the clusters.

Close to 10 million NewspaperSG articles in English, Chinese and Malay were processed and over 900 million similarity associations identified. A new 'News Alike' feature in NewspaperSG makes the related articles available when the users peruse any of these 10 million articles. The enhanced contextual discovery experience makes it easier for the users to discover relevant content.

Keywords: Historic newspapers, text analytics, clustering, contextual discovery, NewspaperSG

---

1 EXCELLENCE IN SINGAPORE CONTENT

Excellence in Singapore content has been one of the core pillars of the National Library Board (NLB) of Singapore. NLB has embarked on large scale acquisition and digitisation of unique and valuable content of literary, heritage, cultural and national importance to Singapore. Once digitised, the rich-media content are made available through innovative content services tailored to the nature of the materials.

The list of content services is as follows:

- Infopedia – http://eresources.nlb.gov.sg/infopedia/
  An electronic encyclopedia on Singapore, providing access to collection of articles on Singapore's history, culture, people, and events.

- NewspaperSG – http://eresources.nlb.gov.sg/newspapers/
  An online resource of current and historic Singapore and Malaya newspapers. You can search our digital archive of English, Chinese and Malay language newspapers published between 1831-2009, or find information on over 200 newspaper titles in the National Library's microfilm collection.

- BookSG – http://eresources.nlb.gov.sg/printheritage/
  An online collection of digitised books and other printed materials held in the National Library Singapore, including rare and historical imprints, selected works from the British Library's Oriental & India Office Collection and works published in Singapore. The collection features materials related to Singapore and Southeast Asia.

- PictureSG – http://eresources.nlb.gov.sg/pictures/
  An online collection of images (photographs, artworks) aims to educate the public on the socio-cultural & historical development of Singapore.

- MusicSG – http://music.nl.sg
  A non-profit digital archive set up to digitise, archive and provide access to all forms of published Singapore musical works. It assembles a collection of music composed or published by Singaporeans, music produced or published in Singapore, and music related to Singapore.

- Singapore Memory Portal – http://singaporememory.sg

The Singapore Memory Project (SMP) is a whole-of-nation movement that aims to capture and document precious moments and memories related to Singapore; recollections not merely from individual Singaporeans, but also organisations, associations, companies and groups.

The web portal – SingaporeMemory.sg - allows every Singaporean to own a memory account to deposit their memories and stories. Memories can be deposited in the form of texts, audio files, video files or images. The web portal will help in community bonding by connecting contributors with shared memories and similar experiences. This will help to draw Singaporeans from all walks of life closer and foster greater social cohesion.

2 USING TEXT ANALYTICS TO CONNECT CONTENT

To enhance the discovery of the vast collection of digital content, NLB started to explore the use of text analytics tools. It aims to identify related content in order to 'push' them as recommendations to the information seeker for a more comprehensive discovery experience (Lim & Chinnasamy, 2013).

The Mahout software from the Apache Software Foundation is an established open source software for scalable data and text analytics [1]. The core algorithms in Mahout are implemented on top of Apache Hadoop [2] using the map/reduce paradigm, a popular framework for massively parallel and distributed processing. One of the Mahout capabilities is the computation of the degree of 'similarity' between pairs of textual content within an input document set.

The documents are each represented as a vector of tokens. For example, the phrase 'The effective text analysis of NewspaperSG articles' can be represented as a vector comprising the tokens 'the', 'effective', 'text', 'analysis', 'of', 'newspapersg' and 'articles'. All the unique tokens from the entire data set form the vector space, and the number of unique tokens in the vector space represents the number of dimensions of the space.

As you can imagine, some tokens are not exactly important. Tokens such as 'the' and 'of' are generally not important in conveying the key message of the article, so they should not carry the same weight as other tokens like 'newspapersg'. Such stop words should be removed as early as possible to improve the accuracy of the recommendation and reduce the amount of time needed for processing.

Even those tokens that remained do not carry the same level of importance in the context of the data set. The weight associated with each token is determined by the established Term Frequency/Reversed Document Frequency (TF/IDF) algorithm, a well established algorithm for search engines and text mining (Salton & McGill, 1986).

