Web文書集合の自動要約に関する研究

高橋, 功 / TAKAHASHI, Ko

(発行年 / Year)
2007-03-24
(学位授与年月日 / Date of Granted)
2007-03-24
(学位名 / Degree Name)
修士(工学)
(学位授与機関 / Degree Grantor)
法政大学 (Hosei University)

2006 年度　修士論文

Web 文書集合の自動要約に関する研究

STUDIES ON AUTOMATIC SUMMARIZATION
FOR WEB PAGES

指導教官　三浦 孝夫 教授

法政大学大学院工学研究科
電気工学専攻修士課程

高橋 功
Kou TAKAHASHI

目 次

第1章　序論

1.1　扱う問題

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, www 上に存在するデータ量は爆発的に増加し続けている．そして,google や Yahoo のような検索エンジンや Web ディレクトリサービスを用いればこの膨大な量のデータには世界中の誰もが容易にアクセスすることができる. こうしたデータを全て閲覧することは物理的不可能であるため, 利用者にとって興味のある情報を得るためには情報の取捨選択が必要不可欠となっている. しかしながら, 利用者にとって目的に合致しないデータや, 利用者にとって既知の内容のデータ, あるいは目的に合致しているが長大な内容のため全体の内容を把握するために多大な時間を要するといったデータも存在する. それゆえに利用者が自分の要求に合う内容のデータなのかを, 実際にデータ全体を閲覧することなく素早く, 正確に, 容易に把握したいというニーズを満たす技術を実現することを本研究の目的とする.

利用者が素早く内容を把握するためには, 元々の対象の内容よりも短く簡潔に表現する必要がある. 例えばニュース記事のタイトルからはニュース記事の内容を把握するために役立つ. そして記事のタイトルのように非常に短い文章を読むことと, ニュース記事全体を読むのでは圧倒的にタイトルのみを読むほうが短い時間で済む. このように素早く内容把握するためには, 対象となるデータをより短く表現するための方法が必要となってくる.

しかしこの場合, より正確に内容を把握するためにはタイトルだけでなくニュース記事全体を読み進めていく必要がある. つまり, 素早く内容把握するためには対象より短く表現されたデータが望ましいが, 正確に内容把握するためには対象の内容を全て読むことが望ましいことから, 内容把握の素早さと正確さはトレードオフの関係にある.

これらのことから, 本研究で実現すべき技術は素早さと正確さを持ちつつ, さらに利用者にとって内容把握のための閲覧が容易な技術が必要となる.

また, こうした技術を評価する方法を確立することは重要項目である. 定量的に評価する方法があれば, いくつかの仮説のなかから最適な方法を選択することが可

能になり, その評価は新たな仮説を導くことにつながるからである. しかしながら, ニュース記事などの自然言語で記述された文書を人間は理解し, 意味を解釈することができる. 一方, 計算機が自然言語で記述された文書の意味を解釈することは非常に困難な問題である. 計算機の処理とは数値による処理であるため, 人間の主観を介在させる余地はないためである. そこで本研究の取り扱う問題の一つとして, 評価対象をどのような客観的数値化手法を用いて評価するのか, その数値化手法は真に有効な手法であるかを検証する必要がある.

次節では, 本研究の目的を実現するために, 過去の関連研究について論じる.

1.2　問題の背景

情報から利用者にとって最も重要な情報を抜き出す手段として，**自動要約** *(Automatic Summarization*）と呼ばれる技術に注目が集まっている [15]. 自動要約は文書，画像，映像，音楽などのマルチメディア情報を対象とし, 短く, 簡潔な形で利用者が内容を素早く把握できることを目指す．こうしたマルチメディア情報を対象とした自動要約手法の応用例としては, Microsoft Office の AutoSummarize オプション, 患者の診断記録に関する医学文献の要約や画像を提供する医師の支援への応用 [19], 聴覚障害者のためにテレビのニュース放送へ字幕を自動生成する応用 [33] や, 会議の内容を復習するために自動音声認識と自動要約を用いて後でブラウズ可能にする応用 [34], などがある. 1950 年代から始まった自動要約の研究において，文書（テキスト）データが最も中心的な要約の対象として研究されてきた．ここで, 文書データは自然言語によって記述されたものであると定義し，データベースシステムにおけるスキーマなどの構造を仮定しない．これまでは単一の文書を要約の対象として盛んに研究がなされ, 文書データ (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することができる.

要約の対象となる文書データが複数の文書であることを**複数文書要約** *(Multi Documents Summarization)* と呼ぶ．Web 上の膨大な量の文書データに対してこの複数文書要約を用いることで, 一見するだけで文書集合全体が何に関するものであるのかがわかることが望ましい. しかしながら, 複数文書要約にはいくつかの問題がある. それぞれの文書で類似の情報が論じられていたり, あるいはまったく別の情報について論じている場合がある. また, それぞれの文書で著者が違うために, 同じ情報について論じているにもかかわらず出現する単語の傾向は異なる場合もあることなどを考慮すると, 単一の文書に対する自動要約手法がそのまま適応可能かは自明ではない.

自動要約分野において評価方法は非常に興味深い問題である．要約を定量的に評価する尺度として**圧縮率** *(compression ratio)* がある [15]. 圧縮率とは原文に対す

る要約文の長さを意味する．圧縮率は要約の生成手法に依存することなく容易に評価することが可能である点で非常に重要である．しかし, 要約が利用者にとって読みやすいかや, 要約が内包する意味といった側面を圧縮率で評価することはできない．これを要約の内的評価と呼ぶ．いくつかの研究では, 異なる人間の要約を比較し, 相互の要約の重なる箇所から理想的な基準となる人間の要約を生成するという手法をとってきた [15]．これまで多くの場合, このような基準となる人間の要約と比較することで内的評価を行ってきた．しかしながら, 要約とは利用者の要求に対して重要な情報を提示することを目指すため, 一つの文書に対して一つの要約が一意に決まるというわけではない．そのため利用者の要求を考慮することで, 内的評価はいっそう複雑で困難になる．

自動要約手法を Web 文書集合に対して適用する一つの例として, 情報検索への適用がある．利用者の目的に合致したデータを取り出すための技術を情報検索と呼ぶ．例えば, 利用者にとって興味のある情報を得るために問合せを WWW 上の検索エンジンに与えると, 検索結果として膨大な量の文書データを得ることができる．しかしながら, このような膨大な量の文書データを閲覧しこの中から利用者にとって興味のある情報を探すことは長大な文書データを閲覧することと同義であり, 膨大な時間を要する作業となってしまう．

