内容的なまとまりをもつWebページ群の自動判定

石田栄美（慶應義塾大学大学院）　久野高志（作新学院大学女子短期大学部）
安形輝　（亜細亜大学）　野末道子（鉄道総合技術研究所）　上田修一（慶應義塾大学文学部）

1 はじめに

1989年にティム・バーナーズ・リー(Tim Berners Lee)らによって提唱され，1990年に実装されたWorld Wide Web（以下，WWW と略す。）は，10 年足らずの間にブラウザーやサーチエンジンの開発と普及によりインターネットの中心的な存在となり，新しい情報メディアとなった。WWW を構成する各ページは，シートの形態をとり，相互にリンクされ，文字，画像，音声を収録できるという特色を持っている。WWWの普及が急激であったため，WWW自体についての研究例や理解は乏しい。

一般に情報メディアは，物理的な実体とその内容とは不可分であるが，WWW も例外ではない。物理的には，「ファイル」と呼ばれる単位が基礎となり，これはディレクトリの階層構造の下に位置しており，地のテキストと付加される画像ファイルなどから構成される。これに対し，「Web ページ」や「ホームページ」は論理的な意味合いを持っている。本や雑誌の「ページ」は物理的な単位であり論理的に完結していないが，「Web ページ」は論理的な単位である。そして「ファイル」と「ページ」はほぼ１対１で対応している。

しかしながら，Web ページには定まった書法がなく，作成者は，あるまとまった内容（著作）を一つのページに格納してもよいし（図１(B)），複数のページに分割することもできる（図１(A)）。さらには，一つのページに異なる複数の内容を掲載することも許されている（図１(B)）。

一つ一つのWebページの閲覧者は，リンクを辿りながら見ていくことができるが，サーチエンジンやリンク集では問題が生じる。サーチエンジンは，ファイル（ページ）を単位としているために，検索結果であるページが，図１の(A)や(C)であることが避けられない。(A)の場合でリンクされていなければ，断片的な情報しか得られないことになる。そこで，(A)のようなまとまった内容がページ群に分割されている場合の処理を考えておく必要が生じ，次世代のサーチエンジンは，ページ群の自動判定機能を持つだろうと予想されている。

本研究では，「内容的なまとまりのあるページ集合」をページ群と定義し，(A)のような複数のページでまとまった内容を表現しているページ集合を自動判定する手法を提案する。ここでいうページ群とは

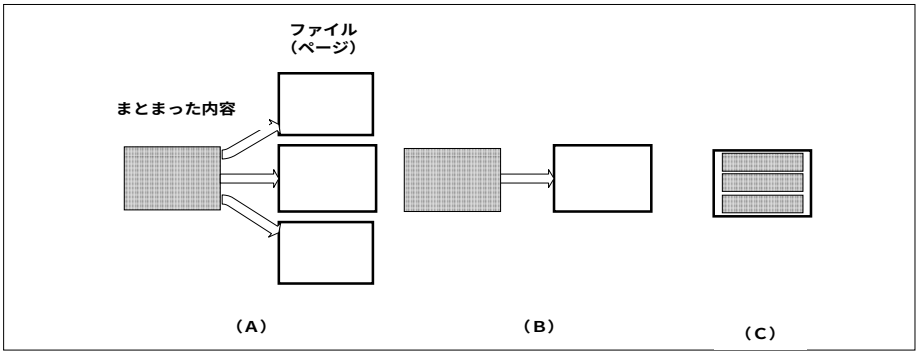

図１　まとまった内容とファイル（ページ）との関係

たとえば，論文の各章がそれぞれのファイルになっているような場合の「論文」全体や，イタリア旅行記が複数のファイルにわかれて書かれていた場合の「イタリア旅行記」全体などのことを指す。

2 先行研究

ページ群の自動判定方法については，すでに注目されており，いくつかの研究がなされている。永藤らは，ページ群を意味的に関連があるページの集合と定義し，リンク関係と内容類似度からページ群の判定を行っている [1)]。また，原田らは，サーチエンジンにおいて，まとまりのあるページをグループ単位で提示することを目的に，経験則からディレクトリ構造をもとにページ集合をまとまりをもったグループにわける方法を提案している [2)]。

これらの研究は，サーチエンジンの検索結果数の増減によりページ群の判定を評価しており，作成されたページ群の内容の分析には至っていない。

3 ページ群の判定方法

ページ群の判定は，Web ページがリンク関係やディレクトリ構造を手がかりとする構造から判断する方法とページの内容類似度など内容から判断する方法，その両方から判断する方法がある。本研究では，内容類似度を重視した判定方法とリンク関係を重視した判定方法の 2 つの方法を提案する。

