階層的要約を用いたWeb文書集合への問合せ

高橋　　功†　　三浦　孝夫†　　塩谷　　勇††

† 法政大学 工学部 電気電子工学科　〒 184-8584 東京都小金井市梶野町 3-7-2
†† 産能大学経営情報学部〒 259-1197 神奈川県伊勢原市上粕屋 1573
E-mail: †{i05r3226,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

あらまし　現在 Web 文書集合への問合せ手法として多くのシステムが問合せの語を含む文章の抜粋を利用する．しかし単純に問合せの語を含む文章の抜粋からでは, どの文書が利用者にとって有用な情報であるかを判断することは困難である．本稿では, 素早く Web 文書の内容を把握するために階層的要約を利用する．そして, 実験により提案手法の有用性を示す.

キーワード　自動要約, Web マイニング,Web クラスタリング, 情報検索

Processing Queries based on Hierarchical Web Summarization

Kou TAKAHASHI†, Takao MIURA†, and Isamu SHIOYA††

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184-8584 Japan
†† Department of Management and Information Science, SANNO University 1573, Kamikasuya,
E-mail: †{i05r3226,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

Abstract　In this investigation we discuss a novel method of query processing based on Web pages summarization while traditional query systems give part of the page contents. Clearly it is hard to grasp what the pages are going to say. In this work we show how useful hierarchical summarizations are to grasp the answers quickly as well as how to evaluate the answers. We show some experimental results using NTCIR Web documents.

Key words　Web Summarization, Web Clustering, Web Mining , Web IR

1. 前　書　き

近年，インターネットの爆発的な普及により World Wide Web(WWW) の世界は急激に拡大し, 世界中の誰もが容易にアクセスできる膨大なテキスト情報をもたらした．このような膨大な Web データ群から利用者にとって有益な情報を見つけるのを手助けするための手法を Web 情報検索と呼ぶ．これまでに Web 情報検索システムとして様々な検索エンジンが提案され,3 億から 30 億と言われる巨大な URL データベースを構築している．この巨大なデータベースを用いた情報検索システムにより問合せと関連する Web 文書の URL を得ることができた．特に Google や Yahoo!等の検索エンジンは利用者が問合せ語を通じて自分の要求を伝えることで関連する Web 文書の URL をランク付けしたリストを検索結果を得ることができる [1], [12]．しかしながらキーワードマッチを用いて問合せ語に適合する Web 文書を検索するとき，問合せ語を直接含まない Web 文書を探し出すことができないという問題点がある．

ランク付けの方法として知られている手法の一つが HITS アルゴリズムである [12]．HITS アルゴリズムは Web ページを *Authority* や *Hub* という二つの視点から取り扱う．あるページから参照されている Web ページを Authority とし, 特定のトピックに関する情報が豊富であることを表す尺度である．一方, あるページを参照している Web ページを Hub とし, authority としての価値が高いページへのリンクが豊富であることを表す尺度である．特に Authority の値に基づいて検索結果をランク付けが利用される．

検索エンジンで利用する場合, ランク付けられた Web 文書の要約として問合せ語の前後の文章の断片を抽出したものを利用する．利用者は得られた検索結果をブラウズし目的に合致する Web 文書を探すのだが, ほとんどのシステムでは問合せ語が含まれている箇所の文章を要約として提示している．そのため, 利用者はこのわずかな情報を頼りに, 類似した Web 文書を多量に含む検索結果の中から求める情報を有する Web 文書を探すことになる．しかしながら, 単純に問合せ語を含む箇所の文章からだけでは，どの Web 文書が有用であるかどうかを判断することは困難である．そのため，Web 文書の内容を素早く把握する方法，即ち Web 文書の自動要約に対するニーズが高まっている．

自動要約は情報源から特定の利用者 (あるいはタスク) にとって最も重要な情報を抜き出すプロセスである [13]．利用者にとっ

て自動要約は文書 (例えばニュース記事など) の内容を素早く把握するために利用できる．これまでに提案されてきた自動要約の手法は 3 つに大別することができる．

抜粋 *(Extracting)* は文書中で最も主題に関連する箇所を識別する手法である．特に, 単語に重みなどのスコア付けを行い，スコアの高い語を含む文章を抽出する方法を重要文抽出法と呼ぶ. 新たに文章を作る必要がなくまた自然言語処理をほとんど必要としないため実現が容易である．

抽象化 *(Abstracting)* はテキストをより一般的な概念に置き換える手法である，言い換えると, 原文には明示的には現れない文章も生じてよい要約技法である．この手法は抜粋よりも一貫性があり高度な要約が期待できるが，対話理解や自然言語処理，オントロジ処理などの高度な技術が必要となる．

クラスタリング *(Clustering)* はオブジェクト集合へのグルーピング手法であり, 同じクラスタ内のオブジェクトは類似し異なるクラスタのオブジェクトは似ていない様に振り分ける [11].

本稿では Web 検索システムにおける問合せ処理と検索結果の表示において我々の提案した自動要約手法を用い, その有用性を示す．我々はこれまでにハイパーリンクと語の共起性から Web 文書の集合を抽出することで, 類似した内容を持つ Web 文書同士の集合を得る手法を提案した [23]．さらにこの Web 文書集合を *semantic textual units* (STU) という意味単位に分割する．文献 [25] において, STU から階層構造を生成し, 階層構造のノードに STU によってラベル付けする手法を提案した, これを階層表現による要約と呼ぶ.

