WWWの記事をその内容によって自動的に分類する

新納浩幸

1. 文書のクラスタリングとその困難性

2. 文書クラスタリングの目的
WWWの記事その内容によって自動的に分類する

また近年、社会の電子化が進み、電子的文書の量が増

えていることから、なんらかの処理をする必要があ

る。一つは文書の内容全体を把握する処理であり、他

は文書の内容全体を記載する処理である。これら

の処理をととめるのが、WWWの対象とした日本語研

究である。

このように、WWWの記事の内容全体を把握する処

理である。一つは文書の内容全体を記載する処理

である。これらWWWの対象とした日本語研究である。

このように、WWWの記事の内容全体を把握する処

理である。一つは文書の内容全体を記載する処理

である。これらWWWの対象とした日本語研究である。
図4 Yahoo!のウェブディレクトリ

図5 「ビジネスと経済」のサブカテゴリー

図6 「アップル」のサブカテゴリー
文書3
(0000000)

文書1
(1111111)

文書2
(2222222)

文書3
(3333333)

文書2
(4444444)

文書1
(5555555)
WWWを対象にした日本語研究

谷の類似性を、どのような方法で数値化できるかである。現状は「[A川][B川]」のモデルを用いて文書をベクトル化し、ベクトル間の余弦値により類似度を定義するのが主流である。そしてそれをベースに様々な改良ができ試みられている状況である。代表的な改良として、単語を新概念に当てはめるN次ベクトルを用いて、類似度を定義する方法がある。例えば、「直系が高い」といった文に「ノードが深い」という文章が低いという文法が成立する。つまり、文書のクラスタリングの改善に繋がる。これによって例えば「石で破壊する方法」と「石を壊す方法」などが選別される。あるいはWWW上の文書においては、文書のメタ情報「リンクをキーワードなどの情報」も取

特集 "WWWを対象にした日本語研究"