The tokens can be single-words (known as 1-gram or mono-gram), or phrases made up of up to n words (n-grams). For example, working with 2-grams will allow Mahout to work with 'text analysis' as a token, and the additional useful information of the proximity of the 'text' and 'analysis' tokens in making its recommendations.

[1] http://mahout.apache.org/
[2] http://hadoop.apache.org/

Each article gets tokenised, and processed to remove case-sensitivity, stop words and numerals, and have the remaining tokens weighted using TF/IDF.

Mahout will then compute the similarity values for all pairs of the weighted vectors based on one of the distance measurements available (Euclidean, Squared Euclidean, Manhattan, Cosine, Tanimoto, etc.). The similarity values range from 0 to 1 (with 1 being most similar), allowing us to pick the good ones to make the recommendations.

The first implementation was rolled out for the popular Infopedia service in June 2013. It was applied to other content-rich services subsequently.

3 TEXT ANALYTICS – A KEY COMPONENT OF THE NLB SERVICE ENABLEMENT ARCHITECTURE

With the successful deployment of text analytics, a robust and scalable text analytics infrastructure has been established within the comprehensive NLB service architecture that powers the various innovative services provided by NLB. The Service Enablement Architecture (Figure 1) has been progressively developed, and the text analytics component is a core part of the Discovery Services.

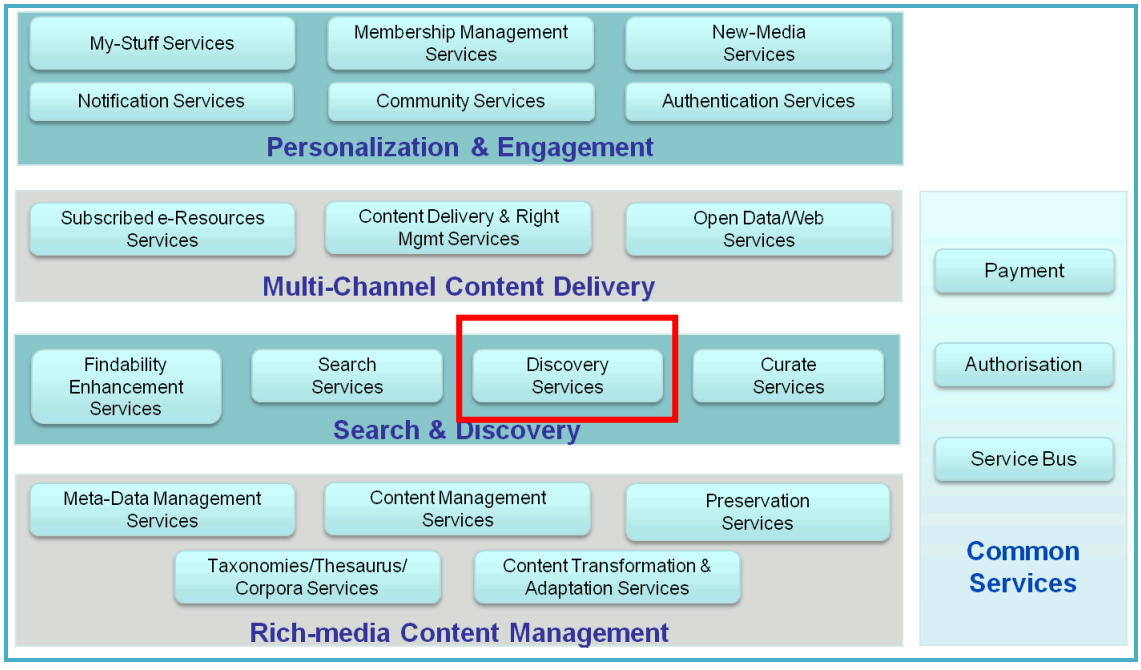

Figure 1: Text analytics is a core part of the Discovery Services

The text analytics capability currently includes the Mahout software operating over a Hadoop cluster with the configuration shown in Figure 2.

