社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－FI－78（4）
2005－DD－49（4）
2005／3／25

階層的キーワードを手がかりにしたＷｅｂ情報整理支援システム

日置 里美†　　廣田善洋†　　遠藤 裕英†

†立命館大学大学院理工学研究科

あるトピックについての知見や情報をまとめる時，インターネットを参照する機会が増えている．この時，まとめ方に近い形で情報が提示されている Web ページを見つけることができると便利である．

本研究ではまとめ方を提示する方法として，アウトライン（階層化されたキーワード群）を用いる．ユーザが提示したアウトラインと Web ページから抽出したアウトラインの類似性を求め，ユーザの視点に近い Web ページを抽出する方法を提案する．次に Web ページのアウトラインから得た情報をもとに新しいアウトライン案(まとめ方)をユーザに提示する手法を提案する．

Reorganizing Web Information along Hierarchical Keywords

Satomi Hioki ††　　Yoshihiro Hirota ††　　Hirohide Endo ††

††Graduate School of Science and Engineering,Ritsumeikan University

Recently, a lot of people have begun to consult the Internet when editing documents on a variety of topics.　Here we propose an outline similarity between the given topic and web pages in order to select useful the web pages for the desired topic.

A view of the topic is given by the table of contents which is regarded as a hierarchical structure of keywords. We call this structure the outline. In addition, the outlines of web pages are collected by finding title and text pairs of each web page's content. The usefulness of a web page towards the given topic is measured by the similarity between the topic outline and web page's outline. Also, by merging the collected outlines, a new and improved outline is proposed.

1. はじめに

日常，現時点の知見をまとめる必要性に遭遇する．そのような場合，情報収集の手段として WWW（World Wide Web）上から関連する情報を集めて草稿を練る機会が増えている．

WWW から情報を収集する手段として検索エンジンを用いることが一般的だが，検索エンジンによる検索結果には不必要な情報も多く，膨大な検索結果の中から必要な情報を見つけ出す作業に多くの時間を費やすことも少なくない．しかも，その情報には利用者が必要とする意図でまとめられているケースは多くない．我々は，いろいろなサイトで開示されている情報を，目的に沿った視点から見直して利用している．

本研究では利用者がまとめたい視点を目次のような階層構造を持つ複数のキーワード群で表現することにする．本研究ではこれをアウトラインと呼ぶ．そこで，与えられたアウトラインに適合する Web ページをどのように見付け出すかということが課題になる．本研究では Web ページを解析し，そこに含まれるキーワードから当該 Web ページのアウトラインを抽出し，ユーザの提示する階層キーワードとの類似性を調べる方法をとる．

しかし，最初にユーザが提示するアウトラインと一致する Web ページが必ず存在するとは限らない．また，ユーザが提示するアウトラインも完璧なものとは限らない．そこで，与えられたアウトラインに類似する Web ページを収集し，個々の Web ページが持つアウトライン情報を取得する．取得したアウトライン情報とユーザが与えたアウトラインを基に階層構造を再

構成しユーザに提示する手法を提案する.

2. 階層構造キーワードを用いた情報収集の問題点

2.1 重要な Web ページを選択する際の問題点

検索エンジンを使うと,不要なWebページが検索結果の上位に出てくることがある.インターネット利用者に向けた調査[1]によると,約9割のユーザが検索結果について「多すぎる」と感じている.また約6割のユーザがなかなか欲しい情報にたどり着けない,情報を得るのに時間がかかる,検索エンジンを使いこなせていないと感じている.検索エンジン各社が検索件数を競うあまり,質よりも量を重視してしまっていることにも一因がある.

Web ページの重要度を評価付けるシステムとしてGoogle が開発した PageRank[2]がある.PageRank は評価の高いページからリンクされているページは評価が高くするというリンク構造に基づいて評価付けを行っている.しかし,この方法では Web ページ自体の内容を評価の基準としていないため,利用者が探している情報が評価の高いページで見付かるとは限らない.

結局,ユーザは検索エンジンから収集した膨大なアドレスから重要だと思われるページにアクセスし,その内容が適切かどうか実際の目で見て判断しなければならない.ある主題についてまとめる上で不要だと思われるページの例として,見出しは存在するが参考になる本文が無い授業や書籍の紹介ページ,個人的主観で記載された日記やブログ,掲示板といったページがある.これらは Web 上に大量に存在し,検索エンジンでも上位にヒットするのでなかなか欲しい情報にたどり着くことができない.

2.2 階層構造キーワードの情報を収集する際の問題点

目次のように階層構造を持つキーワード群を検索エンジンで検索し,その中から重要な情報を収集・整理する際には通常以下のような手順で行う.