本研究では, こうした情報検索結果から素早く文書データの内容を把握することのできる新しい複数文書要約技術を実現し, 合わせてその評価方法を提案することを目指す．電子化された文書が急激に増加する現在においては，文書全体を閲覧するよりも容易に素早く内容を把握できれば情報の取捨選択を助けることとなり，大きな意義を持つと考えられる．

1.3　個別機能要件

本研究では Web 文書集合全体を閲覧することなく, Web 文書集合の内容を素早く把握するために新しい自動要約手法を実現する．

これまで多くの自動要約の研究者の焦点は, 自然言語によって記述され, 構造を有さない文書データを対象とした単一文書要約に向けられてきた．それに対して, 本研究で取り扱う問題としては, Web 文書を対象とし, 複数文書要約を行うこと, そして生成された要約をどのように評価するのかがある．最後に,Web 情報検索の結果へ適用した場合の問題についても扱う．

1.3.1　Web 文書

ここで，まず文書データから文書間の類似度を計算機で統計的に扱えるようにする方法について考える．文書の内容をどのような単位で抽出するかがプロセスの有効性に大きく影響する．このとき，文書を単語の集まりとして考える方法である”set of words”や”bag of words”と呼ばれる方法が一般的である [31]．文書 x は重み $x_1, ..., x_d$ をもった単語の連続として，ベクトル $\vec{x} = (x_1, ..., x_d)$ と表現される．ここで d は文書集合内で出現した単語の数である．このとき, 日本語で記述された文書データから単語を取り出すために形態素解析を用いる方法が一般的である．

本研究では，Web 文書を対象とする．Web 文書は文字列部とタグ部から多重に構成され, 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 Html 言語により記述する．表や図, リストといった構造をどのように扱うかが問題となる．

そして, ニュース記事や公的な文書と違い, Web 文書は文法的な誤りや造語を含むことがあるため, 辞書を利用した形態素解析を適応することができなくなる．これらのことから, どのような単位で文書の内容を抽出するか, Web 文書ベクトルの類似度をどのように定義するのかが問題となる．

1.3.2　複数文書要約

Web 文書集合ではそれぞれの文書で類似の情報が論じられていたり, あるいはまったく別の情報について論じている場合がある．そして, 類似な内容の Web 文書集合の複数文書要約は, 異なる内容同士の Web 文書集合の要約よりも容易に要約を生成することが可能であるし, 容易に要約から内容を把握することが可能であると考えられる．そこで,Web 文書をクラスタリングすることで類似な内容のグループ化を行うことを考える．

クラスタリング (Clustering) はオブジェクト集合へのグループ化手法で, 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 [13]．つまり, クラスリング技法は”類似性”の定義とその実行方法に依存して, 隠れたパターンをどれだけ見出せるかを競い合っているといえる．本研究では，Web 文書を対象としてクラスタリングをする．このとき Web 文書の内容をどのような単位で抽出するかや, Web 文書ベクトルの類似度をどのように定義するのかが問題となる．

そして, 類似した内容を持つ Web 文書集合を得ることができたと仮定する．この Web 文書集合から複数文書要約を得るために，Web 文書集合の内容をどのような単位で抽出して計算機で扱い，どのような方法で要約を生成するかが問題となってくる．

一見するだけで Web 文書集合全体の内容を把握することのできる要約が望ましく, Web 文書集合全体を把握できる要約と, より詳細に個々の文書の内容を把握できるような要約が最も望ましい.

1.3.3　評価

Web 文書集合全体の内容を把握できる要約と, 個々の文書の内容を把握できる要約が両立できたと仮定する. このとき, 利用者は一般的に全体の内容から, より利用者の要求する情報と類似している個々の文書の要約へと読み進めていく.

利用者がより読み進めやすくなる要約であるかを評価するために, 利用者の読解のしやすさといった尺度を定量的に評価することのできる内的評価尺度が必要である.

1.3.4　Web 情報検索結果への適用

これまでの Web 情報検索では問合せに対する検索結果は, 問合せに関連する Web 文書の URL と問合せ語を含む箇所の文章をリストにして表示した. 多くの利用者はこのリストの先頭からブラウズすることで要求する情報を探索すが, 問合せ語を含む箇所の文章からだけでは目的の情報を含む Web 文書を探すことは非常に困難である. Web 情報検索において検索結果をより素早く, 容易に内容を把握することのできる自動要約手法が必要である.

1.4　問題解決に向けてのアイデア

本研究では，以上の問題について以下の構成で論じる.

第 2 章では，類似した内容の Web 文書を同じグループにまとめる手法を提案する. ここでは, 同じリンク先を有する割合が高いほど Web ページ内容が類似しているという考えに基づき，ハイパーリンクの共起性と単語の分布の類似性を考慮した Web 文書クラスタリング手法を提案する. なお，これは IEEE Pacific Rim Conference on Communications, Computers and Signal processing (PACRIM' 05) などにおいて発表した.

第 3 章では，Web 文書集合の自動要約手法を提案する. 類似した内容を持つ Web 文書集合を対象として，html タグなどの構造を考慮して Web 文書から意味単位を抽出し，その間の階層構造を検出する. これにより詳細な内容を求めるならば下位のノードから, 全体の内容を把握するならば上位ノードから文書集合の内容を表現

できる要約を生成できたことを示す．これは筆者が International Association for Development of the Information Society Applied Computing (IADIS-AC) などにおいて発表した．

第 4 章では，第 3 章で提案した階層構造を用いた要約を定量的に評価する手法を提案する．ここでは, 階層のノードの可読性, 階層の可読性, 読解という 3 つ視点から評価する手法を実験的に検証している．これは筆者が ICDT Workshop on Emerging Research Opportunities in Web Data Management(EROW) などにおいて発表した．

第 5 章では，Web 情報検索において検索結果を階層的要約を用いることで, 検索結果から利用者の求める内容を持つ Web 文書への URL を探すために効果的に働くことを示す．これはデータ工学ワークショップ (DEWS07) において発表した．

第 6 章では，本論文をまとめ，また本論文で扱えなかった課題について言及する．

1.5　発表論文

1. 高橋功，三浦孝夫，塩谷勇: ハイパーリンクの共起性を用いたクラスタリング手法, データ工学ワークショップ (DEWS), 電子情報通信学会データ工学研究会，2003,
   Web ページのようなハイパーリンク構造を持つ文書を取り扱う場合，そのハイパーリンクの階層構造が深くなるほど文書の内容は焦点が絞られてくると考えられる。内容の焦点がしぼられているページへのハイパーリンクが共起しているページは内容が酷似しているという考えにもとづき，本論文ではハイパーリンクの共起性と深さ，ベクトル空間モデルにおける類似度を考慮したクラスタリング手法を用い，その手法の有効性を評価する．

