O-006

アウトライン情報に基づくレポート作成支援システム（２）

― 段落情報の抽出と整理 ―

The Report Creation Support System Based on Outline Information

- Extraction and Arrangement of Paragraph Information -

酒井 章嘉† 四津 匡康† 遠藤 裕英‡
Toshihiro Sakai Tadayasu Yotsu Endo Hirohide

１． はじめに

レポートを作成する手順の一つに、アウトラインを作り、それに沿って情報収集を行い、まとめるという方法がある．情報収集では、検索エンジンを使って Web ページを検索し、Web ページを調べて、その結果を人手でまとめるというのが一般的である。

本研究では、人手に代えてコンピュータでまとめる方法を提案しており、本報告では、WWW(World Wide Web)から収集された Web ページから、必要な情報を抽出し、まとめる処理を行う部分について報告する．

２．システムの提案

本システムは大きく分けて四つの部分に分かれる．

１）検索システム：入力されたアウトライン情報をもとに WWW 上から Web ページを検索する．ここで、アウトライン情報とは、レポートの題名、章のタイトルやキーワード、節のタイトルやキーワードなどである．

２）ブロック分割： Web ページを内容に応じたブロックに分割する．

３）抽出システム：アウトライン情報に該当するブロックを抽出

４）整理システム：抽出したブロックを整理する部分

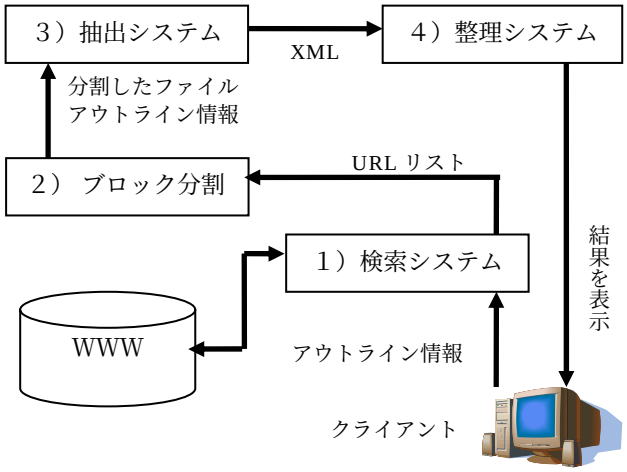

図１ システム全体のイメージ

† 立命館大学大学院
‡ 立命館大学

3. システムの処理

本報告で取り上げる部分は抽出システムと、整理システムに分かれる．以下にそれぞれの処理について説明する．

3.1 抽出システム

Web ページから切り出されたブロックから、アウトライン情報に該当するブロックを選択するシステムである．抽出システムのイメージを図２に示す．以下にシステムの処理の流れを示す．

１）ブロック化された記事ファイルを受け取る
２）アウトラインのキーワードの類似語のリストアップ
３）マッチング処理
４）アウトライン情報に対応するブロックとして選択されたブロックをファイルに出力

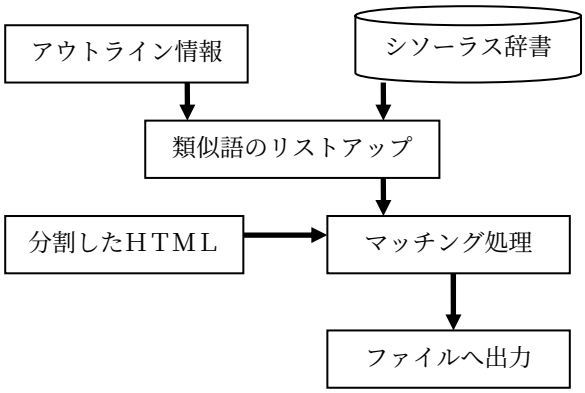

図２ 抽出システムの処理のイメージ

3.1.1 入力情報

抽出処理では、アウトライン情報と、いくつかに分割されたブロック記事ファイルを受け取る．

3.1.2 類似語のリストアップ

マッチング処理に移る前準備として、アウトラインに入力されたキーワードの類似語を、シソーラス辞書を用いてリストアップしておく．[1]

3.1.3 マッチング処理

マッチング処理とは、ブロック記事ファイルの中から、アウトラインに該当するブロックを抽出する処理である．受け取るブロック記事の形式は、記事にタイトルがあるかどうかで大きく２つに分かれ、タイトルがある記事の中でも、サブタイトルがあるかどうかで分かれる．抽出するかどうかは、基本的にブロックのタイトルとアウトラインを見比べて判断する．ブロックのタイトルにアウトラインの

キーワードが含まれていない場合は、ブロック記事を形態素解析し、名詞の出現頻度を数え、最も多い名詞がキーワードなら抽出する．また、ブロックにタイトルがついていない場合は、ブロック記事の中から重要文を抽出し[2]、その文章の中に、アウトラインのキーワードが入っているかどうかで判断する．