① 階層構造を層ごとに分割する
② 分割された層のキーワードごとに検索エンジンで検索し,重要だと思われる Web ページを収集する
③ 収集した Web ページの中から重要箇所を抽出する
④ 抽出した重要箇所を最初に設定した階層構造の視点に沿って並べる.
⑤ 重要な内容に対応する階層構造キーワードが見つかった場合,新たにキーワードを追加し階層を再構成する.

検索エンジンは AND や OR で複数のキーワードを与えることができるが,階層構造を持つキーワードを階層構造を維持した状態で与えることはできない.情報検索をする際には,まず,階層構造を分割し,ひとつひとつの見出しについて検索エンジンで分割された部分のみの検索を行う.この結果から得られた Web ページはユーザが求める目次の一部でしかない.したがって階層構造全体の情報を得るには,分割し検索するという作業を全ての層で繰り返さなければならない.検索結果として得られる情報はキーワードに対するものなので,ユーザの視点でまとめるには,集めた情報を最初の階層構造に沿って組み合わせなければならない.

Web ページから収集した情報はテーマは同じでも,記入されている内容はひとつずつ異なる.そこで重要な情報をもう一度,ユーザの求める視点すなわち最初に設定した階層構造に補完する形で整理し直し,最初に設定した階層構造を再構成する必要が出てくる.

3. 階層構造キーワードによる Web ページ取得方法

Web ページは検索エンジンから取得する.まず,検索エンジンのキーワードの与え方と検索対象とするページについて述べ,次に取得した Web ページの選択方法について述べる.

3.1 検索エンジンのキーワードの与え方

ユーザが与える階層構造キーワードを図1のようにキーワードのツリーで表現する.このような階層構造キーワードがアウトラインである.一番上の階層をタイトルとし,その下の階層のキーワード群を見出しと定義する.

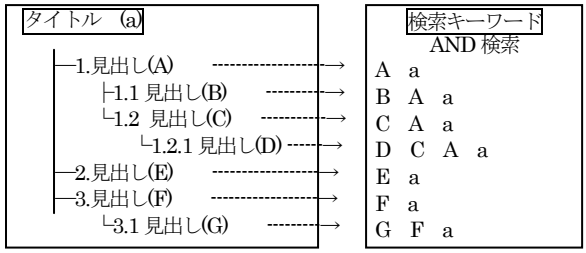

図1 階層キーワード(アウトライン)と検索キーワード

Web ページの検索は見出しごとに行う.図1のように検索エンジンのキーワードに検索対象の見出しとその上の層の見出し全てとタイトルを与え,AND 検索を行い上位 20 件の Web ページを取得する.また,資料を集める際に参考にならないと思われる「日記」や「掲示板」,「ブログ」のような Web ページも事前に省きたい.そこで,このようなページで多く使用されている cgi などのサーバサイドプログラムを使用した

ページは検索の対象としないことにして，Web ページのファイルの拡張子を.html と.htm の２つに限定する．

3.2 Web ページの階層化

Web ページの構成方法にはいくつかのタイプがあるが，本研究では知見の整理を目的としているので，主としてドキュメント主体の Web ページを対象とする．図２にドキュメント主体の Web ページの例とアウトラインを抽出し，ツリー構造にしたものを示す．なお，Web ページのタイトルは一番上の見出しとみなす．

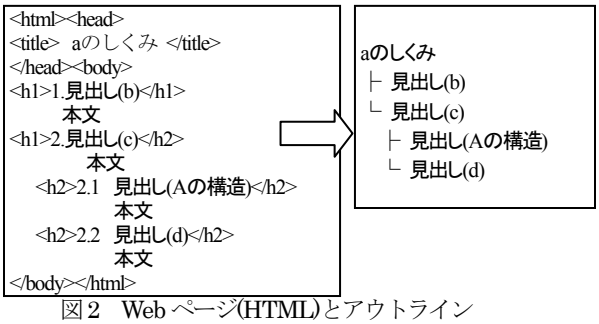

図２ Web ページ(HTML)とアウトライン

ページの構成としては，ページのタイトル(a のしくみ)があり，それに続いて概要的な記述がある．次に，章的な見出し(b)(c)があり，その説明文として本文がある．章の下に中項目の見出し(A の構造)(d)と説明がある．このように Web ページ自体も階層構造になっており，Web ページ中の文章を階層表現する際に読み手にとって最も明解な方法は，見出しを付けることである．本研究ではこの Web ページの階層構造を用いてユーザが与えるアウトラインとの類似度を検証する．そのため，HTML タグからタイトルと見出し，見出しに対する本文を抽出し，Web ページの階層構造を特定する必要がある．