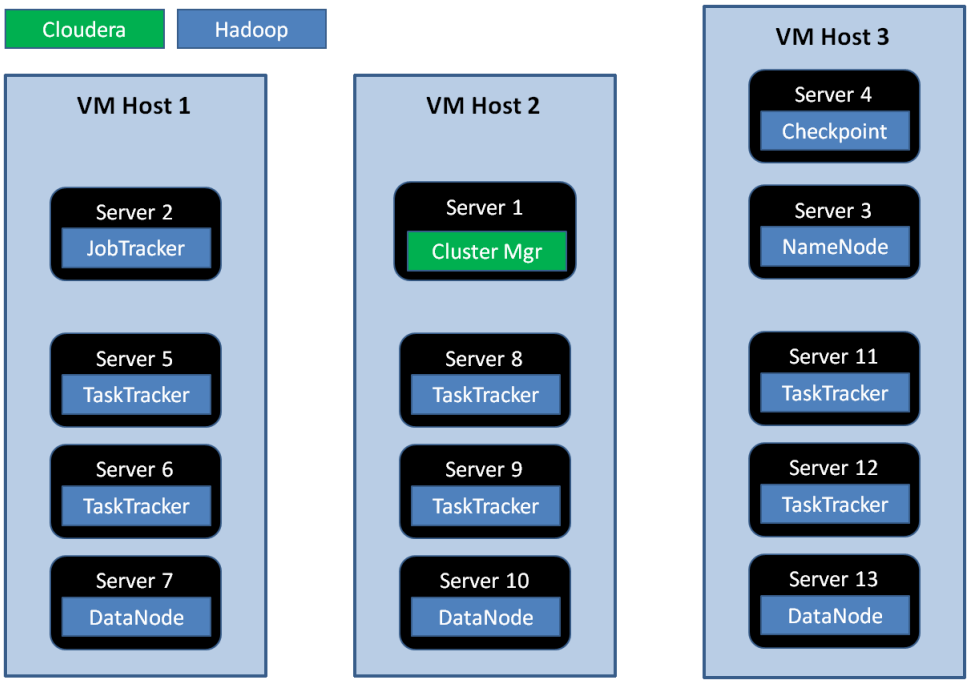

Figure 2: The NLB Hadoop configuration

4 APPLYING TEXT ANALYTICS ON NEWSPAPERSG – THE CHALLENGES

NewspaperSG is an online resource of current and historic Singapore and Malaya newspapers published between 1831 – 2009. Currently, over 20 million newspaper articles have been made available via NewspaperSG.

The searchable newspaper text that appears in the article extract has been automatically generated using Optical Character Recognition (OCR) software. OCR is a process by which software reads a page image and translates it into a text file by recognising the shapes of the letters. OCR enables searching of large quantities of full-text data, but it is not 100% accurate. The level of accuracy depends on the print quality of the original newspaper, its condition at the time of microfilming, and the level of detail captured by the microfilm scanner. Newspapers with poor quality paper, small print, mixed fonts, multiple column layouts, or damaged pages may have poor OCR accuracy.

The two key factors affecting the extent of processing required during the Mahout similarity processing are:

- Size of the input data set

  The NewspaperSG data set comprises over 20 million published articles. Mahout computes the similarity values for every pair of the articles within the data set to pick the top pairs with the highest similarity values. The processing time will increase significantly as the size of the data set increases.

- Number of dimensions

  The number of dimensions is the number of unique tokens identified by mahout after analysing all the articles in the data set. It is highly dependent on the n-gram parameter passed in to the mahout processing. By default, the n-gram is 1, and all tokens will be single words. So 'Coca Cola' will result in 2 1-grams of 'Coca' and 'Cola'. If the n-gram parameter is set to 2, then we will have 3 unique tokens: 'Coca', 'Cola' and 'Coca Cola'. While a higher n-gram may improve the accuracy of the recommendations, it will significantly increase the number of dimensions.

  The historic newspaper collection poses an additional challenge. The OCR errors in the newspaper articles significantly increased the number of dimensions.