3.1 内容類似度を重視した判定方法

ページ間の内容類似度を重視した判定方法は，リンク関係やディレクトリ構造などを考慮せずに内容類似度だけを基準として，ページ群を判定していく方法である。ページ間の類似度を判定するために，ページ集合に対してクラスタリング（完全連結法）を行う [3)]。類似度はページ間の単語の出現頻度から測るが，内容が類似しているページ集合はそのページのスタイルも共通している傾向がみられるので，ページ内で用いられているタグをすべて含めた場合の判定も行うことにした。

ページ群の判定は，図 2 に示した手順で行う。クラスタリングの最初は，一つのページが一つのクラスタとみなす。ページ内の単語の重みは，以下の式で求める。

$$w_{ij} = \frac{F_{ij}}{\max_{j=1,2\cdots M} F_{ij}} \bullet \log \frac{X}{x_j} + 1$$

ここで，F_{ij} は単語 t_j のページ Di における出現回数であり，X は総ページ数，x_j は単語 t_j が出現するカテゴリ数である。

コサイン尺度は，以下の式で求める。

$$S_{ik} = \frac{\sum_{j=1}^{M} w_{ij} w_{kj}}{\sqrt{\sum_{j=1}^{M} w_{ij}^2} \sqrt{\sum_{j=1}^{M} w_{kj}^2}}$$

3.2 リンク関係を重視した判定方法

この方法では Web ページの階層的リンク構造と入口ページ（ページ群を収束する役割を持つページ）の認識によりページ群を判定する。入口ページ識別のための条件は Yahoo!Japan「今週のおすすめ」99 年 8 月 16 日号～9 月 6 日号掲載ページの調査と先行研究を参考に決定した。手順は以下の通りである。

1) 調査対象 Web ページの URL を指定
2) 始点となる入口ページの識別
   a.パスのみで標準で出力されるページ→入口ページ

b.ファイル名が[index,home,main の前方一致]かつ,
拡張子が[htm,html,shtml,stml,stm,sml]である
ページ
→入口ページ
（上記a,b に該当するページがない場合）
c.内部アウトリンク数最大のページ→入口ページ

3) 始点入口ページからのリンク構造を分析

4) 始点入口ページから再起的にリンクを辿れるページ
a.アウトリンクなし→6)
b.外部アウトリンクのみ→6)
c.リンク階層上位への内部アウトリンクのみ→6)
（上記の条件以外→5） ）

5)リンク階層下位への内部アウトリンク先ファイル群
a.他の単一ディレクトリに存在→入口ページ
b.ファイル名（拡張子以外）の先頭あるいは末尾３バイトが
一致→入口ページ
c.タイトルの先頭あるいは末尾３バイトが一致→入口ページ
d.ファイル名（拡張子以外）の数字以外の部分が２バイト以下
で一致，かつ数字部分が連番ー＞入り口
e.タイトルの数字以外の部分が２バイト以下で一致，かつ数字
部分が連番→入口ページ
（上記の条件以外→6） ）

6) 外部アウトリンクが５以上存在→入口ページ

7) 入口ページから辿れるファイル群をページ群と識別

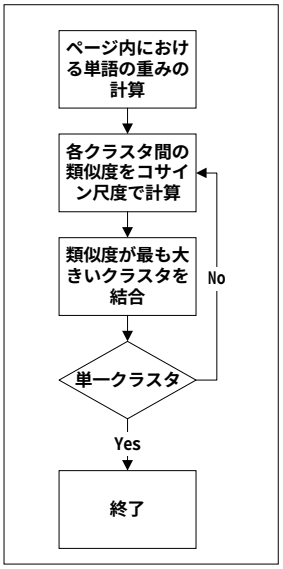

図2 内容類似度を重視した判定手順

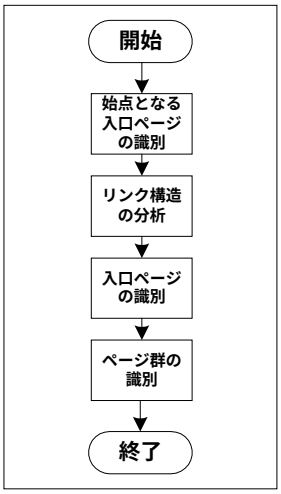

図３ リンク関係を重視した判定手順

4 ページ群の判定実験

提案した二つの判定方法を適用し，Web ページからページ群を判定する実験を行った。

実験用 Web ページとしては，行政，企業，個人といった異なる作成者からなる 4 種の Web ページを用いた。また，ページ群は内容的なまとまりの度合（主題範囲）により，小さな主題から大きな主題までの様々なページ群を想定することができる。そのため本研究では，ページ群を階層的に作成した。それぞれの判定方法により作成されたページ群は，あらかじめ実験用 Web ページを用い人手により判定されたページ群（評価用ページ群）と比較をすることにより評価を行った。以下に評価例を示す。

