Why BASIS tokenization is not working for Chinese records? What can we do?

Alan Ng
Systems Librarian
The Chinese University of Hong Kong Library
Agenda

• about CUHK Library
• Chinese character 101
• how it all started
• about BASIS tokenization
• problems of BASIS
• a workaround solution
• BASIS problem part 2
CUHK Library

- established in 1965
- 7 branches
- ~220 staff
- ~60K current patrons
- 127K journal, 4.6M ebooks, ~2.6M printed vol.
- special collections includes oracle bones, Chinese rare books, modern Chinese literary
Main Library
Chinese character 101

• in English, each word is composed of a string of characters

• but in Chinese language, each character (ideograph) has its own meaning and usage

• each Chinese character can be used independently

• in Chinese context, each character is actually a word
Chinese word 101

• single Chinese word examples:

• 我 (me) 你 (you) 大 (big) 小 (small) 書 (book)

• individual Chinese words combine to become a Chinese term or phrase, e.g.

• 我們 (we) 圖書館 (library) 電子書 (ebook)
Chinese word 101

- forming Chinese phrase from a stem
- 圖書 (book)
- 圖書館 (library)
- 圖書館員 (librarian)
How it all started?

- in 2013, I got a simple user enquiry:
- why searching “陪審” would obtain fewer results than “陪審團” in Primo?
- “陪審” means “the act of being a jury member”
- “陪審團” means the jury
- it just didn’t make sense, started a 13-month journey
What is BASIS?

- not mentioned widely
- in “Primo Version 4.x Highlights”, it is for improving search result relevancy for Chinese language
- implemented since Primo 4.3
- not much further info
BASIS tokenization

- analysis is done during indexing and searching
- the PNX record <search> fields are subject to tokenization
- e.g. the TITLE “諮詢文件：出任陪審員的準則” is broken into:
  - “諮詢” “文件” “出任” “陪審員” “的” “準則”
- Primo only indexes these six tokens
- n-gram tokenization
BASIS tokenization

- the idea was to only index specific Chinese phrases instead of individual word
- based on assumption Chinese users would search for Chinese phrases, instead of arbitrary single Chinese word
- leading to fewer but more relevant search results
- but also ignoring a lot of matches
How a match is found under BASIS?

- not just the PNX data, the search terms will go through the same BASIS tokenization
- to retrieve an entry, you must have a matching token in your tokenized search terms
How a match is found under BASIS?

- to find the TITLE: “諮詢文件：出任陪審員的準則”

- your tokenized search teams must have either: “諮詢” “文件” “出任” “陪審員” “的” “準則”

- if you search with a stem of the token (e.g. “陪審”), you won’t get it

- it is problematic and unacceptable
Case 1 (6 entries for “陪審員”)
Case 1 (3 entries for “陪審”)
What is wrong?

• stem “陪審” gets far fewer matches (6) than the longer term “陪審員” (3)

• both result sets mutually exclusive

• but the search terms share common characters
Case 2 (521 entries for “馬來西亞”)
Case 2 (124 entries for “馬來亞”)
Problems about BASIS

• Chinese language is too complicated

• the tokenization logic is unknown

• the Chinese phrase dictionary used by the BASIS process is unknown

• no recursive tokenization, stems are ignored

• e.g. “陪審員” won’t be further broken down into “陪審”, “員” and “陪” “審” “員”
Problems about BASIS

• also affect Japanese records using the “kanji” (漢字), which shares the same Chinese characters

• Oxford University Library found the same problem in May 2015

• any Primo with Chinese / Japanese record is affected
Chinese / Japanese record discovery fails

- Nevertheless, an exact title match (including punctuations) can always retrieve the record
- because both raw data and search term will go through the exact same tokenization
- it became a “known search”
- purpose of discovery defeated
A workaround solution

- the BASIS current approach simply doesn’t work
- BASIS needs to be OFF
- needs to go back to index every single Chinese character separately (1-gram tokenization)
- Ex Libris acknowledged the problem in Oct 2014
- help CUHK turn off BASIS and reindex
- please refer to support case #00092878 (cuhk)
became a known issue

- listed among the Known Issues for both Primo 4.9 and April 2015 release
- solution still pending
Suggestions

• enhance the tokenization dictionary to handle the stems within the Chinese phrase

• do recursive tokenization, detect the stems

• keep both n-gram and unigram (1-gram) tokenization for enhanced relevancy and complete result set
BASIS problem part 2?

- searching a string of over 30 Chinese character
- huge HPROF file (10+GB) spawned in FE se_bin folder
- "/exlibris/primo/p4_1/ng/jaguar/home/system/bin"
- will hang the Primo FE
- support case #00119959 (cuhk)
- permanent fix expected in Primo 2015 November release
Q & A
Thank you!

Contact: alanng@cuhk.edu.hk