そこで, 本稿ではこの階層表現による要約に問合せをし, この階層構造を検索結果に反映させる手法を提案する．階層構造の各ノードと問合せ語との類似度を計算することで, 問合せ語と適合するノードのランキングを得る．そして, このランキングの中で親子関係にあるノードは, その関係を検索結果として出力することで階層構造を持つ検索結果を得る.

本稿で提案する問合せ手法を用いることで, 利用者が検索結果から合致する Web 文書への URL を探すときに階層構造は効果的に働く．より詳細な内容を求めるならば下位のノードの, 全体の内容を把握するならば上位ノードのラベルを手がかりとしてブラウズすることができる．そしてこの階層構造の各ノードに含まれる URL は, そのノードのラベルによって URL のリンク先の内容を素早く把握することができる．また, さらに検索結果の階層構造の親ノードや子ノードをも抽出の対象とすることで, 問合せ語を直接含まない Web 文書への URL も検索結果に含むことができる.

次章で自動要約を用いた問合せ処理の関連研究について述べる．第 3 章で既に我々が提案した階層表現による要約手法の概要を要約する, 詳細は文献 [25] を参照．第 4 章で本稿で提案する階層表現による要約への問合せ手法を，第 5 章で実験結果を挙げ本手法の有用性を示し，第 6 章は結論である.

2. 関 連 研 究

問合せ処理に自動要約の手法を用いた手法がいくつか提案されている．Google や Yahoo に代表される多くの検索エンジンでは要約として問合せ語を含む箇所の抜粋を行う [1], [12]．そこで, 問合せ語を含む文章を対象として重要文抽出法を用いる手法 [26] が提案された．しかし，この手法で生成された要約は問合せ語を含む箇所の抜粋とあまり変わらない結果を抽出した.

また, 問合せ語を含む箇所ばかりを提示してもどの文書が求める情報を含んでいるかを的確に判断することは困難な作業となるという考えから, 問合せ語に適合する文書集合と適合しない文書集合を利用して要約文を抜粋する手法 [21] もある．この手法は問合せ語に関連する文書集合を抽出し, その集合内で問合せ語とその共起語を計算することで重要文を抽出した．一方, 問合せ語とその共起語以外の語によって, 抽出されなかった文書集合から重要文を抽出することで問合せ語に直接マッチしない文章を抽出することを目指した.

また, 文章ではなく lexial chain という語の並びを抽出する要約手法がある [17]．同じ文章で共起している語は意味的に類似しているという考えに基づき, 類似した語の並びを lexial chain と呼び, 要約として抽出した．この手法はニュース記事によるコーパスで実験がなされている．Web 文書はタグやハイパーリンクといった構造データとテキストデータからなる半構造データであり, 著者が複数存在する文書, 文法の誤りのや造語を含む文書もあることから, この手法の Web 文書への適応は自明では無い.

3. 階層表現による要約手法

我々は新しい自動要約手法として構造化を提案した [25]．**構造化** *(Structuring)* は文書をデータ構造を用いて表現する手法である．データ構造はその構造自身が意味を有するために, 文章を読まなければ全体を把握することのできない従来の自動要約手法に比べ, 明瞭かつ簡潔に表現することが期待できる．しかしながら, どのような構造が文書を要約として適切に整理することができるのかは自明ではない.

また, このとき対象となる処理単位は，単語や語句などのように最小の意味単位であるか，文章などのように一定の意味まとまりを持つものである．前者の場合では精密に扱えるが，単語・語句間の関連の表現が詳細で全体像を捉えにくい．逆に後者では，内容の大筋や概要は記述方法に依存するが，文章間の関連が対応させやすく全体を捉えやすい．文書内容の意味構造を記述するために有向グラフや木構造を用いることで, 文書内容を多段階に表すことを期待できる．木構造では, 根に近いレベルであれば概観や大域的な, 葉に近いレベルであれば詳細や局所的な観点に対応している．*Key Graph* [18] は重要語の関連をグラフ表現する技法である．しかし繋がりに対して大要も細部も同時に表現するため，抽象度に応じて多段階に解釈することは容易で無い．これより, 文章を最小の処理単位として木構造をもちいた構造化に着目する.

この章ではまず,Web 文書集合を類似した集合に分けるために組み合わせクラスタリングについて述べる [23]．次に, この Web 文書集合の自動要約の手法として階層構造を用いた要約手法ついて述べる.

3.1 組み合わせクラスタリング

本節と次節で階層表現を用いた Web 文書集合の階層的要約手法を提案する．最初に, 相互に類似した Web 文書集合を得る手法について述べる．このとき, 主な問題の１つが莫大な量の Web 文書からの適切なページ集合を抜粋する方法である．単純にクラスタリングを用いれば, 小数の巨大クラスタと多数の微細なクラスタが生成されることが一般的に知られている．しかし我々は既にベクトル空間モデルによるクラスタリングとハイパーリンクの共起性に基づいたクラスタリングを組み合わせた Web 文書クラスタリング手法を**組み合わせクラスタリング**と呼び, その有用性を示した [23]．これにより我々は適当なトピックに対応する適切な Web 文書集合を得ることができる．ここでは**組み合わせクラスタリング**の概要を要約する, 詳細は文献 [23] を参照.

