キーワードの含有率を用いたクラウドファイルシステム

木下研究室　　竹村孝太 (200902763)

1　まえがき

インターネット上には多種多様な情報（ファイル，ページ）が散在している．これらの情報から必要なものだけを取り出すのは困難である．インターネット上のファイルをある規則の元に扱いやすいように再構成した上でネットワーク透過的に利用できるシステムをクラウドファイルシステムという．クラウドファイルシステムの構成要素の１つとして視覚的に判りやすいインターフェースを作ることがあげられる．これにより必要な情報のみを取り出すことが容易となってくる．

そこで，ファイルの中身に注目して関係性のあるファイルを集めるシステムを考える．この際の集まる力について“情報同士の関係性”，“情報の重要度”に着目して考察する．

2　提案

2.1　キーワードの含有率

ファイルの内容に着目して関係性のあるファイルを集めるために「キーワードの含有率」を提案する．このキーワードの含有率によりファイルの関係性と重要度を定める．関係性がファイルの集まる力となって群れを作り，重要度により群れの中のファイルの位置が決まる．ファイルの位置は重要度が高いほど群れの中心になり，重要度が低いほど群れの外側になる．

キーワードの含有率とは，ファイル内にキーワードがどの程度含まれているかを示すものである．同じ言葉でも違う意味であったり，違う言葉でも似た意味であったりする言葉の持つ意味の曖昧さ（家族的類似性）も考慮されることが望ましいが，まずは以下のものを含有率の要素として考える．

- キーワードの個数：ファイル内のキーワードの数
- キーワードの色、大きさ：フォントの違い
- キーワードのある場所：ファイルの題名や括弧の中など

今回は上記の要素の中で，キーワードの個数のみに着目して含有率を出すこととする．以下にその式を示す．

- $CBP = C\frac{K}{A} \times 100$　(CBP:含有率　C:個人による補正 K:キーワードの個数 A:ファイル内の総単語数)

式中の個人の補正とは，含有率による重要度と各個人の主観的な重要度の溝を埋めるための係数で各ファイルに任意に設定できるものである．

2.2　キーワードの提示

目的に応じた群れを作るには，キーワードの提示の仕方も重要である．そこで，キーワードの提示方法について考える．

まず，キーワードは単体で意味を成す語であるとする．単体で意味を成さない語ではファイルに関係性を持たせられないからである．次に，キーワードの数や組み合わせを考える．キーワードの意味の幅が狭いなら単体で提示し，意味の幅が広いなら複数のキーワードを組み合わせて提示するのが有効であると考えられる．複数のキーワードを提示する際に，以下のそれぞれの方法を使い分ける必要がある．キーワード A，B の持つ意味の集合 Am，Bm について，

Am ∪ Bm ：家族的類似性の範囲が広く，多くのファイルの関係性がみられる．

Am ∩ Bm ：家族的類似性の範囲が狭く，特に重要なファイルのみが集まる．

3　含有率の算出

キーワードの含有率について，まずはキーワードの個数のみを考慮したものを求める．これには個人の補正も含まれない．それにあたり，茶筌と perl を利用する．茶筌は形態素解析のツールで，ファイル内の文章を形態素ごとに分解し，それぞれを品詞分類できる．そして品詞分類をしたファイルを perl のプログラムにかけることで，任意のキーワードの個数を求める．

プログラムの実行結果

```
C:\Users\takemura kota\Desktop\perlbox>perl
ganyuuritsu.pl
キーワードの数を入れて下さい :5
キーワードを入力して下さい：季節
キーワードを入力して下さい：春
キーワードを入力して下さい：夏
キーワードを入力して下さい：秋
キーワードを入力して下さい：冬
季節:0 春:0 夏:1 秋:1 冬:0
総単語数：231 キーワードの合計数:2
含有率:0.865
```