Web ページを階層化する手順としてまず 1)Web ページ全体を段落に区切り，2)段落から見出しと本文を抽出する，最後に 3)階層構造を構築する．以下，1)～3)の処理について述べる．

1) 段落の抽出

文章の区切りとなるタグ…<P>，
，<BROCKQUOTE>，<HR>，<TABLE>，<TD>，<UL>，<OL>が発見できた場合，まず文章の有無に関わらず出現位置を記憶しておく．抽出候補の部分が文章であるかどうかを判定するために，句点に注目する．句点は章を記述する上で確実に１つ以上存在する文字であり，句点があるならば文章と判断してほぼ間違いない．文章と判断した場合，文章の区切りとなるタグで囲まれた部分を１段落とする．

2) 見出しと本文の抽出

まず，段落の中から見出しと本文を抽出する．見出しの抽出は，①見出しを記述するためのタグとして用意されている<Hn>(n=1, 2, 3, …)に囲まれた箇所，②<強調タグ>見出し語句<／強調タグ><改行タグ>のように強調タグと改行タグを組み合わせた箇所の２点から抽出する．強調タグには，<B>，<I>,<U>,<FONT>,<STRONG>タグがあり，改行タグは，
，<P>，<TD>，<TABLE>タグがある．

次に段落の中で見出しと判断された箇所以外を本文として抽出する．

3) 階層構造化

<Hn>や<FONT>タグで文字サイズを指定している場合はそれを記録し，後の階層構造決定に反映させる．また，文章の区切りとなるタグ<BROCKQUOTE>，<UL>，<OL>タグが現れた場合，階層を１つ下げる．水平線を表す<HR>タグで文章が分けられていた場合は，話題の転換箇所と判断し，次の記述からは最上位の階層とする．階層構造化された Web ページはツリー構造を表現するのに最適な XML 形式で保存する．

3.2.1 階層化 Web ページの整形

Web ページから抽出した階層化データを利用しやすい形に整形する．整形する点は以下の２点である．

> ①見出しの無い本文に見出しを付ける．
> ②見出しに余計な文字，記号が含まれている場合は削除する．

①本研究では Web ページの見出しを抽出してアウトラインの再構築を行うため，見出しの付いていない本文にも仮の見出しを付ける必要がある．本文の内容を完全に把握し，適切な見出しを付ける作業は単純な作業ではないため，今回は最低限の内容を把握することに絞り，本文中の頻出単語を仮の見出しとすることにした．頻出単語を求めるため，本文を形態素解析し，名詞だけを抜き出す．同じ名詞の出現回数を数え，最も多い名詞を仮の見出しにする．最大出現回数が同じ名詞が複数存在する場合は，最初に出現した名詞４個までを羅列して仮の見出しにする．形態素解析エンジンには chasen-2.3.3[注]を使用した．

②Web ページの見出しには，見出し番号や見出し記号が含まれている場合が多く，そのままではアウトラインの見出しと比較する際に一致しているかどうか判定が困難になるため，見出し番号や記号を事前に削除しておく必要がある．本システムで削除の対象とした見出し番号，記号は表１のとおりである．他にも削除した方が望ましい記号が見つかった場合は随時，削除リストに追加可能である．

表１　削除リスト

	削除リスト
見出し番号	半角数字，全角数字，漢数字， ①②③④⑤⑥⑦⑧⑨⑩，ⅠⅡⅢⅣ
見出し記号	．　．，　＊　■　◆　□　◇　・　［　］　＿　♪ ●　○　【　】　（　）　（　）　「　」　『　』　《　》 ＜　＞　：　：　☆　★　※　▼　▽
見出しを表す文字	第　章　項　図

3.3　Web ページの選択

検索エンジンから取得した Web ページには，検索キーワードは含んでいても書籍や授業の紹介ページのように，見出しは存在してもその本文が存在しない Web ページなど資料を集める際に参考になりになりにくいページも多数存在する．このような Web ページを省き，利用者の視点に近い Web ページを見つけ出すため，検索エンジンから取得した Web ページ全てにおいて見出しとその本文に注目しアウトライン一致度を調べる．

アウトライン一致度を調べる手順について説明する．ユーザの入力する階層構造の見出しを分解し，まず一番初めの見出しを基準見出しにする．そして階層化した Web ページについて

① 基準見出しが Web ページ中の見出しとして存在するか
② 基準見出しがあれば，その本文が存在するか
③ 基準見出しの上の層の見出しが①の見出しの上の層に存在するか

を調べ，同様に次の見出しを基準見出しに設定し，調べる．これを全ての見出しについて繰り返す．そして取得した全ての Web ページにおいてアウトライン一致度を調べる．