Web **文書クラスタリング**は類似した内容の Web 文書集合を得ることを目的とするクラスタリングである．我々は Web 文書に対して，”文書特性”と”ハイパーリンク”による構造を利用したクラスタリングが適用する．文書特性を利用したクラスタリングでは,Web 文書は (通常のテキストクラスタリングと同様に) ”単語の多重集合” (Bag of Words) として表現される [11]．各文書はベクトルで表され、全体としてベクトル空間を構成する．ベクトルの各要素は対応する単語の出現頻度に対応し, 文書間の類似度を対応するベクトル間の余弦 (余弦) 値を用いて記述する．索引語により Web 文書をベクトル化し, ベクトル集合をクラスタリング (VSM クラスタリングと呼ぶ) を行う．一方, **ハイパーリンク** (他の Web 文書への参照) は，Web 文書間の意味的な結びつきを明示的な構造で表すと言う点で重要である．この考えに基づき, ハイパーリンクの共起性を利用したクラスタリング (Link クラスタリングと呼ぶ) を行う，この２つクラスタリングの結果を”組み合わせる”ことにより, 同一のトピックを参照し, かつ文書の酷似しているクラスタへと分割する．

ここで組み合わせクラスタリングの例を示す．頂点 (node) を Web 文書に, 辺 (arc) をハイパーリンクに対応させれば, Web 文書集合上のハイパーリンク構造は有向グラフで表現することができる．図 1 のように 6 個の頂点 $a_1 \cdots a_6$ があるとき頂点 a から出る辺の集合 $From(a)$ を a からの出辺集合 (要素数を出次数)，逆に頂点 b へ入る辺の集合 $To(b)$ を入辺集合 (要素数を入次数) という．同じ参照先への出辺数の割合を用いて類似度として階層型クラスタリングを行う．このプロセスから得られるクラスタを *LINK* **クラスタ**と呼ぶ.

次に,6 個の Web 文書集合 $a_1, \cdots, a_6$ に対応して文書ベクトルが図 2 で与えられているとする．これにより得られるクラスタを *VSM* **クラスタ**と呼ぶ.

図 3 は例 1 の Link クラスタを A_1, A_2 を円形で, 例 2 の VSM クラスタ B_1, B_2 を矩形で表している．Link クラスタと VSM クラスタを重ね合わせると, クラスタ $C_{11} = \{a_1, a_4\}$ と, $C_{22}\{a_3, a_6\}$ に分割される, これを組み合わせクラスタと呼ぶ. クラスタ $C_{12} = \{a_2\}$ と $C_{02} = \{a_5\}$ はクラスタが小さすぎるため破棄される．

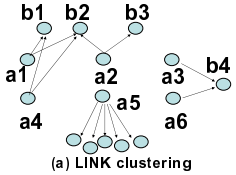

図 1　LINK Clustering

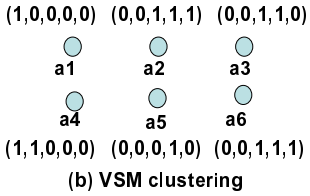

図 2　VSM Clustering

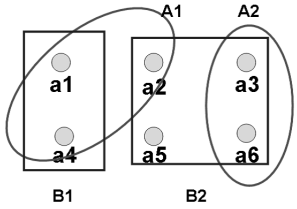

図 3　Combination Clusters

3.2 階層表現による要約

前節で我々はどのように Web 文書集合を得るかを述べた．組み合わせクラスタリングから類似した内容をもつ Web 文書集合を抽出できたと仮定し, その Web 文書集合の階層表現による要約を生成する手法について述べる [25].

Web 文書に対して構造化による要約を適応することを考える．Web 文書は文字列部とタグ部から多重に構成されている．HTML 言語ではタグ付け対象となる部分を要素と呼び, 文章の構造 (見出しやハイパーリンクなど) や, 修飾情報 (文字の大きさや組版の状態など) を記述する．つまり, 整合した Web 文書において要素はタグの持つ意図を反映した完結した意味的まとまりを有すことから, タグで囲われた部分が Web 文書を構成する最小の単位の文章であるとする. 我々はこれを *Semantic Textual Unit* (STU) と呼ぶ．本稿で対象とするタグは `<P>` `<UL>` `<OL>` `<DL>` `<TITLE>` `<TABLE>` `<BLOCKQUOTE>`である.

図 4 に示すように我々は Web 文書集合から STU を抽出し, 階層型クラスタリングを用いることで階層構造を得ることができる．最後に階層構造の各ノードにラベル付けすることで階層表現による要約を得る.