  Taking just the articles from 2 English newspapers (The Straits Time and the Singapore Free Press) of over 6 million articles, the number of dimensions turned out to be around 5,780,000!! One of the largest English word list available is WordNet and the size of that list is only 155,287[3]. The bloated number of dimensions is largely the results of the OCR errors.

Table 1 illustrates the impact of the above factors to the processing time needed for the various NLB content collections.

Document Set	No of documents	No of dimensions	Processing time
Infopedia	1,688	16,524	6 minutes
PictureSG	19,662	21,090	8 minutes
Singapore Memory Portal	62,980	9,069	60 minutes

Table 1: Impact of document set size and dimension on processing time

It is therefore not feasible to process the NewspaperSG collection by the brute-force approach of pair-wise comparison of all article pairs. It will take forever, unless we invest in hundreds of servers and petabytes of storage.

It is in fact not necessary since the majority of the pairs will have few if any common tokens of significant weightage and the similarity values will be close to zero – a sparse matrix.

5 THE INITIAL ATTEMPTS

With the seemingly insurmountable challenges presented by the unique nature of the NewspaperSG collection, we needed alternative approaches.

We needed to reduce the number of dimensions from 5,780,000 to a much more manageable number. Mahout provides a few parameters that can be used to drop tokens from its vector space[4]:

[3] http://wordnet.princeton.edu/
[4] Taken from Mahout in Action by Sean Owen, Robin Anil, Ted Dunning and Ellen Friedman

Parameter	Description	Default value
Minimum support	The minimum frequency of the term in the entire collection to be considered as a part of the dictionary file. Terms with lesser frequency are ignored.	2
Minimum document frequency	The minimum number of documents the term should occur in to be considered a part of the dictionary file. Any term with lesser frequency is ignored.	1
Max document frequency percentage	The maximum number of documents the term should occur in to be considered a part of the dictionary file. This is a mechanism to prune out high frequency terms (stop-words). Any word that occurs in more than the specified percentage of documents is ignored.	99

We have progressively increased the *Minimum support* and *Minimum document frequency* taking the assumption that OCR errors tend to have a lower frequency as compared to proper words. The *Max document frequency percentage* has also been reduced to be more aggressive in the removal of stop words. The results are as follows:

Total number of articles: 6,167,934 **n-gram = 1**

Minimum support	Minimum document frequency	Max document frequency percentage	No of dimensions
2	1	90	5,780,000
7	7	90	1,521,828
7	7	25	1,521,828
10	10	25	1,119,369
30	10	25	473,883

While it is possible to adjust the above parameters to reduce the number of dimensions (i.e., the number of unique tokens remaining after the trimming), it is nonetheless a coarse approach. With aggressive trimming, proper words are inevitably removed while OCR errors, though reduced, will still form a significant proportion of the token space.

With the most aggressive set of parameters (the last line in the above table), the Mahout processing took 4 days for 4 years of newspaper articles using the servers and storage available on our Hadoop cluster. We could then break all the newspapers into chunks of 4 years, and process them one after another, and ultimately complete the processing of the entire NewspaperSG collection. However, the recommendations will be limited to articles within the 4 year bands.

We have also experimented with the use of the Singular Value Decomposition (SVD)[5] to reduce the number of dimensions, but the processing time did not improve.

[5] http://en.wikipedia.org/wiki/Singular_value_decomposition

6 LINKING NEWSPAPERSG ARTICLES THROUGH INFOPEDIA

We then hit upon an interesting idea.

NLB's focus has always been on Singapore content. The priority for our text analytics initiative on the NewspaperSG collection would naturally be those newspaper articles related to Singapore.

Over the years, NLB has built up a comprehensive collection of Infopedia articles covering Singapore's history, culture, people, places and events. All in, the articles cover around 1,700 topics that are close to the heart of Singaporeans. The Infopedia articles, with an average of around 1,000 words per article, are well written, well researched, clear and concise.

To extract the gist of each Infopedia article, we used key term extraction technique to pull out the most important phrases within the article. This was done through the AlchemyAPI software[6]. The key terms were then filtered for common words.