2. Takahashi,K.，Miura,T.，Shioya,I.: Clustering Web Documents Based on Correlation of Hyperlinks (Extended Abstract), International Special Workshop on Databases For Next Generation Researchers in Memoriam of Prof. Kambayashi (SWOD2005), pp.20-23, 2005,
   同じリンク先を有する割合が高いほど Web ページ内容が類似しているという考えにもとづき，ハイパーリンクの共起性と深さ，ベクトル空間モデルにおける類似度を考慮したクラスタリング手法を用い，その手法の有効性を評価する．

3. Takahashi,K.，Miura,T.，Shioya,I.: Combination Clustering for Web Correlation, IEEE Pacific Rim Conference on Communications, Computers and Signal processing (PACRIM' 05), pp.434 - 437, 2005,

同じリンク先を有する割合が高いほど Web ページ内容が類似しているという考えにもとづき，ハイパーリンクの共起性と深さ，ベクトル空間モデルにおける類似度を考慮したクラスタリング手法を用い，その手法の有効性を評価する.

4. Takahashi,K.，Miura,T.，Shioya,I.: Summarizing Web Pages Hierarchically, International Association for Development of the Information Society Applied Computing (IADIS-AC), pp.612-617, 2006,
本稿では，階層構造表現による Web 文書集合の要約手法を提案する．Web 文書から意味単位を抽出し，その間の階層構造を検出することで，文書集合の内容を表現できることを示す.

5. Takahashi,K.，Miura,T.，Shioya,I.: Hierarchical Summarizing and Evaluating for Web Pages, ICDT Workshop on Emerging Research Opportunities in Web Data Management(EROW), 2007,
階層的 Web 文書集合の要約手法を定量的に評価する手法を提案する．Web 文書から意味単位を抽出し，その間の階層構造を検出することで文書集合の内容を表現できることを示し, 階層のノードの可読性, 階層の可読性, 読解という 3 つ視点から評価する.

6. 高橋功，三浦孝夫，塩谷勇: 階層的要約を用いた Web 文書集合への問合せ, データ工学ワークショップ (DEWS), 電子情報通信学会データ工学研究会, 2007,
階層的要約を用いた Web 情報検索手法の提案.

第2章　ハイパーリンクの共起性を用いたクラスタリング手法

2.1　前書き

これまで数多くの Web クラスタリング手法が提案されている [5]．その目的は様々であり,Web 上でのクラスタリング, Web ログ・セッション入手,Web セッションクラスタリング, Web コミュニティ検出 (Authority や Hub),Web 文書クラスタリング, 検索エンジン結果の集約など多岐に渡っている．

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 [13]．つまり, クラスリング技法は”類似性”の定義とその実行方法に依存して, 隠れたパターンをどれだけ見出せるかを競い合っているといえる．これまで知られたクラスタリング技法は, 大きく分割方式 (オブジェクト集合を分割し, ある基準で評価する), 階層化 (オブジェクト集合をある基準で階層的に分解する), 密度に基づく手法 (結合度・密度関数による評価) などに大別され, 類似性の定義は距離の定義として考察されることが多い．

Web **クラスタリング** は Web 情報の利用度の向上,Web 探索経路の短縮, 利用者要求への対応・応答の向上, 検索性能 (Recall/Precision) の向上, 内容提示の質的向上, 利用者動作の意図の理解, データ表現標準への対応,Web 情報構造の改善などを目的としたものであり, 上述クラスタリングと変わるものではない．Web 文書をクラスタリングすることによって, 相互に関連する Web ページのグループを検出し Web コミュニティを得るものや, 類似した内容をもつページ集合にまとめ, Web 検索の性能向上・検索結果の質的向上を図ることができる．

対象とするデータは,Web 文書 (ページ内容) だけでなくて, 利用者動作を記述するログ情報も含まれる．このとき,Web 文書集合をグループ化するための特徴値として何を用いればよいのであろうか？これまで, 検索エンジンの結果を解析するという立場からの提案は多い．検索エンジンへの要求に何らかの共通性があり, これを手がかりに”強く関連した文書は同じ質問に反応しがち” という特性から, 利用者意図の表現を探る多くのクラスタリング手法が提案されている．例えば

Scatter/Gather[6], STC[35], Carrot2[26] などが代表的である.

本稿では, 情報検索手法を用いてカテゴリカルクラスタリングを Web 文書に適用する手法を提案する. 本稿で扱うデータのほとんどはカテゴリカル (”Gold, Silver, Blonde” など) であり, 数値データではない. このため距離や順序概念が考えにくく, 過去に提案された多くのアプローチがなじみにくいの理由のひとつになっている. 本稿ではハイパーリンクおよび Web 文書内に生じる索引語の分布頻度や共起性を分析し, また関連性によるグラフ構造分析を用いたクラスタリング手法を提案する.

第 2 章は Web 文書のクラスタリング手法を, 続く 3 章では提案手法を示す. 実際の Web 文書データを用いた実験結果を 第 4 章で述べる.

2.2　Web 文書クラスタリング

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. 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される.

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 [13]．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, 文書間の非類似度を対応するベクトル間の余弦 (cosine) 値を用いて記述する．文書に生じる各語については，語幹を抽出し (stemming) 不要語 (stop word) を除去するなどの事前操作により，文書の特定を効率よく行う必要がある．しかし，文書数が増えるにつれ高次元化していくという問題点があるため，特徴的な語 (索引語 index term) を選び出して次元数を限定するなどの工夫が必要である.

一般の文書と比較して，Web 文書に際立つ特徴について配慮せねばならない. 例えば，少ない語だけで特徴的な Web 文書[1]や，空間配置，CSS/XML, 色彩, フォント, マルチメディアといった Web 文書の特殊性を吸収する必要があろう．これらの特性のうち**ハイパーリンク** (他 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．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．頂点が Web ページ, 辺がハイパーリンクに相当し, 参照の数はトピックの注目度に対応する. ただ良く知られているように，参照/非参照の頻度 (構造情報) によるクラスタリングを行うと，巨大サイトへの参照のみでクラスタが形成されることが多い．つまり，少数の巨大・準巨大クラスタ

[1]例えば“ 飛べ赤星！”とだけ記述された Web 文書がある.