アウトライン一致度はポイント付けにより数値化する．①の場合一致する見出しがあれば＋2，基準見出しを含む見出しであれば＋1 ポイントを付ける．見出しは無いが基準見出しが本文中に存在すれば 1 ポイントを付ける．②で本文が存在すれば＋2 ポイントを付ける．③で上の層と一致する見出しが存在すれば＋2 ポイント，上の層を含む見出しであれば＋1 ポイントを付ける．1 つの見出しにつき最大 6 ポイントなので Web ページの判定ポイントを次の数式により求める．

$$\text{判定ポイント} = \frac{\text{合計ポイント}}{\text{見出しの数}\times 6}\times 100$$

全ての Web ページで判定ポイントを求め，ポイントの平均点以上の Web ページのみ選択対象にする．

4. Web 情報の整理とアウトラインの再構成

Web 情報を整理し，ユーザに提示する手法として最初に入力したアウトラインに Web から収集した情報を補完し結合させて，アウトラインを再構成する手法を提案する．

Web ページの製作者の視点に基づいて個々に作成されたページとユーザの求める視点が必ずしも一致するとは限らない．そこで，ユーザの求める視点を分解し，分解した個々のキーワードをともに Web ページから重要情報を収集しそれらを結合する手法を用いる．実際には，Web ページから抽出した重要な見出しを，最初のアウトラインの階層構造を核にして言語的な上下関係を用いて結合させ，アウトラインを再構成させる．

次にアウトラインの再構成の手順について述べる．

①Web ページの見出しからアウトラインの基準見出しとその上の層の見出しを探索する．
②Web ページから抽出した見出しに言語的な上下関係を利用し，最初のアウトラインの見出しを補いつつ，Web ページから抽出した見出しをツリーに追加する．
③複数の Web ページから抽出・作成した，抽出見出しツリーをひとつに結合し，見出し結合ツリーを作成する．
④アウトラインの見出しごとに作成した，見出し結合ツリーをひとつに統合する．

図 1 のアウトラインと図 2 の Web ページを例にして手順を説明する．

① まず，アウトラインの一番初めの見出し A を基準見出しとし，A と一致する見出しが Web ページにあるか探索する．もし Web ページにアウトラインの見出し A と一致する見出しが無ければ，A を含む見出しを探索し，抽出する．次に，見出し A の上の層のタイトル a と一致する見出しが，抽出した Web ページの見出しＡまたはＡを含む見出しの上の層にあるか探索し，あればＡと a の間に存在する見出しも含めて抽出する．

② ①で抽出した Web ページの見出しをツリー構造にする．Web ページからは A を含む見出し「A の構造」とひとつ上の層の「c」，そして c の上の「a のしくみ」の 3 つの見出しを抽出した．ここで，3 つの見出しの間にアウトラインの見出し A と a を挿入する．まず A を含む見出し「A の構造」の上の層に A を追加する．次に a を含む見出し「a のしくみ」の上の層に a を追加する．アウトラインの見出しを含む見出しの

上の層にアウトラインの見出しを追加する理由は，異なる Web ページの見出しをアウトラインの見出しを基準に整理するためである．もし，アウトラインの見出しを追加しなければ，異なる Web ページから抽出したアウトラインを含む見出しをまとめる上の層が存在しないことになる．それでは，見出しツリーを結合することができず，見出しツリーがバラバラになってしまう可能性がある．アウトラインの見出しを，アウトラインの見出しを含む見出しの上に追加すれば，言語的な上下関係を保つことができるので，これで仮の上下関係を構築する．図 3 に例を示す．

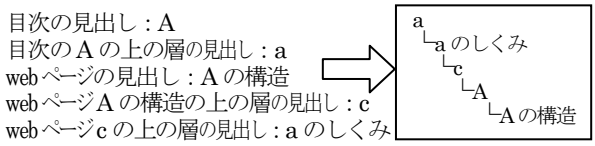

図 3　アウトラインの見出しを挿入した
Web ページの抽出見出しツリー

③　さらに他の Web ページでも同様にアウトラインの基準見出し A と，ひとつ上の層の見出しタイトル a を基準に探索を繰り返し，Web ページごとに上下関係の見出しツリーを作成する．Web ページごとに異なった見出しツリーが作成されるので，この上下関係のみの抽出見出しツリーをひとつのツリーに結合し，基準見出し A の結合ツリーを作成する．異なるツリーを上の層から順に同じ階層に同じ見出しが存在するか比べ，もし同じ深さの階層に異なる見出しが現れた場合，異なる見出しから枝分かれさせる．