For illustration, we would use the Infopedia article on the 'Housing and Development Board (HDB)'[7] as shown in Figure 3. The list of key terms for this article was as follows:

> *HDB flats, public housing, new flats, home ownership, home ownership rate, housing shortage, HDB Hub, Background HDB, housing programmes, housing authority, housing problem, housing units, HDB home ownership, housing estate, estate renewal strategy*


singapore Infopedia

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

Home › Politics and Government

Housing and Development Board (HDB)

Email Share Like 3 Tweet 0 +1 0 Feedback on article

The Housing and Development Board (HDB) is the national public housing authority of Singapore and it is currently headquartered at HDB Hub in Toa Payoh. It was formed in February 1960, shortly after Singapore attained self-government, to alleviate the severe housing shortage facing the state at the time. Through the years, the emphasis of its housing programmes has shifted from quantity of housing to quality of life. Since 1985, its flats have been housing over 80% of Singapore's resident population.

Background

HDB was formed as a direct result of the failure of its predecessor, the Singapore Improvement Trust (SIT), to provide adequate housing for the people. SIT was established in 1927 by the colonial government at a time when Singapore was facing an acute housing shortage. By the time SIT was replaced by HDB in 1960, the housing problem had worsened significantly as its building programmes had fallen far short of what was required to keep pace with the fast-growing population.

The first step towards the formation of HDB was taken in August 1958 with the introduction of the Housing and Development Bill in the legislative assembly. The bill was passed in January 1959 and came into effect on 1 February 1960, thus bringing HDB into existence to succeed SIT as Singapore's housing authority. Its primary function was to build and manage housing units for the low -income groups, and it had to produce the maximum output in the minimum time and at the lowest

Figure 3: Infopedia article on the Housing and Development Board (HDB)

[6] http://www.alchemyapi.com/
[7] http://infopedia.nl.sg/articles/SIP_1589_2009-10-26.html

The list of key terms was next used to search the NewspaperSG repository, and up to 20,000 top matches were identified for each Infopedia article. Every one of these 20,000 newspaper articles would contain at least one of the key terms. Moreover, since the search was based on the key terms from the Infopedia articles, the issue of OCR errors has been generally eliminated.

Continuing with the illustration, the following NewspaperSG search query was formed:

> *"HDB flats" OR "public housing" OR "new flat" OR "home ownership" OR "home ownership rate" OR "housing shortage" OR "HDB Hub" OR "Background HDB" OR "housing programmes" OR "housing authority" OR "housing problem" OR "housing units" OR "HDB home ownership" OR "housing estate" OR "estate renewal strategy"*

When the query was executed on 25 Jun 2014, 82,068 results were found, and the breakdown was as follows:

By Newspaper Titles	By Article Type*	By Word Count	By Year
The Straits Times: 67,934	Advertisements: 51,303	< 51 words: 463	2000 - 2009: 31,680
Today: 5,477	Articles: 27,627	51 - 300 words: 13,524	1990 - 1999: 22,568
The Business Times: 4,168	Illustrations: 12,372	301 - 800 words: 35,876	1980 - 1989: 16,914
Berita Harian: 1,756	Letters: 2,689	801 - 1,000 words: 13,130	1970 - 1979: 5,918
Singapore Monitor: 1,569	Miscellaneous: 433	> 1000 words: 19,075	1960 - 1969: 1,940
The Singapore Free Press: 766	Obituaries: 16		1950 - 1959: 1,456
Others: 398			Others: 1,592

** Illustrations are associated with Articles*

From the breakdown by newspaper titles, we could see that the results were drawn from various newspapers. We would be able to provide the users a comprehensive view of the perspectives and coverage by the various newspaper publishers on the same news item. As interestingly, we were able to analyse the news items across years, and present the results in a timeline manner showing how the story unfolded over time.

Each cluster of 20,000 newspaper articles around an Infopedia topic was then put through the established Mahout similarity process, and the related articles identified. With this 'divide-and-conquer' approach, the data preparation and Mahout processing tasks proceeded fairly smoothly at a rate of around 60 clusters per day, with the 1,682 clusters processed successfully in 28 days. The bulk of the processing time required was in data preparation. We repeated the Mahout processing using the same prepared 1,682 clusters covering 1,521,059 newspaper articles. The processing was completed within 9 days in the re-run.