[例1]PIONEER R&D 技術解説（http://www.pioneer.co.jp/crdl/tech/dvd/）

このページは，技術解説のページであり，目次のページと MPEG のディレクトリ，DVD のディレクトリに分かれている。そのため評価用ページ群は，DVD と MPEG に関して書かれたものに大きく二分した。

内容からみた判定結果は，単語のみのファイルで判定した結果の 9 割以上のページが評価用ページ群と一致し，タグ付きのファイルは約半分が一致していた。単語だけの場合，別のページ群として判断さ

れたページは，目次のページと DVD のフォーマットに関して説明しているページであった。

リンク構造を重視した判定結果は，「5-c. タイトルの先頭あるいは末尾３バイトが一致→入口ページ」という条件一致を中心に，ほぼ正確な DVD，MPEG ページ群の分離から各詳細解説の入口ページまでが判定できた。

[例 2]茨城県情報政策課広報紙 infopia（http://www.pref.ibaraki.jp/infopia）

このページは，茨城県情報政策課の広報紙「インフォピア」に関するページで，第 7 号から第 12 号まで各号ごとにディレクトリが分かれている。各号はそれぞれ，特集，情報化レポート，New System，ブレイクタイム，Information などの項目から成る。評価用ページ群は，号番号を問わず同種項目ページの集合から成るとも考えられたが，逐次刊行物としての要素を重視し，各号，各項目とした。

内容からみた判定結果は，ページ群が特集や情報化レポートなど項目ごとのページ群になったため，評価用ページ群とは一致しなかった。ただし，先にもの述べたように評価用ページ群をそれぞれの項目ごととした場合は，ある程度の一致がみられる。また，タグ付きで判定した場合と単語のみで判定した場合とを比較すると，後者の方が項目ごとにまとまっている傾向がみられた。

リンク構造を重視した判定結果は，5-c. の条件一致により各号，5-b. の条件一致により同一号内同項目のページ群が判定できた。しかし，インフォピア 7 号と 8 号は他のいずれの条件にも当てはまらず，それぞれ同一号とは判定されず「インフォピア」としての大ページ群に属するものとなった。

結論として，内容からみたページ群判定結果は，ディレクトリ構造やリンク関係に関係なく，内容ごとにページ群を判定できることがわかったが，ページ群の内容だけからページ群を判定してしまう問題点も明らかになった。また，タグ付きファイルと単語のみのファイルによるページ群判定の実験結果から，全体的な傾向として単語のみの方がタグを含めた場合よりも好結果となる。その原因として，単語が内容類似度に寄与する割合が大きいこと，内容に関係なく表やリストなど同じスタイルを持つものが類似していると判断されてしまうことなどが考えられる。一方，リンク構造を重視した判定結果は，Web 上での物理的最小単位としてのページ（ファイル）のグループ化を意識して構築された Web ページに対して有効であると考えられる。またこの判定結果から，Web ページのディレクトリ構造をそれほど考慮しなくてもリンク構造から論理的単位としてのページ群が判定できるという可能性が確認できた。

5 おわりに

本研究では 2 種類の観点からなる判定方法により，Web ページからのページ群自動判定を試みた。実験の結果から Web ページには内容ごとに構造化されているページもあれば，そうでないページもあるなど様々であるため，それぞれの方法が適用可能な Web ページのタイプあることがわかった。今後は 2 種類の手法の使い分け，あるいは組み合わせによって，内容と構造の両側面に対応したページ群の判定が必要であるといえる。

【引用文献】

1)永藤拓宏; 遠山元道. ページ群への分割を利用した WWW 検索エンジン. 第 9 回データ工学ワークショップ（DEWS'98）,(1998.3.5-7)

2)原田昌紀ほか. WWW ページ間の階層構造の推定と検索システムへの応用. 情報処理学会研究報告情報学基礎 54-14, p.105-112(1999)

3)岸田和明. 情報検索の理論と技術. 東京, 勁草書房, 1998. 314p.