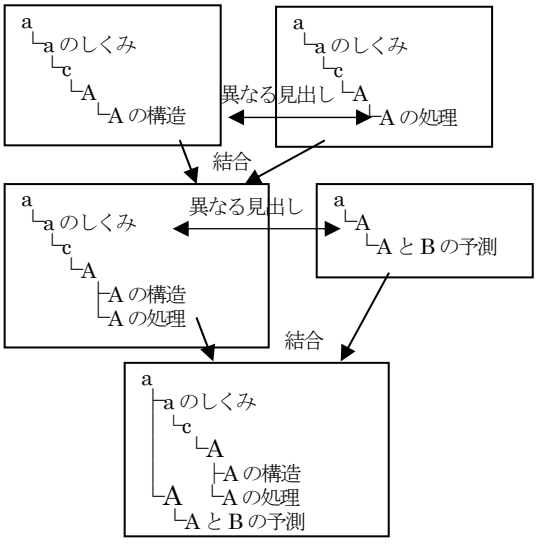

図 4　A の結合ツリー

④　その他のアウトラインの見出し B から G についても同様に見出しツリーを作成したものをひとつに統合する．上の層から順に同じ階層に同じ見出しが存在するか比べ，もし異なる見出しが現れた場合，異なる見出しから枝分かれさせる．図 5 に A の結合ツリーと B の結合ツリーを統合した結果，図 6 に全ての見出し結合ツリーをひとつに統合した結果を示す．

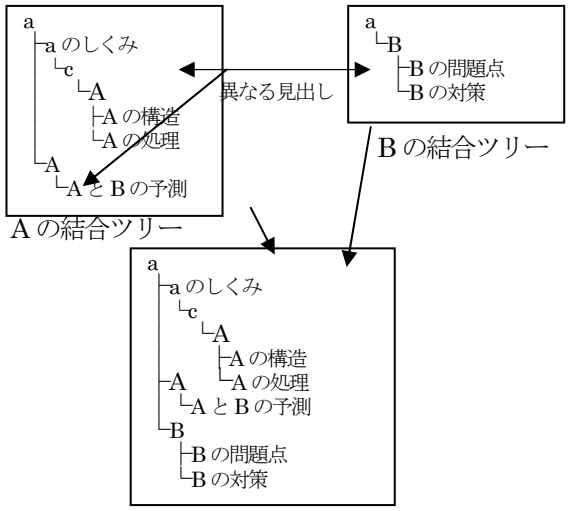

図 5　A と B の結合ツリーの統合

アウトラインの見出し C から G のツリーについても同様に統合すると図 6 の統合ツリーのようになる．

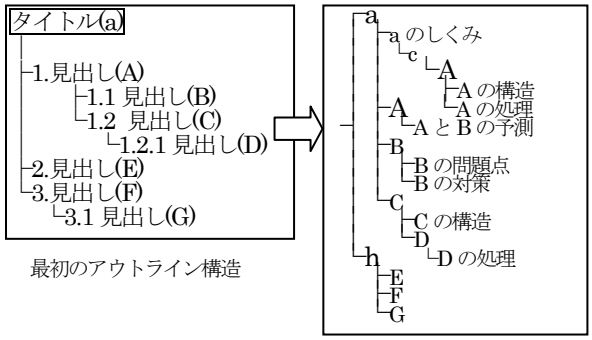

図 6　統合ツリー(再構成したアウトライン)

図 6 の統合ツリーのように，Web ページから抽出した新しい見出しで最初に設定したアウトラインの間を補い，階層構造を変更すると再構成したアウトラインが完成する．

5. システムの実装

本研究のシステムでは階層構造を持つアウトラインを入力し，WWW から Web データを抽出した後，Web ページを階層化する．階層化した Web ページは XML 形式で保存され，ポイントの高い Web ページから見出しを抽出する．抽出した見出しをアウトラインの見出しを軸に並び替え，アウトラインを再構成しユーザに表示する．再構成されたアウトラインは拡張，削除ができるようになっている．図 7 にシステム全体の構