Figure 4 shows the newspaper articles automatically identified by text analytics for the Infopedia article on 'Major oil spills in the Straits of Singapore'.

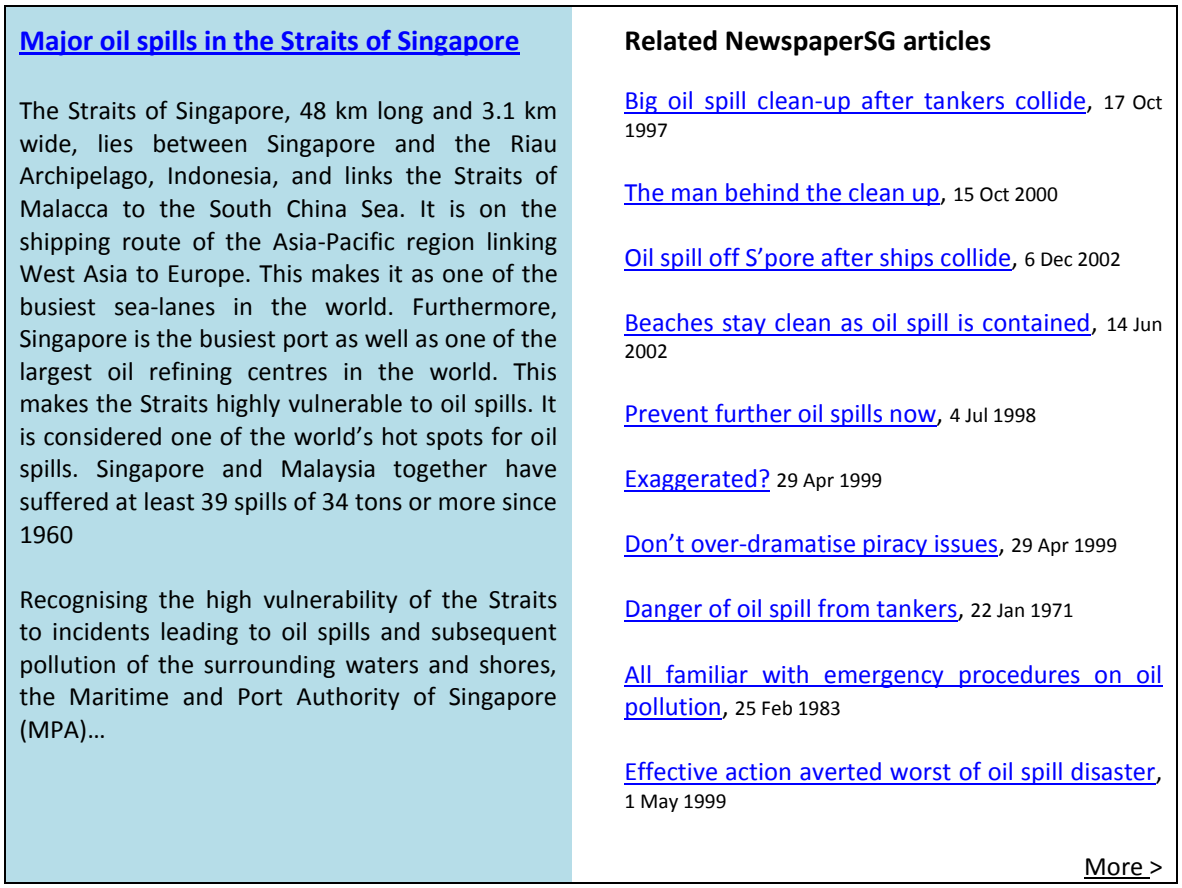

Figure 4: Related NewspaperSG articles on for a specific Infopedia article

The results showed a high level of relevance for the systems generated recommendations, and was representative of the recommendations for the rest of the articles.