と多数の泡沫クラスタが生成されやすく，実質的に特徴的なクラスタ集合を得にくいという問題点がある.

2.3　提案手法

本稿で提案する Web 文書クラスタリング手法は, Web 文書の文書特性とハイパーリンク構造を反映したものであり, 直感的で単純な方法である. HITS アルゴリズム [14] の解析等でよく知られているように, オーソリティ (authority) とは非参照 Web 文書 (ページ) のうち特定のトピックにおける的確な情報を持つと承認されているものを意味する. このため同じ authority を参照する Web 文書は同一トピックに言及している可能性が高く, 当該トピックにに関して類似していると考えてよい.

この考え方を用いて, 本稿では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. 同時に，索引語により Web 文書をベクトル化し (当該ベクトル空間上で) 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) を生成する. この 2 つの結果を”重ね合わせる”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.

2.3.1　Link クラスタリング

はじめに Link クラスタリングを定義する. このため有向グラフ G を用いた形式化を行う. 有限集合 $N = \{a_1, .., a_n\}$ および $E \subseteq N \times N$ が与えられたとき $G = \langle N, E \rangle$ を有向グラフという。ただし，N の要素を頂点，$(a, b) \in E$ を始点を a, 終点を b とする辺という. G では，始点および終点それぞれが同じである辺は唯一つしかなく，またサイクル (a, a) は無いとする. 頂点 a から出る辺の集合 $From(a) = \{b \in N \mid (a\, b) \in E\}$ 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 $To(b) = \{a \in N \mid (a\, b) \in E\}$ を入辺集合 (要素数を入次数) という. 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. 参照数は入次数に対応しており，トピックの注目度に対応する. 一般に $|From(a)| \gg 0$ となる a はハブ (hub), $|To(b)| \gg 0$ となる b はオーソリティに対応する. なお，本稿では出次数が 0 の頂点は (トピックが独立しており) 除外する.

Link クラスタリングの手順を示す. 実際の手順は完全連結法を用いた，階層型クラスタリングによる. 2 つの頂点 a_i, a_j に対して，a_i と a_j の値 d_{ij} を次式で与

える．

$$d_{ij} = 1 - \frac{2\,|From(a_i) \cap From(a_j)|}{|From(a_i)| + |From(a_j)|} \tag{2.1}$$

d_{ij} は a_i, a_j の双方から参照されている頂点数 (共起数) の割合を用いて定義されていることに注意したい．実際，この距離は頂点の辺の数に依存せず同じ終点への辺の割合と対応している．d_{ij} は，共起の割合が大きいほど 0.0 に，少ないほど 1.0 に近づく．このため，(共起性に関する) 非類似度と呼ぶ．$n \times n$ の行列 $D = ((d_{ij}))$ を非類似度行列と呼ぶ．定義から D は対称行列である．

非類似度行列 D を用いてクラスタリングを行う [13]．このクラスタリング手法を *Link* **クラスタリング**，その結果の各クラスタを ”Link クラスタ” と呼ぶ. 以下では，クラスタのうち要素数が閾値 θ 以下のものを破棄する．実際，頂点の数少ないクラスタは内容を判断することができず，誤った判断を招く可能性が高い。

前節で指摘したように，Link クラスタリングを実行するとき，(検索エンジンサイトなどの) 巨大なハブ・オーソリティだけからなるクラスタが検出され，効果的で意味のある結果が得られないことが多い．本研究では，Zipf の法則を用いて，効果的なハイパーリンクだけを抽出する[2].

例 1 ここでは *Link* クラスタリングの例を示す．図 *2.1* のように *6* 個の頂点 $a_1 \cdots a_6$ があるとき *Zipf* の法則によりいくつかの頂点を破棄する．この例では閾値 $\theta = 1$ としている．頂点 $a_1, .., a_6$ の出次数はそれぞれ *2,2,1,2,5,1* である．*Zipf* の法則より，$f_k = \frac{\sqrt{8\cdot 2+1}-1}{=} 1.56$ であり，a_5 *(*ハブとみなすことができる*)* が除去される．これ以外の頂点の非類似度行列を D とする。これは以下のように求められる．*Link*

[2]Zipf の法則とは高次元化を抑制するための次元縮小技法の一つで, 高頻度の単語で成り立つ Zipf の第 1 法則と, 低頻度の単語で成り立つ Zipf の第 2 法則がある．低頻度の単語をどの程度削除するかの基準として, まず「中程度の頻度」を決める必要がある．頻度 1 の単語数を F_1 とすると,2 つの法則を同時に満たす中程度の単語頻度 f_k は, 以下の式で求められる．

$$f_k = \frac{\sqrt{8F_1 + 1} - 1}{2} \tag{2.2}$$

ここで得られた出現頻度 f_k が索引語の頻度順位において中間地点であることを仮定すれば, 以下の手順で索引語数を決定できる．

1. 出現頻度 f_k を持つすべての語を索引語とする
2. 第 1 順位から $f_k - 1$ 個の頻度を持つ語までのすべてを索引語とする．全部で K 個の語があるとする
3. $f_k + 1$ 以下の出現頻度の語のうち, 上位 K 個を索引語とする

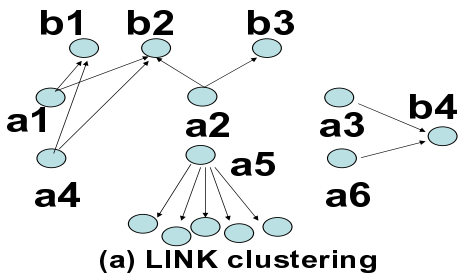


図 2.1: 例 1:Link クラスタ

クラスタリングを行う.

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 2 & 3 & 4 & 6 \end{array} \\ \begin{pmatrix} 0 & 0.5 & 1 & 0 & 1 \\ 0.5 & 0 & 1 & 0.5 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0.5 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

このクラスタリングによりふたつのクラスタ $A_1 = \{a_1, a_2, a_4\}$, $A_2 = \{a_3, a_6\}$ が生成できる. 頂点 a_5 は削除されたため, 孤立点 *(1* 点だけからなるクラスタ*)* とみなして削除する.

