Contributors:
Mr. Siang Hock KIA (Deputy Director)
Mr. Chee Kiam LIM (Senior Solution Architect)
Mr. **Balakumar CHINNASAMY** (Solution Architect)
Ms. Cally LAW (Project Manager)
Mr. Peter PAK (Senior Project Manager)

National Library Board, Singapore (NLB)

Presenter:
Mr. Siang Hock KIA (Deputy Director)

Title:
News Alike – Text analytics to link related NewspaperSG articles
In support of the following sub-theme: Text mining newspapers and news content

Abstract:
NLB explored the use of text analytics to associate its vast and growing collections of unique and valuable Singapore content. The first service to benefit from text analytics is Infopedia in June 2013. Infopedia is an electronic encyclopedia on Singapore's history, culture, people and events. We have progressively extended the feature to other NLB content-rich services.

The text analytics on the NewspaperSG collection however posed seemingly insurmountable challenges due to its sheer size of over 20 million published newspaper articles, and the presence of OCR errors as a result of the digitisation process.

An innovative divide-and-conquer approach has been implemented. We started by automatically extracting the key terms in the 1,700 Infopedia articles. Each set of key terms was then used to identify 20,000 best-matched newspaper articles from NewspaperSG. As a result, 1,700 newspaper article clusters were formed. Text analytics were then applied to each of the clusters. Over 130 million high quality related article associations were identified. The text analytics processing took 9 days to complete. The entire process is fully automated and scales linearly.

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. Without this capability, researchers would need to perform numerous searches with various keyword combinations, and scan through large number of articles to identify the relevant ones. Text analytics processing is akin to doing all that, albeit more thoroughly and consistently. It delivers to the researcher a cluster of the most related resources - a dossier of sort. The researchers can now spend their valuable time analysing the content.

This paper will focus on the lessons learnt and the process in the successful implementation of the text analytics in NewspaperSG and Infopedia.

Biography of presenter:

Siang Hock is currently the Deputy Director overseeing IT Architecture and Innovation at the National Library Board of Singapore (NLB). In this role, he and his teams are heavily involved in the conceptualisation, proof of concepts (PoCs), design and development of various innovative services at NLB.