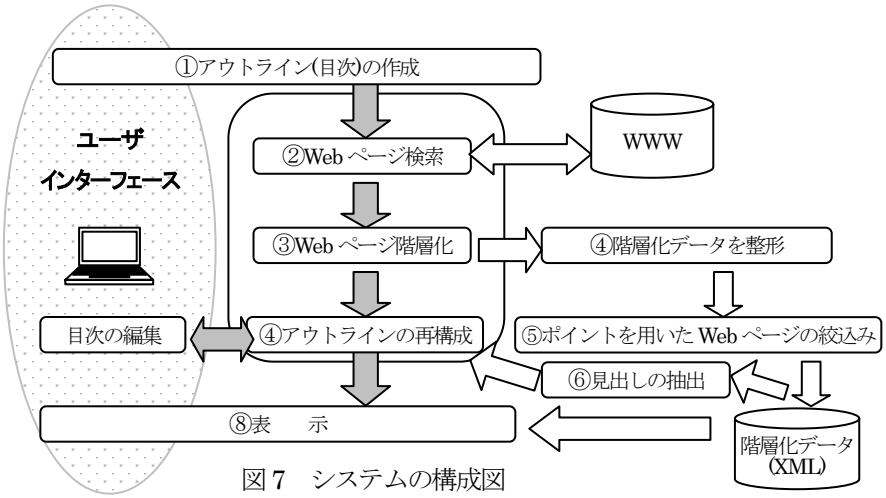

図 7　システムの構成図

成図を示す．ユーザインターフェースは Java サーブレットによって作成し，ブラウザで表示する．アウトラインの入力画面を図 8 に，再構成したアウトラインの表示画面を図 9 に示す．

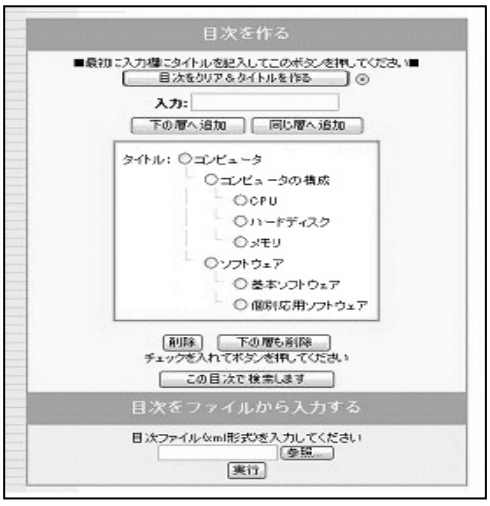

図 8　アウトラインの入力画面

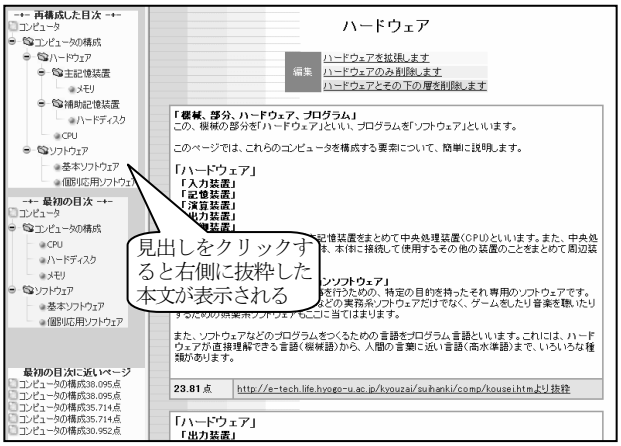

図 9　再構成したアウトラインの表示画面

6. 評価と考察

6.1　再構成したアウトラインの妥当性の調査結果

セキュリティ，生命科学，技術経営，国民年金のタイトルを持つ 4 つのアウトライン(合計 22 個の見出し)において再構成したアウトライン(合計 100 個の見出し)の妥当性の評価結果を表 2 に示す．

上下関係の妥当性，並列関係の妥当性については，意味的な判断と言語的な判断の 2 種類を検証した．上下関係の計算方法は次の式を用い，言語的な上下関係は上の層の見出しを下の層の見出しが含んでいるかどうかで判定する．

$$\text{上下関係の妥当性} = \frac{\text{正しい上下関係の対の数}}{\text{上下関係として表示された見出しの対の数}}$$

並列関係の計算方法は次の式を用いる．

$$\text{上下関係の妥当性} = \frac{\text{正しい並列関係になっている見出しの数}}{\text{並列関係として表示された箇所の見出しの数}}$$

言語的な並列関係は並列関係に表示されている見出しが前方一致または後方一致の関係にあるかどうかで判定する．正しい並列関係になっている見出しの数とは，並列関係として表示されている箇所の見出しの数から孤立した関連のない見出しを引いた数である．並列関係として表示されている箇所の中に並列関係となる組が複数存在する場合は，一番数の多い組を基準とする．

見出しの妥当性は全ての見出しの中の参考になる見出しの数を求めた割合である．

表 2：再構成したアウトラインの評価

タイトル	上下関係の妥当性		並列関係の妥当性		見出しの妥当性
	意味的	言語的	意味的	言語的	
セキュリティ	94.1%	76.5%	62.5%	18.8%	88.9%
生命科学	80.5%	69.4%	50.0%	50.0%	78.9%
技術経営	94.4%	88.9%	44.4%	33.3%	63.6%
国民年金	75.0%	70.0%	55.6%	66.7%	82.6%
平　均	86.0%	76.2%	53.1%	42.2%	78.5%