2.3.2　VSM クラスタリング

Web 文書クラスタリングでは Web 文書からアンカータグを取り除いたプレーンテキストを対象にする. ここでは大量の文書を扱うために, ページごとの単語の出現頻度を扱うと，Web 文書の文字数による偏りが生じる可能性がある. このため，本稿では各 Web 文書の重みを 0(未出現), 1(出現) の 2 値で表現したベクトル $\vec{p_i}$ を用いる.

m 個の Web 文書集合 $P = \{\vec{p_1}, \cdots, \vec{p_m}\}$ に対し $\vec{p_i}$ と $\vec{p_j}$ の非類似度 d_{ij} を次で定義する.

$$d_{ij} = 1 - \frac{(\vec{p_i} \cdot \vec{p_j})}{|\vec{p_i}\cdot| \, |\vec{p_j}|} \tag{2.3}$$

この d_{ij} から非類似度行列 $D = ((d_{ij}))$ を定義し，先と同様に完全連結法による階層型クラスタリングを行う. この手法を *VSM* **クラスタリング**，その結果のクラスタを ”VSM クラスタ” と呼ぶ.

例 2 *VSM* クラスタリングの例を示す．*6* 個の *Web* 文書 $a_1, \cdots, a_6$ に対応して文書ベクトルが図 *2.2* で与えられているとする．このとき, *VSM* クラスタリングを行う, ここで閾値 $\theta = 1$ とする.

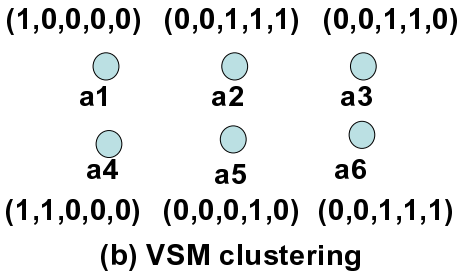


図 2.2: 例 2:VSM クラスタ

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0.5 & 1 & 1 \\ 1 & 0 & 0.67 & 1 & 0.67 & 0.67 \\ 1 & 0.67 & 0 & 1 & 0.75 & 0.75 \\ 0.4 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0.67 & 0.75 & 1 & 0 & 0.75 \\ 1 & 0.67 & 0.751 & 1 & 0.75 & 0 \end{pmatrix} \end{array}$$

各ベクトルの非類似度を上記の行列 D で表し，階層型クラスタリングを行う. この結果，*2* つのクラスタ $B_1 = \{a_1, a_4\}$, $B_2 = \{a_2, a_3, a_5, a_6\}$ が生成される.

2.3.3　クラスタの重ね合わせ

文書特性とハイパーリンク構造の特性の双方を備え, さらに Link クラスタ結果を効果的に分割するために，2 つのクラスタリング結果を重ね合わせる方法を考える.

Link クラスタリングによる n 個のクラスタ $\mathcal{A} = \{A_1 \cdots A_p\}$, VSM クラスタリングによる m 個のクラスタ $\mathcal{B} = \{B_1 \cdots B_q\}$ に対して，さらに A_0 と B_0 をそれぞれの手法で破棄された頂点の集まりとする．このとき, **クラスタの重ね合わせ** を次式で定義する.

$$C_{ij} = A_i \cap B_j \tag{2.4}$$

ただし C_{ij} の要素数が閾値 θ を下回れば破棄する．このとき C_{ij} は最大で $p \times q$ 個得られる．

クラスタの重ね合わせのアルゴリズムは以下の通りである．

1. $C_{ij} = \emptyset$ とする，$i = 1, .., p, j = 1, .., q$
2. 各 $a_i \in N$ に対して (3)-(5) を行う
3. $a_i \in A_k$ となる k を求める
4. $a_i \in B_{k'}$ となる k' を求める
5. a_i にクラスタ $C_{kk'}$ を割り当てる

例 3 重ね合わせたクラスタの例を示す．図 *2.3* 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．閾値 $\theta = 1$ としたとき，*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．クラスタ $C_{12} = \{a_2\}$ と $C_{02} = \{a_5\}$ は閾値以下の頂点数のため破棄される．

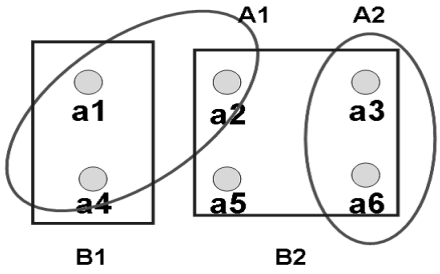


図 2.3: 例 3:重ねたクラスタ

2.4　実験

2.4.1　実験環境

本研究では,実験データとして,NTCIR-3 Web 文書データ[3] を使用する．NTCIR-3 は 2001 年 8 月 29 日から 2001 年 11 月 12 日の間に収集した .jp ドメインの拡張

[3]http://research.nii.ac.jp/ntcir/

子 html と text データを集めたテスト・コレクションである. 100Gbyte を越えるデータを含む NW100G-01,10Gbyte-01 のデータを含む NW10G-01 の 2 つがある．このテスト・コレクションはとくに Web 文書を対象とした検索, 分類, 情報抽出などの情報活用システムの比較評価, ならびに,Web テストコレクションの構築を目的としている．NW100G-01 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 件の日本語の Web サイトのデータを用いる.

本稿では，Zipf の法則に基づき頂点となる 3234 件を抽出する. またその頂点が持つ 2,429,984 個の辺と 3,825,293 個の単語に Zipf の法則を適用し 1285 個の辺,449 個の単語を抽出する．これより Link クラスタ及び VSM クラスタを求め, 2 つのクラスタから重なり合うクラスタをつくる. またクラスタの要素数が閾値 θ が 5 以下ならば破棄する.

2.4.2　実験 (1):VSM クラスタ 結果と考察

VSM クラスタを用いて得られた代表的なものを表2.1 に示す. 表2.1 より，VSM1,3,4 クラスタは個人のページが多く, VSM2,5 クラスタは企業, 大学等の公式ページが多く集まっていることがわかる．VSM によるクラスタリングでは類似した単語集合を持つページ同士がクラスタになりやすいため, ページ内で口語体, あるいは文語体であるという差が結果に現れている.

クラスタ名	
VSM1	個人サイト 203 個 大学, 企業　63 個
VSM2	個人サイト 40 個 大学, 企業　113 個
VSM3	個人サイト 127 個 大学, 企業　111 個 (図書館, 書籍に関するもの 25 個)
VSM4	個人サイト 109 個 (写真, イラストに関するもの 51 個) 大学, 企業　54 個
VSM5	個人サイト 72 個 大学, 企業 165 個

表 2.1: VSM の代表的なクラスタの内容

2.4.3　実験 (2):Link クラスタ 結果と考察

表 2.2 に Link クラスタリングの結果を示す。ここで参照しているハイパーリンクからそれぞれのクラスタは以下のようなトピックが含まれていると考えられる.

