Faculty of Engineering
Tokushima University

特定文書集合へのインタラクティブテキストマイニング

[キーワード：テキストマイニング，接尾辞配列]　講師　吉田稔

図1：テキストマイニングによる検索支援の概念図

図2：テキストマイニングによる検索支援システム

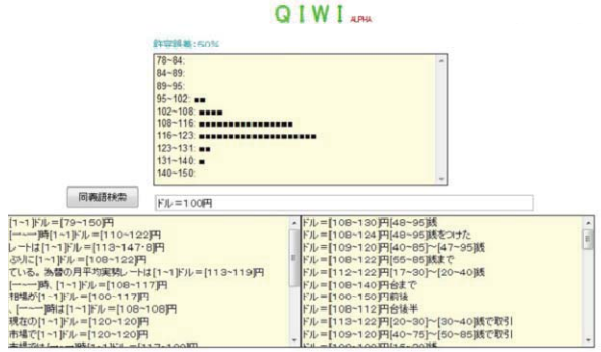

図3：テキスト中の数値表現マイニングシステム

内容：

近年，WWW上や組織内に蓄積される電子的文書の量は増大の一途を辿り、特定組織のWeb文書集合、Wikipedia、さらには企業内文書集合等、様々な文書集合（コーパス）において電子的文書のサイズが増大し、把握が困難となりつつある。

この状況に対し、我々は、「リアルタイムテキストマイニングによる検索支援システム」（図1）を提案している。テキストマイニングとは、与えられたテキスト集合の中での、「言葉の使われ方」（主に、言葉に関する統計的情報）について分析するタスクである。接尾辞配列というデータ構造を活用することで、入力されたクエリに対し、「用例抽出」「同義語抽出」という二種類のテキストマイニングをリアルタイムに行い、マイニング結果を提示することで、検索支援に役立てる（図2）。

また、「テキスト中の数値情報マイニング」に関する研究も行っている。テキスト情報の中には、「25歳」「10000円」等、多くの数値表現が含まれている。我々は、数値範囲を検索クエリとして用いる検索機能を備えた新たなテキストマイニングシステムを提案している（図3）。

分野：知能情報学

専門：テキストマイニング

E-mail: mino@tokushima-u.ac.jp

Tel.　088-656-9689

Fax:　088-656-9689