6.2　再構成したアウトラインの妥当性の考察

1)　上下関係の妥当性の考察

Web ページから抽出した上下関係を利用しているため，意味的に調査した結果，最高 94.4%，平均でも 86%とほぼ良好な結果が得られた．言語的に判断した場合，意味的な判定より若干結果が悪くなったのは最

初に設定したアウトラインが言語的な上下関係になっていなかったためである．

2) 並列関係の妥当性の考察

並列関係の妥当性の評価は平均で 53%とあまりよい結果が得られなかった．これは，異なる Web ページから収集した見出しを，上の層の見出しが同じであるということで，無理やり同じ層に表示してしまったためである．上の層が同じでも，異なる Web ページから収集した見出しが並列関係にあるとは限らないため，アウトラインをひとつにまとめる上で，よい結果が得られなかった．

しかし，言語的に異なる Web ページから収集した見出しで，前方一致または後方一致の関係にある見出しは約半数存在した．異なる Web ページから抽出した見出しを前方一致または後方一致の関係でまとめ，さらに整理できるような処理を付け加えることができればもう少し評価がよくなると予想される．

3) 見出しの妥当性の考察

見出しの妥当性は平均で78.5%と概ね良好な結果が得られたが，参考にはならない見出しが若干含まれていた．これはポイント付けにより高得点が付けられたページの見出しである．このような見出しを機械的に100%省くことは難しいので，最終的な判断はユーザが行う必要がある．

6.3 選択した Web ページの妥当性の調査結果

技術経営，生命科学，セキュリティの 3 タイトルのアウトラインから収集した Web ページ 300 ページを資料を集める際に参考になるページと参考になりにくいページの２つに手作業で分類した．参考になるページとは，調べたい分野について詳しい説明のあるページ，論文形式のページなどである．参考になりにくいページとは，書籍，授業，講演内容の紹介ページ，商用ページ，日記，掲示板，リンク集である．人が判断する参考になるページとポイント付けで判断した有効ページ(判定基準は平均点)とどれだけ一致するか調べる(正答率)．さらに一致しなかったページについては①ポイントが高いが重要ではないと判断したページ，②ポイントが低いが重要だと判断したページの 2 点について検証した．本システムでは平均ポイントで分けたが，平均ポイントで分ける方法が妥当かどうか検証するため，平均の前後 1 ポイントでも同様に検証した．使用したアウトラインは以下の図 10 の 3 種類であり，15 個の見出しからそれぞれ Web ページを 20 ページ収集し，合計 300 ページを調査した．表 3 に結果を示す．

タイトル：技術経営	タイトル:生命科学	タイトル：セキュリティ
1. イノベーション	1. 遺伝子	1. プライバシーの保護
2. 研究開発戦略	1.1. 遺伝子情報	2. 暗号化
2.1. 研究開発管理	1.2. ヒトゲノム	3. ウィルス
3. 企業統治	1.3. クローン	3.1. ウィルスの種類
4. 技術経営戦略		3.2. 予防策
4.1. 特許戦略		

図 10　選択した Web ページ妥当性検証のために使用したアウトライン

表 3：平均ポイントの前後 1 ポイントで分けた結果

タイトル	判定ポイント		正答率	ポイントが低いのに参考になると判断したページ数	ポイントが高いのに参考にならないと判断したページ数
技術経営 120 ページ	−1	1.98	70.8%	10 ページ	25 ページ
	平均	2.98	76.7%	15 ページ	13 ページ
	+1	3.98	79.2%	15 ページ	10 ページ
生命科学 80 ページ	−1	10.92	73.4%	8 ページ	13 ページ
	平均	11.92	73.4%	8 ページ	13 ページ
	+1	12.92	75.0%	13 ページ	7 ページ
セキュリティ 100 ページ	−1	8.97	65.0%	16 ページ	19 ページ
	平均	9.97	65.0%	16 ページ	19 ページ
	+1	10.97	72.0%	17 ページ	11 ページ

6.4 選択した Web ページの妥当性の考察

今回検証した 300 ページのうち，ポイントの平均点で正しく分けることができたページは216ページであった．平均ポイントの前後 1 ポイントで分けた結果と比較すると，1 ポイント高くすると，参考にならないページがふるい落とされ，正答率は若干上がるが，重要でもポイントが低く抽出できなかったページが増えてしまう．逆に判定ポイントを下げれば，重要でもポイントの低いページは抽出できるが，その分参考にならないページも抽出されてしまう．重要なページをできるだけ抽出したいのでポイントを下げたいが，必要の無いページまで抽出してしまうと不要な見出しが増え，再構成したアウトラインが煩雑になってしまう．そこで平均ポイントで分ける方法が妥当だと考える．

正しく分けることができなかったページは，ポイントが低いのに参考になると判断したページは 39 ページ，ポイントが高いのに参考にならないと判断したページは 45 ページであった．