3.1.4　ファイルへ出力

次の要約システムへ、出力する内容は、抽出した文章、アウトラインにあてはまる章番号、一致したキーワード、形態素解析を行った場合は、解析結果の情報である．これらを XML 形式で出力する．

3.2　段落整理システム

抽出システムによりアウトラインに対応する複数のブロックを、ユーザが把握しやすいように類似したブロックをグループ化する.そして各グループの特徴を表すキーワードを抽出する．

次にグループ化したブロックの数が１つの場合、ブロックの内容を最も表していると考えられる文を要約文として出力する．

ブロックの数が複数の場合、ブロック間において重複した文が存在することがあるため、グループの中の複数のブロックを要約し１つの文として出力する．

以上の処理を行うことにより、レポートを作成する際にユーザが知らなかった情報を容易に判断することができ、必要な情報の取捨選択の支援を行うことができる．

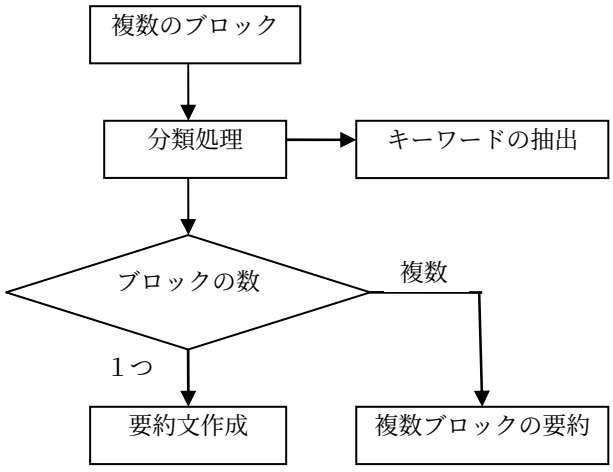

図３　ブロック整理システムの処理のイメージ

3.2.1　分類処理

クラスタリング手法を用い分類し、そして各クラスタのキーワードを tf*idf を用いることにより抽出する．

クラスタリングだけでは分類されたグループ間の関係が判断しにくいため、対象分野の概念関係を記述した Ontology のノードとクラスタのキーワードで対応するものがあれば Ontology のカテゴリにクラスタを当てはめる．

3.2.2　要約文作成処理

単一ブロックにおいて、キーワードだけを表示するのでは内容の把握に不十分であるため、キーワードを多く含む文を抽出し表示するのが望ましい．

アウトラインに対応する複数のブロックに頻出する語、段落に頻出する語に重みをつけて、最も重みのある文を要約文として出力する．[3]

3.2.3　複数ブロック要約処理

形態素解析し、自立語を抽出する.ヒューリスティックスにより比較するブロックの形態素の一致度が 60%を超える場合は共通個所として同定する.[4]
相違個所においては共通個所に係る文を抜き出す方法を考えている．

5. 実験結果・考察

検索システムにより収集された１４ページから、ブロック記事を１０４個受け取り、提案したシステムの処理に基づいて、（A）抽出に成功、（B）抽出に失敗、（C）削除に成功、（D）削除に失敗、の４通りに分け、処理の精度を調べた．

表１：実験結果

（A）抽出に成功	49
（B）抽出に失敗	15
（C）削除に成功	28
（D）削除に失敗	13

この結果から、成功した数は７７個であり、率にして７４%であった．

タイトルがないブロックや、ブロックが大きい場合には、ブロックごとに抽出を行うと不要な部分も抽出してしまい、削除に失敗するというケースが存在した．

また、抽出に失敗するケースとしては、ひとつのブロックの中で、アウトライン上の複数の内容について書かれている場合があった．

6. おわりに

本研究では、レポート作成支援を目的とし、情報を収集し、整理を行う手法を提案した．アウトラインに対応するブロック記事の抽出では、Web ページをブロックという意味的なまとまりに分け、アウトラインに対応したブロックを抽出する方法を提案した．

今後の課題としては、システムを実装し、評価データを増やした評価と、精度の向上が必要である．

参考文献

[1]佐藤慎哉：“web ページ中のテキストと表からの重要情報抽出”,情報処理学会研究報告、03-NL-153-9、2003
[2] 大竹清敬・ 岡本大吾・児玉充・増山繁：“重要文抽出，自由作成要約に対応した新聞記事要約システム YELLOW”，情報処理学会研究報告、2001.7
[3]Inderjeet Mani 著/奥村学・難波英嗣・植田偵子訳“自動要約”共立出版
[4]難波英嗣・奥村学：“ここまで来たテキスト自動要約” IPSJMagazine Vol43 No.12 Dec 2002