> Link1 クラスタは「大学, 地域の話題, 書籍」に関するページへの参照が多く, 頂点は「個人サイト (日記, 旅行)43 個, 大学 7 個, 企業 10 個」を持つ.
> Link2 クラスタは「OS, セキュリティ」で頂点は「個人サイト (日記, 無料掲示板)30 個, 大学の研究室 13 個・企業 23 個」を持つ.
> Link3 クラスタは「プロバイダ」で, 頂点は「個人サイト (日記)24 個, 大学 2 個」. いずれのクラスタも参照がトピックと対応しているとするならば, トピックと頂点は対応している見ることができる.

しかしトピックが複数にわたるクラスタでは全体としてまとまりが悪く，一見して集約することは容易ではない.

2.4.4　実験 (3):重ね合わせたクラスタ 結果と考察

重ね合わせたクラスタの結果を表 2.3 に示す。ここでは 7 つのクラスタを得た.

Link1 クラスタは 3 つのクラスタに分割されている. すなわち，(VSM1 クラスタと重なった) 個人のページ（日記・旅行）のクラスタ，(VSM4 クラスタと重なった) イラストをメインとする個人ページのクラスタ，(VSM5 クラスタと重なった) 大学・NTT や公務員に関するページのクラスタである.

各々は各トピックと対応するクラスタに分割されている. 実際 (Link1 クラスタの) トピック「地域の話題, 書籍, 大学」は VSM1 では個人サイト,VSM4 ではイラストをメインとする個人サイト, VSM5 では大学のクラスタであった. 重ね合わせたクラスタがそれぞれのトピックに対応しているクラスタに分割されている.

Link2 クラスタでは (VSM1 クラスタと重なった) 大学の研究室（電気, 土木）と個人サイトのクラスタ，(VSM2 クラスタと重なった) 大学の研究室（化学）, 個人サイトのクラスタ，(VSM3 クラスタと重なった) 大学の研究室（電気, 医）と無料掲示板のクラスタという 3 つのクラスタに分割された. Link2 クラスタのトピック「OS, セキュリティ」には，いずれの重ねたクラスタにおいても大学の研究室サイトが含まれる．VSM1,3 クラスタは個人サイトが多いクラスタであるため重ねたクラスタでも個人サイトをみることができている.

Link3 クラスタは，VSM1 クラスタと重なり個人サイトのみのクラスタが得られた．反面，大学の研究室などのページが除去された. Link3 も VSM1 も「個人サイト」というトピックのため，それ以外のページが除去されている.

クラスタ名	Link1	Link2	Link3
クラスタ内の頻出辺	jp.freebsd.org	sun.com	try-net.or.jp
	aozora.gr.jp	linux.org	freeweb.ne.jp
	washingtonpost.com	sgi.com	so-net.ne.jp/postpet
	un.org	netbsd.org	webring.ne.jp
	tsukuba.ac.jp	linuxhq.com	ixla.com
	suzuki.co.jp	hp.com	alles.or.jp/ queen
	shogakukan.co.jp	freebsd.org	geocities.co.jp
	sanyo.co.jp	w3.org/Daemon	altan.hr/snow
	pref.niigata.jp	tripod.com	6.big.or.jp/ neon
	pref.fukuoka.jp	specbench.org	ushikai.com
	pref.aomori.jp	sleepycat.com	azaq.net
	pref.akita.jp	sequent.com	hp.bird.to
	nytimes.com	sco.com	7.big.or.jp/ jawa
	nikkeibp.co.jp	rsa.com	eva.hi-ho.ne.jp/takeuchi
	nasda.go.jp	openbsd.org	w3.org
	mycom.co.jp/career	nikkansports.com	odn.ne.jp
	monbu.go.jp	netcraft.com/Survey	nifty.com
	mext.go.jp	ncr.com	jra.go.jp
	maruzen.co.jp	my.host.com	zakzak.co.jp
	kyushu-u.ac.jp	multihost.com	winamp.com

表 2.2: Link の代表的なクラスタの内容

2.4.5　議論

これまでの結果の考察から，クラスタ結果を重ね合わせることで，より明確で的確なクラスタへと分割することができたと言える．類似したクラスタへと分割できた．しかし 各クラスタの頂点の数とクラスタ数の表 2.4 によると, クラスタを重ね合わせによりクラスタ要素数が減少し，破棄クラスタが増加している．このうち要素数 1 のクラスタは 1121 ある．これらはリンクの保持数もしくは単語の保持数が突出している Web 文書が多いため, 雑音であると考えられる．

要素数 2 ないし 4 のクラスタは，Link クラスタリングにより少ない要素数であったクラスタを更に分割したものが大半である．実験では Link1 クラスタと VSM2 クラスタは共に大学というトピックであり, 重ねたクラスタでは要素数が 3 であった．この 3 つにクラスタはいずれも大学に関するページである．同様に Link2 クラスタと VSM3 クラスタを重ねたときも要素数が 3 であり, 大学の研究室のページ

	VSM1	VSM2	VSM3	VSM4	VSM5
Link1	個人サイト			個人サイト	企業
Link2	大学 個人サイト	大学 (化学)	大学 （電気, 医）		
Link3	個人サイト				

表 2.3: 重ね合わせたクラスタの内容

が 3 つであった．これらは閾値以下だったため破棄した．しかしこれらは正しくクラスタに分割できており，真に雑音かどうかは即断できない．重ねたクラスタにおいては閾値を下げる等の判断が必要である．

頂点の数	Link	VSM	重ね
1-5	171	54	1747
6-10	118	19	23
11-30	84	27	6
31-50	6	2	6
51-70	2	0	1
71-90	-	2	-
91-110	-	0	-
111-130	-	0	-
131-150	-	0	-
151-170	-	3	-
171-190	-	0	-
191-210	-	2	-

表 2.4: 頂点数とクラスタ数の対応表

2.5　関連研究

クラスタリングは統計, パターン認識, データベース, データマイニングといった分野で研究されている. 類似しているオブジェクト同士をまとめていきデータ集合を分割するのに利用される. 類似度を分析するのに通常ユークリッド距離などを用いるためオブジェクトは数値属性で表現されているのが一般的であるが, カテゴリカル属性を扱う手法も研究されている. Web サイトなどの半構造データにおいてはカテゴリ属性を扱う必要がある. そこでリンク概念を用いたカテゴリカルクラスタリン