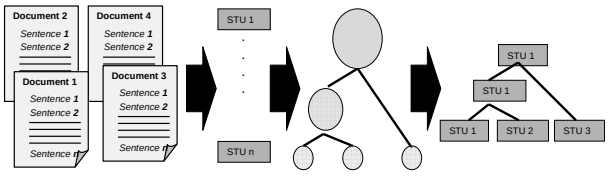

図 4　overview

STU を生成する時, 我々は二つのタグの入れ子構造とリンク

に着目する. HTML 言語ではタグが多段階の入れ子構造になることを許すため, 通常, ある要素にタグを複数指定する場合はタグを入れ子構造にする. 次のようにタグが入れ子構造になっている場合, どのように STU を抽出するかを示す.

```
<blockquote>
    <p>要素 1</p>
    <p>要素 2</p>
</blockquote>
```

要素 1, 要素 2 はそれぞれ<P>に囲まれた要素であり, また{要素 1, 要素 2}は<blockquote>の要素でもある. このとき抽出される STU は :STU1 = {要素 1} , STU2 = {要素 2} , STU3 = {要素 1, 要素 2} の 3 個の STU が抽出できると考える. 即ち,STU 内のタグを解析することで内部のタグによる要素もまた STU であるとみなす. この結果, クラスタリングは Web 文書の内部構造も反映させた結果を生む.

Web 文書の関連を表すタグ<A HREF>は, リンク先の内容を示唆しているとみなし, リンク先の Web 文書構造と<A HREF>を置き換えて処理する.

STU のモデル化にベクトル空間モデルを用いる. 本稿では単語として, 連続する漢字・カタカナを利用する. Web 文書にはしばしば文法的に正しくない表現が含まれるため, 形態素解析などの文法的な体系付け手法は適さない. また文書に出現する単語を減らし, ベクトル表現の次元数を縮小するために, Zipf の法則を用いる.

階層型クラスタリングは, 各クラスタ間の距離が計算され最も距離の近い二つのクラスタが逐次的に併合される. 一つのクラスタに併合されるまで繰り返すことで最終的に階層構造を得る. この結果の階層構造は類似度とクラスタ構成方法に依存する. 本稿では群平均法 (average linkage method) による構成方式を用いる. そしてクラスタリングによって得られた各文書ベクトルの平均値を計算し, 平均ベクトルから最も近い STU を重心 (centroid) とする. このとき各クラスタは重心 STU によってラベル付けする. 最終的に, Web 文書集合から重心 STU でラベルづけられた階層表現を得る.

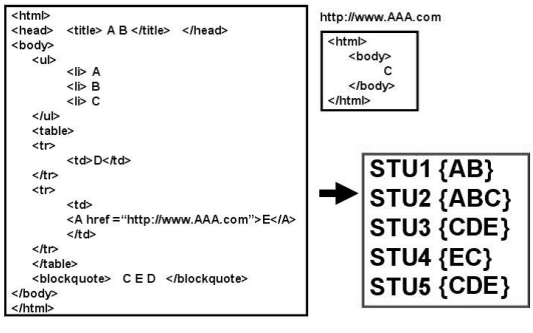

図 5 Taking STUs from Web Pages

図 5 のように、2 つの Web 文書と単語 A,B,C,D,E に対して本手法の例を示す. STU1 は<TITLE>, STU2 は<UL>, STU5 は<BLOCKQUOTE>タグに囲まれているので STU として抽出する.<A>で囲まれた単語 E とリンク先の単語 C から STU4 ができ, また STU3 は STU4 を入れ子に持つ. こうして 5 つの STU が生成され. これらの STU を群平均法による階層型クラスタリングした結果を図 6 に示す.

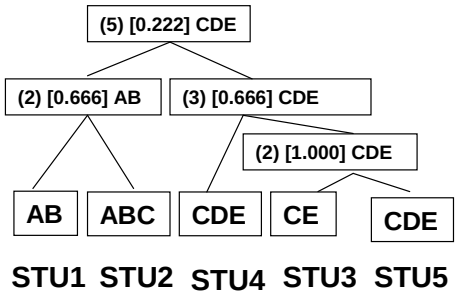

図 6 Hierarchy using STUs

3.3 階層表現による要約の評価方法

自動要約の結果の評価については既にいくつかの提案がある [13]. まず最初に着目された評価尺度は圧縮率である. 圧縮率とは原文に対する要約文の長さを意味し, 高い圧縮率の要約では原文の内容を全て包括しているわけではない. また, こうした評価尺度は階層構造を用いた要約に適応することができない. 我々は階層表現による要約を定量的に評価する方法として CHR Method を提案した [24]. この手法は階層表現のノードの可読性, 階層の可読性, 読解の三つの視点に基づいて評価する手法で, ここでは特にノードの可読性の概要を要約する.

ノードの可読性とは重心の STU がノード内のトピックを包括した内容を表しているかを評価する尺度である. この尺度のために粒度という内包する要素の相互距離によって定義される概念を導入する. 粒度が細かいノードは, ノード内の要素数にかかわらず類似した内容を示す. ノード内の STU が類似した内容であるならば重心の STU によって容易に内容を把握することができる.

