IDEOGRAPHIC STATUS
TEXT STORAGE AND RETRIEVAL IN CHINESE AND ENGLISH

Richard Jones, Douglas Kelly and Sek Kit Leung
Computer Power Group
Cnr Geils Circuit & Denison Street, Deakin, ACT 2600

INTRODUCTION

Full text retrieval is concerned with locating relevant documents by looking for word and phrase matches from an automatically generated index or concordance. This involves indexing virtually every word in sufficient detail so that its position relative to any other indexed word can be obtained. This approach is in contrast to keywording which is (at least until recently) a purely manual process and which relies on selection of descriptors often from a fixed vocabulary controlled by a thesaurus.

THE STATUS SYSTEM

The STATUS full text management system has been available as a commercial offering since 1974. It is owned by Harwell Computer Power Ltd. (HPC), a UK company. Computer Power Group (CPG), the largest Australian software company, has had a strong and varied role in STATUS since 1981. It is the major retailer of the product covering Australia, the USA and the Pacific Rim. It is part owner of HPC, and perhaps most significantly, CPG's Canberra laboratory continues to develop and deliver new technology for the product.

A major feature of STATUS is its intelligent query capability (IQ) developed by CPG. This provides a practical way of measuring the relevance of a document to a query, and four years after its introduction, still sets STATUS apart from its competitors, as has been seen in the remarkable success of the product in the USA. AIDA, CPG's revolutionary new document analysis capability is being examined at HCP, to determine its relevance to STATUS.

THE PROBLEM

From the beginning STATUS has had a multilingual flavour. Its earliest sales were in the Netherlands and Scandinavia. The software has been able to handle a wide variety of European alphabet extensions for many years. However, in early 1986, the CPG office in Hong Kong asked if STATUS could be extended to handle Chinese and other ideographic character sets. This caused some consternation.

The technical challenge that lay under what seemed (to them at least) a simple request, was how to move STATUS from

handling languages with alphabets of fewer than 80 characters (including upper and lower case) to an ideographic script with in excess of 30,000 characters, no collating sequence and no word break. Further, unlike the Roman alphabet which has one (or including IBM two) internal representation across all computers, there is no international standard. Different vendors pursue their own representations of differing lengths from two to four bytes. There is now an emerging standard; it remains to be seen if this becomes a commercial reality.

STATUS IN CHINESE

Between July 1987 and May 1988 the first version of Ideographic STATUS was developed in Canberra. The development platform was a Wang computer running the ideographic form of their operating system, and the two person development team started from the Wang version of STATUS.

The decision was taken to overcome the lack of a word break by indexing on individual characters, and defining 'stop characters'. The index was expanded to store both European words and Chinese characters in the same text base. Display turned out to be quite straightforward, as did the provision of STATUS macros and synonym lists, which can be multilingual. The user interface was translated. Perhaps the major problem for the non-Chinese member of the team was writing Chinese characters, and the Greek alphabet was used to good effect!

SICR - A STANDARD IDEOGRAPHIC REPRESENTATION?

In October 1989 the project was transferred to CP Taiwan and a Research and Development (R&D) group established there to continue development. Different vendors in the market-place approached CPG to port the software to other machines and the problem of a lack of internal representation became very apparent. An internal STATUS standard was defined: SICR, the STATUS Ideographic Character Representation. This has enabled Ideographic STATUS to be moved to DEC, IBM and HP equipment while maintaining the common base of software. The characters are translated back and forward into the local machine representation for input and printing.

STATUS IN THE MARKET-PLACE

The ideographic version of STATUS has opened up a new market for Chinese language text retrieval systems. In Taiwan, access to vast collections of domestic and international scientific and technical material is provided by the National Science Council. Through their Science and Technology Information Centre (STIC), a nationwide service (STICNET) is providing indexes and abstracts of

international databases such as BIOSIS, CA SEARCH, COMPENDEX, MEDLINE, ERIC, INSPEC and NTIS, using STATUS databases.

Since 1988, online searching of domestic databases, including current periodicals, technical reports, monographs, etc., has been available in the Chinese language. Using Ideographic STATUS on Wang, the STIC is providing several thousand users nationwide with over 700,000 titles in Sci-Tech information as Chinese and English text together.

Also in Asia, the Attorney General's Department in Singapore and Hong Kong have developed legal information retrieval systems using STATUS. Their intended direction is towards law libraries and legislative drafting with Chinese language support. The Hong Kong Legal Department now has Ideographic STATUS for Chinese and English text support. The Bilingual Law Drafting System is in preparation for the return of Hong Kong sovereignty to China.

THE FUTURE

The commercial success of Ideographic STATUS means that the technology can be enhanced. One option that has been discussed includes the use of some of the recent heuristics for breaking a string of characters into words to speed up the index and reduce its size. Another is the feaibility of introducing IQ technology. Perhaps further away is AIDA.

CONCLUSIONS

Ideographic STATUS has become the *de facto* standard for Chinese text retrieval systems and continues to attract a lot of attention. It will continue to make sales for many years.

It is interesting that little thought was actually given at the time to questioning if the heuristics that underpin full text retrieval in European languages are also valid in Chinese. They seem to be, and perhaps this is due to their very primitive nature. However the intriguing question is, are there heuristics in Chinese which improve the retrieval process that we have merely translated from English in what is perhaps a simple minded way linguistically? Anyone out there with any ideas?