7 CLUSTERING - AN ELEGANT APPROACH TO MAHOUT SIMILARITY PROCESSING FOR ENTIRE NEWSPAPERSG COLLECTION

While the use of Infopedia articles to identify related NewspaperSG articles was a breakthrough, we wanted to do more. We wanted to be able to process the entire NewspaperSG articles, and the Infopedia approach came with the following limitations:

- Infopedia articles cover only key Singapore history, culture, people and events. There will invariably be gaps even for Singapore news. Moreover, regional and international news cannot be catered for.

- The Infopedia articles are primarily written in English, with very few written in Malay and Chinese. We are therefore not able to easily apply the approach to the Malay (Berita Harian/Minggu) and Chinese (联合早报，南洋商报，星洲日報) newspapers within NewspaperSG[8].

[8] The Tamil newspaper Tamil Murasu has been digitised, pending OCR processing. We will perform similarity processing once the OCR has been completed.

The success in the divide-and-conquer approach described in the previous section showed the way forward. We needed an automated way to break the entire NewspaperSG collection into smaller clusters that can then be efficiently processed through the text analytics software, while retaining a high level of accuracy.

And we did not have to look far, as Mahout came with a number of commonly used clustering algorithms. Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters)[9].

The Mahout K-Means Clustering algorithm was used with Cosine Distance as the distance measurement. The table below shows some sample clusters identified, and the top 50 stemmed words in the clusters.

Size of cluster	Top 50 terms
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

The clustering certainly worked well. It worked equally well when applied to the Chinese and Malay newspaper articles. 6.28 million English articles, 2.28 million Chinese articles and 0.98 million Malay articles were processed. In all, over 900 million similarity associations were identified. Table 2 shows the processing time taken.

By Language	# of articles	Processing time (hrs)
English articles	6,277,574	478.0
Chinese articles	2,280,158	223.0
Malay articles	978,665	25.5

Table 2: Processing time for NewspaperSG

On the NewspaperSG site, a new 'News Alike' service has been introduced to make the vast amount of the similarity associations available. Figure 5 shows an example of the service.