我々は粒度の荒いノードにペナルティを課す尺度を導入し可読性を評価する. ノード内の粒度は細かく, ノード間の粒度は粗いようにノードが構成されていることが理想的であることから, ノード内の粒度 G_{in} とノード間の粒度 G_{out} からノードの可読性 C_{det} を定義する.

$$G_{in} = 1 - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i}^{n} sim(x_i, x_j) \tag{1}$$

クラスタ要素を $x_k(k : 1 \ldots n)$, 要素同士の類似度を $sim(x_i, x_j)$ とする. クラスタ $Cl_r(r : 1 \ldots s)$ としたとき Cl_i と Cl_j の群平均法に基づくクラスタ間類似度を $sim(Cl_i, Cl_j)$ とした場合、

$$G_{out} = \sum_{i=1}^{s} \sum_{j=i}^{s} sim(Cl_i, Cl_j) \tag{2}$$

こうした二つのコストの線形和によってノードの可読性を評価するコストは定義される.

$$C_{det} = G_{in} + G_{out} \tag{3}$$

4. 階層表現による要約への問合せ

本稿では階層的要約を用いた問合せ処理を行うことにより, 従来の Web 検索結果では内容の把握が困難であった問題や, 問合せ語を含まない Web 文書への問題を解決する. 前章で我々は類似した Web 文書の集合とその階層表現による要約を生成する手法の概要を述べた. これらの手法がすでに適応されたと仮定して問合せを行う. つまり,Web 文書は類似した内容を持つ文書同士を集合に分け, 各 Web 文書集合には階層表現による要約を適応した状態である. 我々の提案する問合せ処理は次の手順で行う.

- ルートへの問合せ
- 下位ノードへの問合せ
- 階層の抽出

最初に, 各 Web 文書集合の階層表現のルートノードに対して問合せとの類似度を計算する. 階層表現による要約ではルートノードが Web 文書集合全体の単語の分布の重心を保持することから, ルートノードは Web 文書集合全体を抽象化した内容であると考えることができる. そして, 問合せとの類似度が閾値以下場合はこの Web 文書の集合はこの後の処理の対象からはずすことができる. つまり, ルートノードは Web 文書集合内を把握するための要約として利用することができる. このとき利用する類似度は, 余弦類似度と bGIOSS 類似度 [10] のいずれかを用いる。問合せ語とノードとの余弦類似度は以下のように示される.

$$cos(q\,N) = \frac{q \cdot N}{\|q\|\|N\|} \tag{4}$$

q と N はそれぞれ問合せ語とノードの単語ベクトルを表す. bGIOSS 類似度 [10] は以下のように示される.

$$bGIOSS(Q\,N) = \|N\| \sum_{i=1}^{n} \frac{\|q_i\|}{\|N\|} \tag{5}$$

N はノードの単語ベクトルを, 単語数 n 個の問合せ語は $Q = \{q_1 \ldots q_n\}$ と表す.

次に, 下位ノードへの問合せを行う. ルートノードが問合せ語と閾値以上の類似度であった Web 文書集合の全ノードに問合せ語との類似度を計算する. ここでも余弦類似度と bGIOSS 類似度のいずれかを用いる。そして, 我々は問合せ語と類似したノードのランキングを得ることができる. ここで STU と Web 文書の関係に着目する. STU は Web 文書をタグ構造とリンク構造から分割して生成したものである. 分割前の STU は必ずいずれかの Web 文書に含まれていることから, STU は分割前の Web 文書の URL と関連づけることができる. また,STU がリンク構造を持つ場合はそのリンクとも関連づけられる. すべてのノードは STU を保持していることから, ノードは URL と関連付けることができる. そして, ノードの保持する URL の要約として重心 STU の文章をラベル付けすることができる. これより, 我々は問合せと類似した重心 STU と URL を持つノードのランキングを得ることができる.

最後に, 階層の抽出を行う. ノードの類似度ランキングにおいて, 高い類似度のノード同士が親子関係にある場合はその関係ごと抽出することで問合せの結果が階層構造を持った要約として出力することができる. さらにこの抽出した階層構造の親や子のノードが問合せ語と類似していなくても, 階層表現の評価方法として利用したノードの可読性 c_{det} がより低いコストのときには抽出の対象として扱う. これにより, 問合せ語を直接含まない Web 文書の URL が階層構造として抽出することが期待できる.

ここで階層表現による要約への問合せの例を示す. 問合せベクトルを $Q = \{0\,0\,0\,0\,1\}$ とする. まず, ルートへの問合せとして, 図 7 の四つのクラスタに余弦類似度を用いる. クラスタ 1 は類似度 0.2, クラスタ 2 が 0.33, クラスタ 3,4 は 0.0. 閾値は 0.1 以上とすると, クラスタ 3 と 4 を次の処理の対象から外れる.

次に, 下位ノードへの問合せで, 図 8 に示す 4 つのノードに着目する. ノード 1 は類似度 1.0, ノード 2 は 0.2, ノード 3 は 0.5, ノード 4 が 0.33. ノードのランキングはノード 1,3,4,2 の順となる.

最後に階層の抽出は, ノード 1 とノード 2 は親子関係であることから, その階層構造ごと抽出する. そしてノード 1 位からノード 2 の親ノードと子ノードの可読性を計算し, ノード 1 の下位ノードとノード 3 の親ノードが低いコストであった場合に出力される問合せの結果を図 9 で示す.

