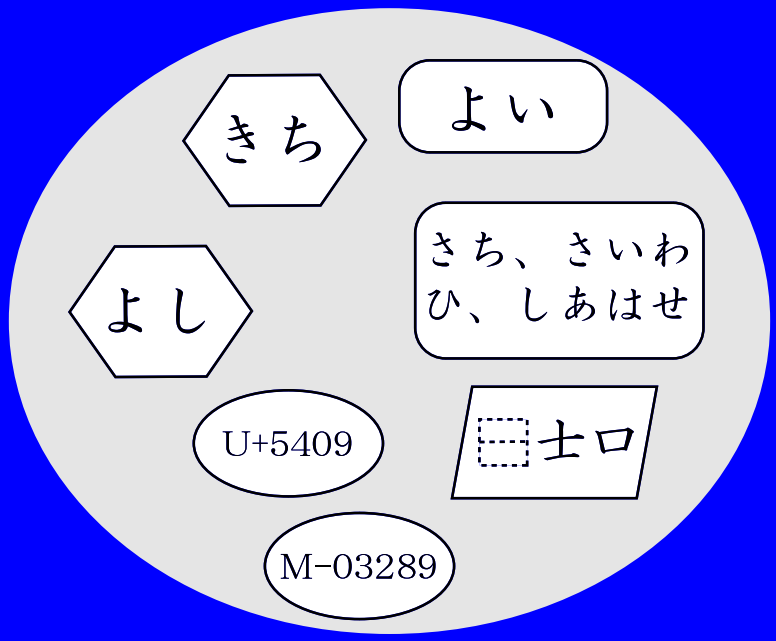
Character Processing based on Character Ontology

MORIOKA Tomohiko

2005-01-22

CHISE Project

Chaon model

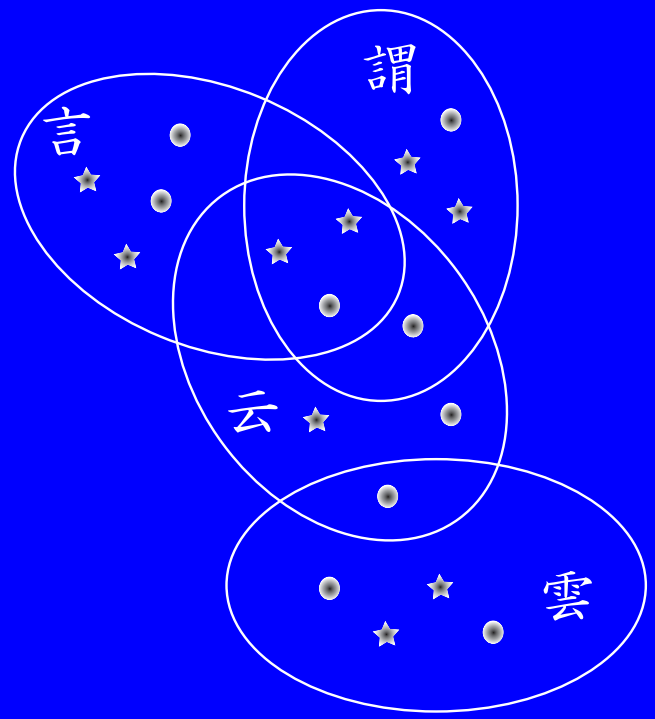


Each character is represented by its features

Comparison between character representation methods

type	easiness	expressivity
integer	◎	×
set	○	○
tree	△	○
network (graph)	×	◎

Set operations for character features



Character operations in Chaon model

- Define/put characters · character features
- Get character features
- Search characters by character features

Implementations: by CHISE project

Fundamental library: libchise

Editing system: XEmacs CHISE

Scripting languages: Ruby/CHISE, Perl/CHISE

Multilingual Typesetter: Ω/CHISE

Automatic Ideographic Glyph Generator: KAGE

Character databases for CHISE

1. General database distributed as a part of XEmacs CHISE (CHISE basic character database)
2. Database about structure information of Ideographs (CHISE IDS database)

CHISE basic character database

- Basic database for multi-scripts
- ‘define-char’ format

  Character features are represented by Lisp (XEmacs CHISE) function ‘define-char’
- About 100000 characters (in terms of character objects in CHISE database)
- UCS Unified Ideographs may be separated (now introducing multilevel unification rules)

Category of character features

1. General properties (like items in dictionaries)
2. Mapping to ID of character
   - two-way mapping (=`CCS`)
   - one-way mapping (=>`CCS`)
3. Relations between characters (`->`*`active-direction`*, `<-`*`passive-direction`*)

CHISE IDS database

Machine readable expression about combinations of components of Ideographs (解字)

- *Ideographic Description Sequence* (IDS) of ISO/IEC 10646 is used as the format
- non-UCS characters are also used as components (if it is not found in UCS)

File Edit View Cmds Tools Options Buffers

Open Dired Save Print Cut Copy Paste Undo Spell Replace Mail Info Compile Debug News

U-00020020	廿	⿶廿乂
U-00020021	先	⿱一先
U-00020022	此	⿱一此
U-00020023	𠂤	⿱一⿴几丨⿱七一
U-00020024	𡗜	⿱⿰大大一
U-00020025	亘	⿱一⿰丨日一
U-00020026	州	⿱一州
U-00020027	巴	⿴巴三
U-00020028	及	⿰下及
U-00020029	丂	⿰⿰丂丂丁
U-0002002A	甘	⿱甘丶
U-0002002B	甘	⿱甘丿
U-0002002C	丙	⿶丙人
U-0002002D	禾	⿱禾刀
U-0002002E	𠂉	⿱一八⿶口乂
U-0002002F	凶	⿱一凶

IDC

	2FF
0	2FF0
1	2FF1
2	2FF2
3	2FF3
4	2FF4
5	2FF5
6	2FF6
7	2FF7
8	2FF8
9	2FF9
A	2FFA
B	2FFB

Ideographic description characters

These are visibly displayed graphic characters, not invisible composition controls.

2FF0 IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO RIGHT

2FF1 IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO BELOW

2FF2 IDEOGRAPHIC DESCRIPTION CHARACTER LEFT TO MIDDLE AND RIGHT

2FF3 IDEOGRAPHIC DESCRIPTION CHARACTER ABOVE TO MIDDLE AND BELOW

2FF4 IDEOGRAPHIC DESCRIPTION CHARACTER FULL SURROUND

2FF5 IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM ABOVE

2FF6 IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM BELOW

2FF7 IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LEFT

2FF8 IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER LEFT

2FF9 IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM UPPER RIGHT

2FFA IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER LEFT

2FFB IDEOGRAPHIC DESCRIPTION CHARACTER OVERLAID

Demo

- XEmacs CHISE
- IDS

Conclusion

- Chaon model: each character is represented by a set of character features
- Chaon mode has enough expressivity and not so complex (easy to implement)
- CHISE Project are developing implementations based on Chaon model and some results are available
- CHISE character databases