グ手法として ROCK(Robust Clustering using linKs)[11] と CACTUS(CAtegorical ClusTering Using Summaries)[8] がある．ROCK とは 2 つの対象で共起しているリンクが Jaccard 係数などを用いて閾値以上であるとき対象は類似しているとする手法である．2 つの対象だけでなく, それらの近隣の影響を考慮することで少数の例外的な対象の影響を受けにくい特徴がある．

$$\sum_{i=1}^{k} n_i \times \sum_{x_p, x_r \in C_i} \frac{link(x_p, x_r)}{n_i}$$

リンクの類似度の大きな対象同士を同じクラスタに分類する目的でこの評価関数を最大化する.link(x_p,x_r) は対象 x_p と x_r の間のリンク数を表す．

これに対して, CACTUS は類似性に基づく近隣関係ではなく対象集合中の属性値の共起性に基づいた連結関係を用いる. 共起性の強い属性についての要約情報があればデータ全体の情報がなくてもクラスタを抽出できる性質を持つため記憶容量を削減できる．

2.6　結び

本稿では, ハイパーリンクの共起性とベクトル空間モデルを用いたクラスタを重ね合わせる手法を提案した. 2 つのクラスタリングの結果を重ね合わせることにより類似したクラスタを生成することができた．しかし，重ね合わせにより細分化されたクラスタについては，別途評価を行う必要があり，今後の問題として残されている．

第3章　階層的Web文書集合の要約

3.1　前書き

近年，インターネットの普及により World Wide Web(WWW) の世界は急激に拡大し膨大なテキスト情報をもたらした. そのため，個別の Web 文書要約だけでなく Web 文書集合全体から内容をすばやく把握する方法，即ち Web 文書集合の自動要約に対するニーズが高まっている.

この問題に取り組むためのアプローチとして *Web* **クラスタリング**が挙げられる. これまで数多くの Web クラスタリング手法が提案されているが [5]. その目的は様々であり,Web 上でのクラスタリング, Web ログ・セッション入手,Web セッションクラスタリング, Web コミュニティ検出 (*Authority* や *Hub*),Web 文書クラスタリング, 検索エンジン結果の集約など多岐に渡っている.

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 [13]. つまり, クラスリング技法は”類似性”の定義とその実行方法に依存して, 隠れたパターンをどれだけ見出せるかを競い合っているといえる. これまで知られたクラスタリング技法は, 大きく分割方式 (オブジェクト集合を分割し, ある基準で評価する), 階層化 (オブジェクト集合をある基準で階層的に分解する), 密度に基づく手法 (結合度・密度関数による評価) などに大別され, 類似性の定義は距離の定義として考察されることが多い.

ほとんどの Web 文書が扱うデータはカテゴリカル (例えば *ComputerScience* ,*Biology,athematics* など）であり, 数値データではない. このため距離や順序概念が考えにくく, 過去に提案された多くのアプローチがなじみにくいの理由のひとつになっている.

要約はクラスタリングとは少々異なり，情報ソースにおける重要な内容を簡約な形で提示することである. ここでは一見するだけで内容が把握できることが望ましい. 人手による要約は，大意と要旨の二つに区別することができる [24]. 大意とは原文の表現をできるだけ用いて順序を変えずにまとめたものであり，要旨とは原文の主題や結びに焦点を絞り原文の表現形式にとらわれずにまとめたものである.

これに対応して，自動要約の研究では**抜粋** *(Extraction)* と**抽象化** *(Abstraction)* という手法が提案され研究されている．抜粋とは対象文書 (あるいは集合) から文章を抜き出す手法を意味する. 例えば，重要文抽出法では，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である．新たに文章を作る必要がなくまた言語知識をほとんど必要としないため実現が容易である．反面，代名詞や接続詞などの関連が整合性を有さないことがあり，語句間の関連性を壊す可能性があること，箇条書きや表などの構造情報の取り扱いに統一性がなく，この結果長大な題材をうまく取り扱えない．これに対し，対象文書 (集合) の内容を把握し，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を抽象化と呼ぶ．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．過去には，新聞記事のように出現内容を規定できるものや，教科書などのように語句が整形された文書に限定して適用されている．

これらの技術でテキストを対象として要約するため，リンクやタグなどの半構造を有する Web 文書に適用することは容易でない．そこで, 本稿では Web 文書クラスタリングと Web 文書要約という二つの考えに基づき, 階層型クラスタリングを用いた Web 文書集合の要約手法を提案する．

第 2 章では自動要約について現状を述べ，第 3 章では階層的抽象化手法について論じる．第 4 章, 第 5 章で提案手法を述べ，第 6 章で実験結果を挙げ本手法の有用性を示し，第 7 章は結びである．

3.2　自動要約技術

Web 文書に対する自動要約技術では，例えばハンドヘルドデバイス上の操作が提案されている [4]. 装置上の小さなディスプレイでブラウジングのために Web 文書を *Semantic Textual Unit* (STU) に分割し，その先頭一行目を出力する．深く読むには，先頭 3 行, 全文, と出力のレベルを変える．この基礎となる STU は，意味的まとまりをもつ文章単位であり，段落や画像の説明属性 (alt 部分) の値を意味している．このほか，**文脈**を考慮した Web 文書の要約手法も提案されている [7]. 文脈とは，要約の対象となる文書へのリンクを有する Web 文書集合であり，文脈内のリンク周辺の STU を抜粋し，これを要約とする．

これらは，いずれも単一文書を対象としている. 対象は同一著者であることから文章表現や語句使用に一貫性があり，機械的な解析になじむことによる. これまで述べた方法で複数文書に対することは自明では無い [15].

Web クラスタリングは相互に関連する Web ページ集合を得る手法であり, Web

の類似度の概念はある種の**距離**[1]. で導出される. こうして形成されたクラスタに対しては異なった要約手法が提案される. **クラスタのラベル付け**はクラスタの意味する内容をラベルが表現することから, Web ページの内容の要約と対応していると考えることができる. 多くの場合, 頻出語をラベルとして利用するが [15], サーチ・エンジンから得られたページ集合のように強い類似性もつ集合の場合, 頻出語からは各クラスタを区別することができない [16, 17]. 頻出語に代わるアプローチとして重要語を用いることですばやく内容を把握することができる. さらに一つの語ではなく**語の並び**を用いることができればより内容理解を容易にする. 重要語を抽出する手法として KeyGraph[21] や, 語の並びを考慮した SuffixTree[35], そしてこの二つを組合せからラベルを生成する手法も提案されているが, 結果が時制に依存しているとされる [18].