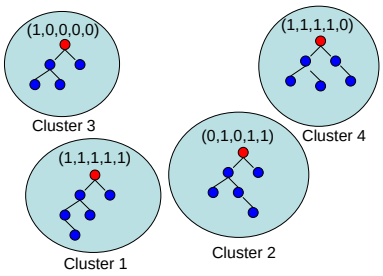

図 7 querying to root node

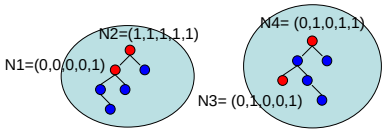

図 8 querying to nodes

5. 実　　験

5.1 実験環境

本稿では, 実験データとして NTCIR-3 を使用する. NTCIR-3 は.jp ドメインの html 及び txt データを集めたテストコレクションである. この中から 2001 年 9 月 29 日から 2001 年 10 月 5 日までに収集した 9929 件の Web 文書を対象とする. 階層表現に問合せ処理を行うためにこの Web 文書に組み合わせクラスタリングを行い, 我々は 6 つのクラスタが得ることができた. そしてこれらのクラスタに階層表現による要約を適用する. この階層表現への問合せ処理の評価を以下 3 点において行う.

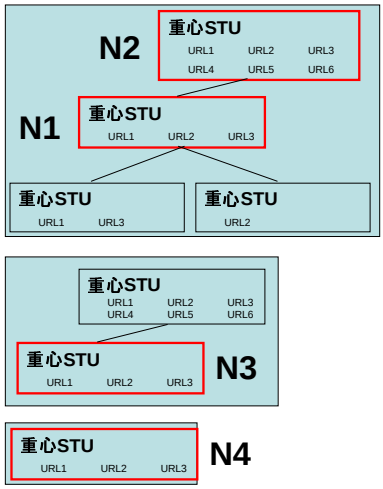

図 9　querying result

- HITS アルゴリズムの URL との適合率と再現率
- 余弦類似度と bGIOSS 類似度の比較
- 抽出した木構造の詳細

5.2　HITS アルゴリズムの URL との適合率と再現率

まず最初に二種類の問合せにおける HITS アルゴリズムの URL との適合率と再現率の比較を行う．ノードが HITS アルゴリズムの URL を多く含む割合 (カバレッジ) が高ければ理想的なノードとなる．このカバレッジを評価するために再現率を用いる．一方,HITS アルゴリズムの URL 以外の URL はノードにとってノイズとみなすことができる．よってノイズの少なさを評価するために適合率を用いる．

階層表現は上位ノードになれば Web 文書全体の内容をカバーする抽象的な要約となり, 下位ノードでは個々のトピックの詳細な要約となる．同様に, ノードが保持する URL も下位になるほど問合せと強く類似した URL のみが含まれることになるが,URL のカバレッジも悪くなる．そこで HITS アルゴリズムによって実験データの 9929 件の Web 文書にランク付けを行い, このランクとノードの保持する URL を比較することで提案手法の精度を評価する．

HITS アルゴリズムは Authority 値が高いほど特定のトピックに関する情報が豊富であることを表すことから, ノードの保持する URL と HITS アルゴリズムによるランクで問合せ語を含む URL の URL の個数から適合率と再現率を求めることは HITS アルゴリズムの考えと合致しない．そこで URL の Authority 値によって重み付けをした適合率と再現率を以下のように定義する．

$$\text{適合率} = \frac{w(URL_N \cap URL_{HITS})}{w(URL_N)} \tag{6}$$

$$\text{再現率} = \frac{w(URL_N \cap URL_{HITS})}{w(URL_{HITS})} \tag{7}$$

ある URL の Authority 値を $w(URL)$, ノードが保持する URL 集合を URL_N ,HITS ランクで問合せ語を含む URL 集合を URL_{HITS} とする．

問合せ語 { フィルタ } で問合せたとき, 階層の深さに対する適合率と再現率の最大値との関係を図 10 に示す．

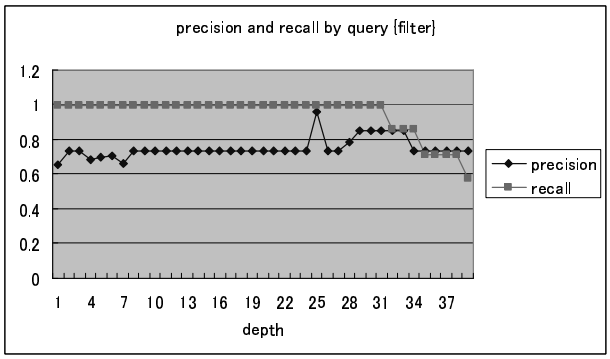

図 10　precision and recall by query {filter}

問合せ語 { フィルタ, 実験 } で問合せたときの関係を図 11 に示す．

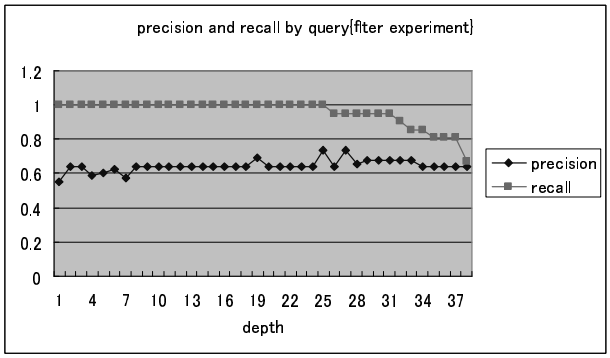

図 11　precision and recall by query {filter , experiments}

再現率は図 10,11 共に深さ 30 前後でほぼ 1.0 となっている．これは HITS アルゴリズムによるランクで authority 値の高い URL を包括することができていることを示している．適合率は深さ 25 前後で最大の値を示している．深さ 25 より上位のノードではノイズとなる URL を含んでしまうために適合率が下がっている．このため深さ 25 前後のノードを抽出する類似度が望ましいことがわかる．