[9] http://en.wikipedia.org/wiki/Cluster_analysis

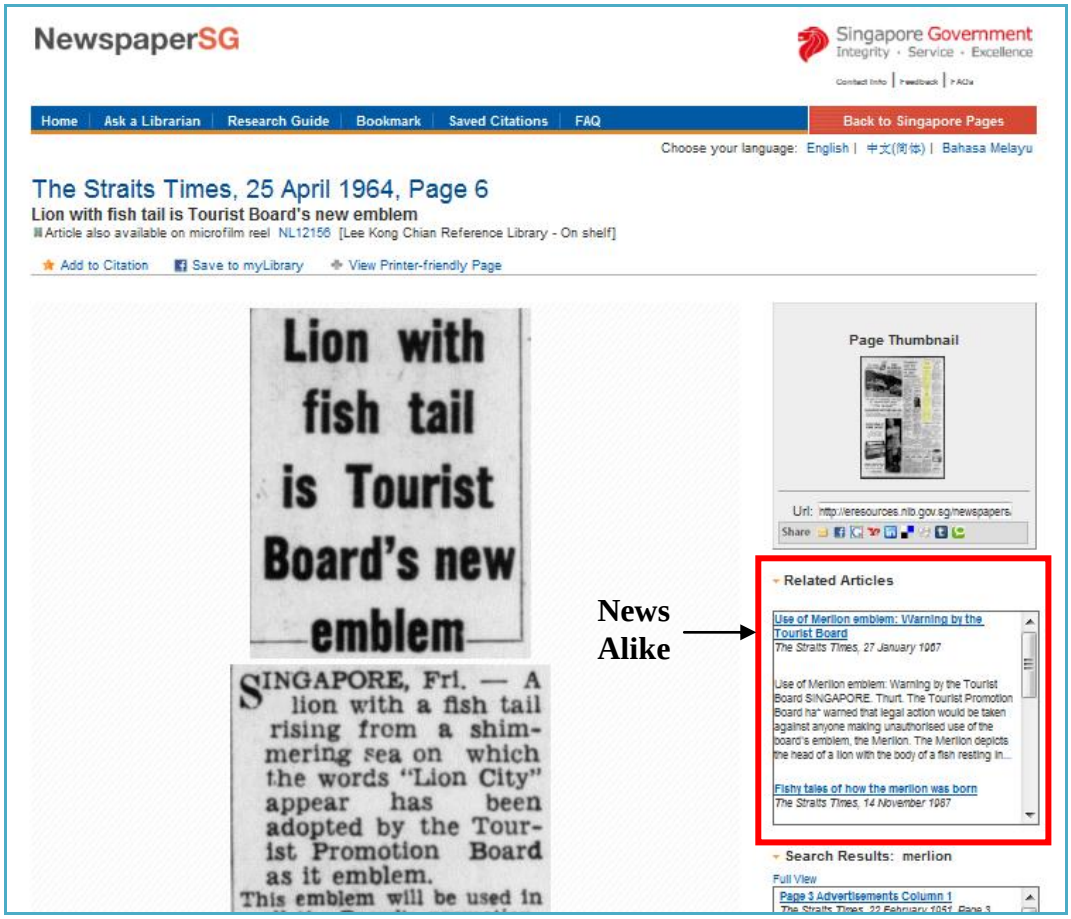

Figure 5: 'News Alike' related articles feature on NewspaperSG

Newspaper is an important repository of records in any country, and there are many initiatives all over the world to establish newspaper archives[10].

The use of text analytics makes it possible to identify a much larger pool of related articles, as compared to one that is manually identified. Researchers in particular will find it very useful to be able to view a large number of recommendations on the topics they are researching on.

Without this capability, the researcher would need to perform numerous searches with various keywords (and keyword combinations), scan through a large number of articles in the search result lists to identify the relevant ones. The text analytics processing is akin to just doing all that, albeit more thoroughly and consistently, and delivering to the researcher a cluster of the most related resources - a dossier of sort. The researcher can now spend his valuable time analysing the dossier for new insights.

The ease of discovery (just a click away) will also encourage the casual information seekers to read more, and in the process deepening their understanding and appreciation of the trusted resources available in NLB.

[10] http://en.wikipedia.org/wiki/Wikipedia:List_of_online_newspaper_archives

8 WHAT'S NEXT

The NewspaperSG collection contains newspapers in the four official languages in Singapore: English, Chinese, Malay and Tamil. As a multi-racial nation, it will be very useful if we are able to link newspaper articles across languages. We have started exploring the use of machine translation for this purpose.

9 CONCLUSION

Identifying related resources within a sizeable digital collection manually is very resource intensive, and would become intractable when the collection size grows beyond tens of thousands. Automated means to associate related content is becoming a critical tool for the next generation of digital libraries.

By leveraging on established text analytics and clustering techniques, it is now possible to identify similar resources based on their textual content, even if the number of resources goes into tens of millions, using a moderately sized hardware set up. The entire process can be fully automated, and will scale linearly.

NLB has managed to overcome the twin challenges of the large number of newspaper articles and the high dimension of the NewspaperSG collection. As a result, we rolled out the 'News Alike' service on the NewspaperSG online historic newspaper portal that provides highly relevant recommendations for close to 10 million newspaper articles.

The techniques employed can be readily adopted or adapted to text mining other massively big digital collections when the brute-force approach is no longer feasible. Organisations with a penchant for open source software will find many options available to them.

The next generation digital libraries will need to move beyond one that depends primarily on search, to one that brings the most relevant content across all collections into an engaging discovery and learning experience. The fact that the first Google search results page generates over 92% of all traffic from Google searches is a clear indication of the info-seeking behaviour of typical users. The casual information seekers' attention span is short, and the easier it is for them to get to the desired information, the better the reach of our resources.

By connecting content to content, we connect people to knowledge.

10 REFERENCES

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 World Library and Information Congress, 17 - 23 August 2013, Singapore.

Salton G and McGill MJ (1986), Introduction to modern information retrieval. McGraw-Hill. ISBN 0-07-054484-0.