マルチリンガルエディタにおける検索機能の設計(1)

1ZA－6

基本要件の抽出と発展型の提案

片岡朋子　上園一知　筧捷彦†
早稲田大学メディアネットワークセンター
†早稲田大学理工学部

1. はじめに

近年におけるネットワークの著しい発達に伴い、世界のどこにあっても任意の言語・文字を混在使用（入出力、文書処理、通信）できることがますます必要となっている。基本文書処理において言語情報は不要であり、世界の文字に共通の原理で挿入・削除・検索・置換等を行うことが可能である[1]。

本稿では検索機能について、「基本検索」と「発展型検索」とに分類し、前者は文字の同定情報のみに基づくが、後者は文字の表示に関する情報に基づくものであり、「文字属性」と「フォント属性」を峻別することで、個々の検索機能を明示できることを見る。

2. 文字属性とフォント属性

一文字のもつ情報としては、「文字 ID」以外に「表示に関する情報」「文字集合 ID」「文字コード単位(mb)への変換情報」が存在する[2]。表示情報として、「図形配置開始位置」「改行方向」「行配置開始位置」「禁則処理情報」「Direction」「Form」「Variant」「Ligature」などがある。これを「文字属性(character attributes)」と定義する。

一方、「フォント属性(font attributes)」としては、「ポイント数」「Weight（例：ボールド）」「Posture（例：イタリック）」などがある。

検索の種別・範囲を明確にするために、文字属性とフォント属性を区別し議論を進める。

3. 基本検索機能

任意の文字種の混在を可能とするマルチリンガルエディタの行う文字列処理とは、一般化された文字の情報にのみ基づく、挿入・削除・置換・検索等をいう。文字の表記方向が混在する可能性があるため、メモリ上の文字の順序と表示上の文字の順序が必ずしも一致せず、一定の考慮を必要とするが[2]、ここでは割愛する。

文字列処理は一般に表示を伴って行われるために、特に検索においては文字の検索か図形の検索かを明確に区別せず、あるいは意識さえすることなく行っている可能性がある。全世界の文字を調査・分析した結果、すべての文字種に共通の特性を抽出し、一般化したが、このうち文字の同定に必須な「文字ID」に基づく検索を「基本検索」と定義する。

4. 発展型検索機能

2節で定義した基本検索機能以外に、文字属性、フォント属性およびその組み合わせから成る検索機能がある。これを発展型検索機能と呼ぶことにする（表1）。

4.1 文字属性に基づく検索

ペルソ・アラビック系文字は文字が連続して筆記され、同一文字が単語中の位置（語頭/語中/語末）により異なるフォームをもつのが特徴であり、特定

Designing Search Functions for a Multilingual Editor (1):
Its Essential Functions and the Extended Search Possibilities.
Tomoko I. Kataoka, Kazutomo Uezono and Katsuhiko Kakehi
Waseda University
{tomoko, uezono, kakehi}@kake.info.waseda.ac.jp

文字の検索以外に、特定フォーム（例：語中形のみ）の検索機能が必要である。

表1：発展型検索機能

文字属性に基づく	フォント属性に基づく	文字属性＋フォント属性に基づく
特定フォーム検索	同図形異文字検索	ゴチック体特定フォーム検索
特定バリアント検索	特定ポイント文字列検索	特定ポイントのリガチャ検索
リガチャ検索	ゴチック文字列検索	ゴチック＋斜体特定バリアント検索
Discontinuous sequence 検索	イタリック文字列検索	など
	アンダーライン文字列検索	
	ゴチック＋イタリック文字列検索	
	など	

モンゴル文字系では更に、特定フォームに異体字(variant)が存在するので、この検索も可能とすべきである[3]。多くの文字種に存在する合字(ligature)中の各文字検索は基本検索であるが、リガチャ自体は図形すなわち発展型検索に属する。

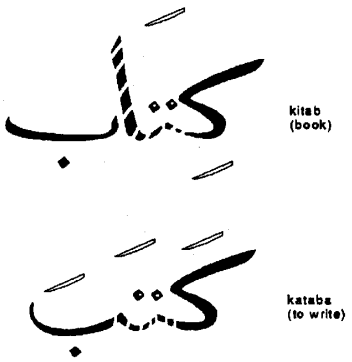

図1：Discontinuous sequence 検索

セム言語族に一般にみられる現象として、動詞や名詞などの「非連続３子音」があり、語源や派生語の研究において必須の検索となる。アラビア系文字は子音文字を土台にした表記体系であり、母音記号を記述しない場合は単純な検索であるが、母音記号を記述した文字列と記述しない文字列で同様の検索結果を得るために、特定の Diacritic Mark を無視する検索方法が保証されなければならない。

4.2 フォント属性に基づく検索

モンゴル系文字では、異なる文字が同一の表示図形（同一フォント）で表現されることがあり、目的によっては表示に着目した文字列検索も可能とすべきである。

フォント属性情報を利用して、文字列の内容には無関係に特定ポイント（以上・以下）の文字列を検索したり、いわゆる強調を施した文字列を検索する場合もある。

4.3 文字属性とフォント属性に基づく検索

特定フォームかつ強調フォントで表示された文字、特定ポイントのリガチャ検索など、文字属性とフォント属性の双方の情報を含めた検索もサポートする必要がある。

5. おわりに

文字列検索を「基本検索」「発展型検索」とに分類し、また、文字属性とフォント属性とを峻別することで、発展型検索の種類を明確に定義できた。

既に X Window System 上で開発済みの国際化ライブラリを用い、挿入・削除等を含め、基本検索機能を実装したエディタを作成、発展型機能を準備する予定である。

参考文献

[1] T. Kataoka, et al. The Worldwide Multilingual Computing (4): Essentials for the Multilingual Text Manipulation, Proceedings of the 51st General Meeting, IPSJ, 1995, pp.251-252.

[2] 上園一知他: マルチリンガルエディタにおける検索機能の設計(2): 検索機能の実現, 本冊, 1999.

[3] T. Kataoka, et al. Internationalized Text Manipulation Covering Perso-Arabic Enhanced for Mongolian Scripts, EP'98, 1998, pp.305-318. Springer-Verlag.