5.3　余弦類似度と bGIOSS 類似度の比較

次に二種類の問合せにおける余弦類似度と bGIOSS 類似度の比較を行う．問合せ語 { フィルタ } で問合せたときの階層の深さと余弦類似度と bGIOSS 類似度の最大値との関係を図 12 に示す．

問合せ語 { フィルタ, 実験 } で問合せたときの関係を図 13 に示す．

図 12 から, 問合せ語が 1 語の場合, どちらの類似度も類似した傾向を示し, URL の適合率と再現率が高かった深さ 25 前後で類似度が高い値をとっていることからどちらの類似度も理想的な抽出に貢献している．しかしながら, 図 13 では bGIOSS 類似度がより下位の階層のノードで高い類似度を示している．これは bGIOSS 類似度は問合せ語の語数の数だけノードのサイズで正規化を行うことから, よりサイズが小さいノードを選び

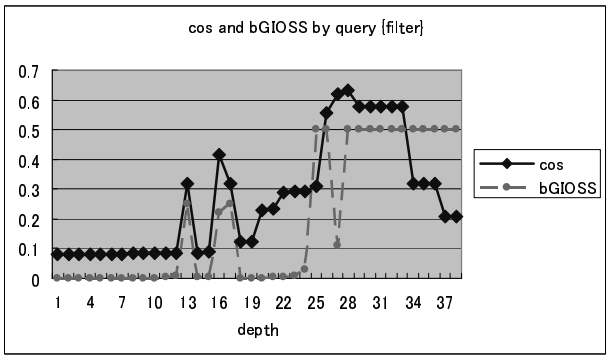

図 12　cos and bGIOSS by query {filter}

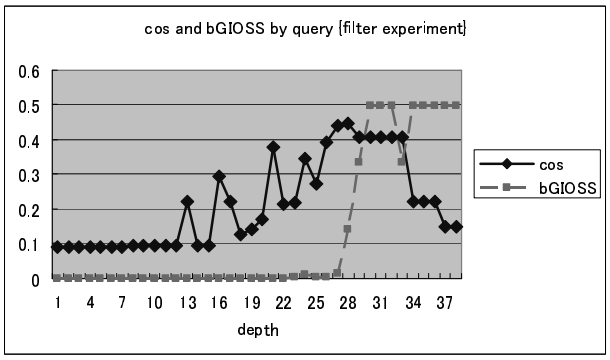

図 13　cos and bGIOSS by query {filter , experiments}

やすくなるという傾向にあることがわかる．それ故, 深さ 25 前後で類似度が高い値をとっている余弦類似度が有効であった．

5.4　抽出した木構造の詳細

問合せ語 { フィルタ } で問合せたとき余弦類似度が高い上位 4 つのノードの抽出を行った結果を図 14 に示す．

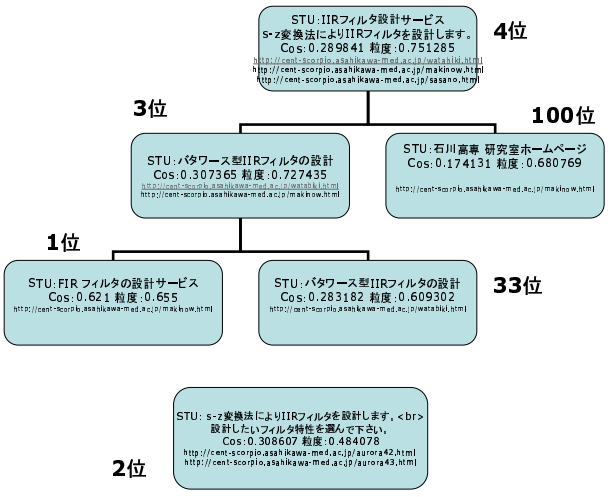

図 14　hierarchy by query {filter , experiments}

1 位,3 位,4 位のノードが親子関係であることから階層表現が抽出した．そして 3 位と 4 位のノードよりも子ノードが C_{det} 低い値をとっていることから 33 位と 100 位という問合せ語と合致していないノードも抽出できている．下位ノードの重心 STU の内容はフィルタに関する内容であり, 上位ノードではフィルタと電気工学科に関する内容であることから, 上位ノードになるほど抽象的な内容になっていることが確認できる．ノードが保持する URL に関しても下位ノードではフィルタに関する Web 文書への URL であり, 上位ノードでは電気工学科に関する URL や大学の研究室への URL などを含んでいる．これより階層表現を用いて問合せの結果を表示手法は従来の Web 検索結果では困難な内容把握や, 問合せ語を含まない Web 文書などの問題を解決している．