1) ポイントが低いのに参考になると判断したページについての考察

ポイントが低いのに参考になると判断したページは，Web ページの内容はアウトラインのタイトルと同じ分野だが，文章中にアウトラインの見出しと同じ言葉を使っておらず，別の言葉で言い換えていたり，アウトラインの見出しとは別の視点で論じている場合が 9

ページあり一番多かった．本システムではアウトラインの見出しを基準にして Web ページから見出しを抽出したので，このようなページからは見出しを抽出することができなかった．類語辞書や専門用語辞書をシステムに組み込み，言い換えられた言葉や関連する言葉でも抽出できるようにすれば，新しい論点をユーザに提供できると考えられる．

また，ひとつのキーワードについて詳しく書いているページが 6 ページありポイントが低かった．本システムではアウトライン全体の見出しと一致しているかどうかで判定したので，もし，ひとつの見出しのみ一致しても高いポイントは得ることができなかった．どの程度詳しい内容を記載しているかを判定するため，本文の文字量や本文中で使われているキーワードの出現回数などでも判定できるようにし，詳しい内容が記載されているページはポイントの重みを重くすれば抽出できるようになると思われる．

Web ページ自体に見出しが無く，階層化されていないページも 300 ページ中，3 ページ見つかった．この場合も，本文の文字量やキーワードの出現回数を調べ，内容によってポイントの重み付けを考慮すれば抽出できるようになると思われる．

Web ページを階層化する際に重要な部分が抜け落ちてしまったページ，正しく階層化できなかったページ，階層化した XML データに不具合があったページ 21 ページについては Web ページ階層化システムの精度の向上が必要である．

2) ポイントが低いのに参考になると判断したページについての考察

ポイントが低いのに参考になると判断したページは，詳しく内容を記載した講演会のページ，授業のページが 23 ページあり一番多かった．また，セキュリティやプライバシーの保護といった見出しから抽出したページは企業のプライバシー理念のページが 12 ページ存在した．他にも参考になりにくいページにどのような特徴があるか調べ，たとえポイントが高くなっても参考になりにくいページの特徴が当てはまっていれば事前に省くことができるように改善すればさらによい結果が得られるのではないかと考える．なお，本システムでは検索対象とするページのファイル形式を htm と html に絞ったため個人の日記のサイト，掲示板，アマゾン.com の書籍の紹介ページは概ね省くことができ，もし検索でヒットしたとしても，ポイントが低くふるい落とされた．

また，今回はリンク集自体には資料となる本文が記載されていないということで，参考にならないページに分類した．リンク集が参考になるページかどうかの判断は個人の判断による場合が大きいので，どちらか正確に分類することは難しい．

7. 終わりに

本研究では Web ページから階層構造を持つキーワードを収集し，さらに収集した情報を整理・統合し，ユーザに提供できるシステムを構築した．本研究の特徴は，階層構造のキーワードに適した Web ページを収集し，有効なページを抽出できる点と最初に与えるアウトラインを Web ページから収集した情報をもとに再構成する点の 2 点である．

階層を持ったキーワードで検索したい場合，特にレポート作成時の情報収集に本システムを使用すれば，見出しひとつひとつで検索するよりもはるかに効率的に，しかも重要な情報のみ整理した形で提供できるので非常に便利なツールになると考える．

今後の課題は，アウトラインの再構成と Web ページからの抽出の精度を上げることである．具体的にはアウトラインを再構成する際に，言語的な繋がりで見出しをまとめ，さらに整理した形で提供できるようにしたい．また Web ページからの抽出の際には言い換えられたキーワードでも抽出できるようにし，新しい視点を提供できるようなシステムにしていきたい．

参考文献

[1]インターネットコム(株)による記事：http://japan.internet.com/research/20020201/1.html
[2]Google：http:www.google.com/
[3]茶筌：http://chasen.naist.jp/hiki/ChaSen/
[3]廣田善洋，花房誠一：“アウトライン情報に基づくレポート作成支援システム (1) -情報収集と段落抽出-” FIT(情報科学技術フォーラム)0-005 P479-480 (2003)
[4]酒井章嘉，四津匡康：“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ー” FIT 0-005 P481-482（2003）
[5]藤井敦，伊藤克亘，秋葉友良：”CYCLONE: 最強事典サイトの構築”, 情報処理振興事業協会 2002 年度成果報告集 第二版 http://www.ipa.go.jp/SPC/report/02fy-pro/report/983/paper.pdf
[6]George Chang, Marcus J.Healey, James A.M. Mchugh, Jason T.L.Wang 著 武田善行，梅村恭司，藤井敦 訳：『Web マイニング』，共立出版，2004