40

Full-Text date-base development in ROC

Prepare for
The Information Technology and Pacific Rim session
of
LITA 1992, Denver Colorado

By
Ching-chun Hsieh
Professor and research fellow,
Institute of Information Science
Academia Sinica, Taipei

Sept 16, 1992

Full-Text data-base development in ROC

The study of full-text data-base has been started since the early 1980s in Taiwan, ROC. Owing to the intrinsic characteristics of Chinese language, the technology developed for Chinese full-text data base is quite different from that for English. Some major differences are listed in Table-1.

Recently, the Chinese language full-text data-base technology is gaining in maturity. There are 4 full-text DBMS available in Taiwan. Full-text data bases are available in classical literature studies, legislative information services, news services, abstract search in library services, etc. [1] The number of available full-text data bases are around 30 now in Taiwan, and more than 70 for mainland China, Hongkong and Taiwan all together. [2] But, so far, there is no full-text DBMS developed from mainland China of Hongkong.

The demand of full-text data base is increasing rapidly in Taiwan, and we expect more full-text data bases will be available in the coming year. As a conclusion, let me list some major development events that you might be interested to know as follows:

1. Full-text data-bases in CD is available since 1991, for example, the 25 dynasties data-base.[3]

2. Word identification has become a sophisticated technology since 1991. There are 4 such package are available now, with worst case error rate between 3% to 0.3%.[4]

3. Some achievements from computational linguistics study will certainly improve the full-text processing capabilities in the future. They are [4]:

   (1) The number of words in machine readable dictionary has exceeded 100k since 1990.
   (2) A modern mandarin corpus database of more than 10 million characters are available since 1990.
   (3) A prototype of an information based unification grammar complier for sentence parsing is successfully developed in June, 1992.
   (4) Some statistical properties of Chinese characters in classical literatures has been explored since 1991.

4. A Chinese version of the Standard General MakeUp Language (ISO 8879 SGML) parser and editor will be available by the end of 1992. [5] This parser and editor will provide a text sharing environment for all text processing utilities, including the full-text databases.

5. A preliminary result from a news auto - classification project shows that nearly 90% correctness can be achieved. This means the document- term relationship in Chinese language might be more helpful than that of English for text retrieval.[6]

Thank you for your attention. Any comment from you will be appreciated and I will be glad to answer any questions.

Notes

[1] 卜小蝶,〈全文資料庫糸統之技術發展與中文應用之探討〉,中央圖書館館刊 (Proceedings of the National Central Library) vol25, No1, June 1992.

[2] From the list provided by 社會科學中文電腦研討會 (廣東省汕頭市 Sept. 3-6,1992), and by personal contact in Taiwan.

[3] By the Computing Center, Academia Sinica, Taipei, July 1992.

[4] Those related information can be obtained from the ROC Computational Linguistic Society, P.O.Box 1-7, Nankong Taipei, ROC.

[5] Please cantact professor S.M. Chuu, Yuan-Ze Institute of Technology Dept. of Computer Science, Nei-Li, Chung-Li, 320 Taiwan ROC.

[6] Please contact the author.

Table 1 Full-text processing related properties between English and Chinese

Items	English	Chinese
• General		
1. text size	relatively large	relatively mall
2. Statistical structure knowledge of language	more sophiscated	need more study
3. linguistic knowledge of language	more sophiscated	need more study
• Language structure		
4. Capital letter	yes	--
5. Indexing level	word,phrase	character,word,phrase
6. Variant	word	character(serious)
7. String length between punctuation symbols	relatively long	relatively short
8. name set of person	reatively closed	open
9. foriegn terms	--	need special care
• Technigues		
10.hyphentation	needed	--
11.word identification	simple	complicated
12.stop list	available	depends on word identification capabil
13.pattern(string)matching	complicated,need suffix stripping to produce word stems, relatively slow	straight forword and relatively fast
14.text signature	yes	different approach

Table 2 Full-text DBMS

item	by	example of use
1. Chinese Text Processor (CTP)	The Computer Center Acadenica Sinica, Taipei in 1988	25 Dynasties db 十三經 db 大藏經 db
2. 虹成全文系統	虹成公司,Taipei,1989	四書(four books) 唐詩等
3. 龍泉系統	Prof. 陳郁夫 National Normal Univ. Taipei,ROC,1991	左傳、儒林外史 說苑等
4. Join Full Text Retrieval System (JFTR)	Join Computer Corp.,Ltd 1992	台灣省教育論文資料庫 Full-text db of essays in education