6.　結　　論

本稿では Web 文書集合を階層表現により要約し問合せをする手法を提案した．提案手法による問合せは従来の Web 検索では困難だった素早い内容把握を容易に実現可能なことを実験により示した．今後の課題としては, 階層的要約の生成時に必要となる記憶域についての議論が必要である．本研究では階層的要約の生成には階層型クラスタリングを用いた．このとき全要素の類似度行列を必要とするため大量の記憶域を消費してしまう．そして要約の対象となる Web 文書に変更が生じた場合, 階層的要約を再度生成しなければならない．これらの課題に対して次元縮小を用いたり, やあるいは動的な要約生成法プロセスの必要性は, 実際に実用することを考えると対処が必要となる．

謝　　辞

本実験に対しては国立情報学研究所より NTCIR-3 Web 文書データの提供をいただきました. 関係各位に深く感謝します.

文　　献

[1] S, Brin. L, Page.: TThe Anatomy of a Large-Scale Hypertextual Web Search Engine, Computer Networks and ISDN Systems, 1999

[2] Buyukkokten, O., Garcia-Molina, H.and Paepcke, A.: Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices, In Proceedings International WWW Conferenc(2001)

[3] Chakrabat,S.: Mining the Web, Morgan Kaufmann, 2003

[4] Cutting, D., Karger, D., Pedersen, J. and Tukey,J.W.: Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, SIGIR, 1992

[5] Delort, J.-Y., Bouchon-Meunier,, B.and Rifqi, M.: Enhanced web document summarization using Hyperlinks ", Proceedings of the 14th ACM conference on Hypertext and hypermedia, pages 208-215, New York, NY, USA, ACM Press (2003).

[6] Ganti,V., Gehrke, J. and Ramakrishnan, R.: CACTUS Clustering Categorical Data Using Summaries, Knowledge Discovery and Data Mining (KDDM), 1999

[7] Gibson,D., Kleinberg, J. and Raghaven, P.: Clustering categorical Data, An Approach Based on Dynamic systems, VLDB, 1998

[8] Grossman,D. and Frieder,O.: Information Retrieval – Algorithms and Heuristics, Kluwer Academic Press, 1998

[9] Guha, S., Rastogi, R. and Shim, K.: ROCK: A Robust Clustering Algorithm for Categorical Attributes, ICDE, 1999

[10] P. Ipeirotis, and L. Gravano, .: When one Sample is not Enough: Improving Text Database Selection Using Shrinkage, Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, 2004

[11] Jain,A.K., Murty, M.N. and Flynn, P.J.: Data Clustering: A Review, ACM Computing Surveys 31-3, 1999

[12] Kleinberg, J.M. : Authoritative Sources in a Hyperlinked Environment, JACM 46-5, 1999

[13] Mani, I.: Automatic Summarization, John Benjamins, 2001
[14] Mori, M., Miura, T. and Shioya, I.: Labeling Temporal Cluster of Web Pages, DBSJ Letters 3-2, 2004, pp.109-112 (in Japanese)
[15] Mori, M., Miura, T. and Shioya, I.: Extracting Events From Web Pages, proc. AISTA, 2004
[16] Mori, M., Miura, T. and Shioya, I.: Abstracting Temporal Clusters, proc. ITA, 2005
[17] Okumura, M., Mochizuki, H. and Nanba, H. : Query-biased Summarization Based on Lexical Chaining, In Proceedings of PACLING'99, pp.324-334, 1999.
[18] Yukio Ohsawa, Nels E. Benson and Masahiko Yachida: KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor Proc. Advanced Digital Library Conference (IEEE ADL'98), pp.12-18, 1998
[19] Radev, D. and Fan, W. : Automatic summarization of search engine hit lists, proc ACL'2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, 2000, Hong Kong
[20] Radev, D., Jing, H. and M. Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, Information Processing and Management, 2004, pp.919-938
[21] Sakurai,T.and Utsumi,A.: Query-based Multidocument Summarization for Information Retrieval. in Proc. of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering, and Summarization, pp452-458, 2004
[22] Stefanowski, J., Weiss, D.: Carrot2 and Language Properties in Web Search Results Clustering, Atlantic Web Intelligence Conference, 2003
[23] Takahashi, K., Miura, T. and Shioya, I.: " Combination Clustering for Web Correlation ", IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), pp.434 - 437, 2005
[24] Takahashi, K., Miura, T. and Shioya, I.: Hierarchical Summarizing and Evaluating for Web Pages, ICDT Workshop on Emerging Research Opportunities in Web Data Management(EROW), 2007
[25] Takahashi, K., Miura, T. and Shioya, I.: Summarizing Web Pages Hierarchically, International Association for Development of the Information Society Applied Computing (IADIS-AC), pp.612-617, 2006
[26] Tombros,A.and Sanderson,M.: Advantages of query biased summaries in information retrieval. In Proceedings of the 21st annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ' 98), 2-10. 1998
[27] Trieschnigg, D. and Kraaij, W.: Scalable Hiearachical Topic Detection, SIGIR, 2005
[28] Zamir, O. and Etzioni, O.: Web Document Clustering – A feasibility Demonstration, SIGIR, 1998
[29] Zipf, G, K.: The human behavior and the principle of least effort, Addison Wesley, 1949