3.3　階層的抽象化手法

抽象化要約では，対象の重要部を抜き出し，その性質を把握するという二つのステップから構成される. ここで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. 重要部の判定は，単語・語句の場合は出現頻度や共起性等で，文章の場合は (概念クラスタリング等のように) 文章集合の重心からの距離を用いて表される. 従って，文書内容の性質の把握のためには，重要な単語・語句間の関連抽出，あるいは文書クラスタリング手法とその解釈が重要な役割を持つ. 例えば, 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*[21] 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.

Web に適応する場合, 主な問題の 1 つが莫大な量の Web ページからの適切なページ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合せた Web 文書クラスタリングの有用性を示した [27]. そ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

[1] ここで**距離**とは距離の公理を満たすものと定義される.

きる．本稿では，この Web 文書集合に対して階層クラスタリングを適用し，文章の階層表現により要約する．各クラスタの重心を計算することでラベル付けし，クラスタの階層関連を抽象度と対応させる．

3.4　組合せクラスタリング

本章と次章で階層表現を用いた Web 文書集合の要約手法を提案する．最初に, 相互に類似した Web ページ集合を得る手法について述べる．我々はこのために**組合せクラスタリング**を示す．次に各クラスタの解釈の方法について述べる．この時, 各クラスタをさらにクラスタリングを適用するという手法をとる．最後に Web 文書集合の解釈と対応した階層表現を得ることができる．

組合せクラスタリングは Web 文書の文書特性とハイパーリンク構造を反映したものであり, 直感的で単純な方法である．この手法は既に提案されているためここでは概要を示す [27].

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．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される．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 [13]．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, 文書間の非類似度を対応するベクトル間の余弦 (cosine) 値を用いて記述する．一般の文書と比較して，Web 文書に際立つ特徴について配慮せねばならない．一方, **ハイパーリンク** (他 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．

以上の考えに基づき, 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．同時に，索引語により Web 文書をベクトル化し (当該ベクトル空間上で) 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) を生成する．この 2 つの結果を”重ね合わせる”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 Link 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 3.1 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．頂点 $a_1, .., a_6$ の出次数はそれぞれ 2,2,1,2,5,1 である．そして, 頂点の非類似度行列を D の要素 d_{ij} は a_i, a_j から次式で定義される．

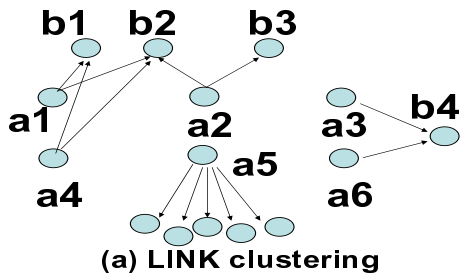


図 3.1: LINK Clustering

$$d_{ij} = 1 - \frac{2\,|From(a_i) \cap From(a_j)|}{|From(a_i)| + |From(a_j)|} \tag{3.1}$$

d_{ij} は a_i, a_j の双方から参照されている頂点数 (共起数) の割合を用いて定義されていることに注意したい. 次元縮小のために非常に小さい値の頂点は除去され, 5 つの頂点に対して次の**非類似行列** D を得る.

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 6 \end{array} \\ \begin{pmatrix} 0 & 0.5 & 1 & 0 & 1 \\ 0.5 & 0 & 1 & 0.5 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0.5 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

各頂点をそれぞれクラスタとみなし,d_{ij} が最大となる頂点 a_i, a_j を併合して一つのクラスタにする. もしくは, クラスタ C_1, C_2 では $a_i \in C_1$ と $a_j \in C_2$ で最小値となる d_{ij} をクラスタ間非類似度と定義し,d_{ij} が最大となるクラスタを併合する. これを一つのクラスタになるまで繰り返す. このプロセスを *LINK* **クラスタリング**と呼び, 得られるクラスタを *LINK* **クラスタ**と呼ぶ. LINK クラスタリングよりふたつのクラスタ $A_1 = \{a_1, a_2, a_4\}$, $A_2 = \{a_3, a_6\}$ が生成できる. 頂点 a_5 は孤立点 (1 点だけからなるクラスタ) とみなして削除する.

VSM クラスタリングの例を示す. 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 3.2 で与えられているとする. m 個の Web 文書 $P = \{\vec{p_1}, \cdots, \vec{p_m}\}$ が与えられた時, 二つのベクトル $\vec{p_i}$ $\vec{p_j}$ の**非類似度**は次式で定義される.

$$d_{ij} = 1 - \frac{(\vec{p_i} \cdot \vec{p_j})}{|\vec{p_i}\cdot|\,|\vec{p_j}|} \tag{3.2}$$

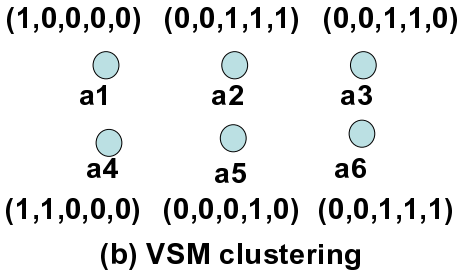


図 3.2: VSM Clustering

$m \times m$ の**非類似行列** $D = ((d_{ij}))$ を得て, complete linkage method による階層型クラスタリングを適応する手法を VSM clustering と呼び, 得られるクラスタを *VSM* **クラスタ**と呼ぶ. 図 3.2 の非類似行列 D は:

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0.5 & 1 & 1 \\ 1 & 0 & 0.67 & 1 & 0.67 & 0.67 \\ 1 & 0.67 & 0 & 1 & 0.75 & 0.75 \\ 0.4 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0.67 & 0.75 & 1 & 0 & 0.75 \\ 1 & 0.67 & 0.751 & 1 & 0.75 & 0 \end{pmatrix} \end{array}$$

この結果，2 つのクラスタ $B_1 = \{a_1, a_4\}$, $B_2 = \{a_2, a_3, a_5, a_6\}$ が生成される.

組合せクラスタリングの例を示す．図 3.3 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. 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.

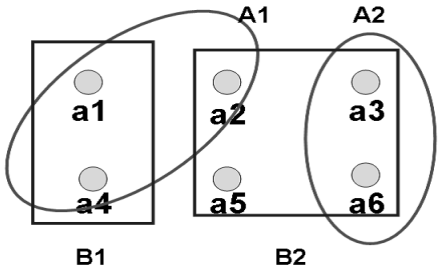


図 3.3: Combination Clusters

3.5　文書集合の階層クラスタリング

前章で我々はどのように Web 文書集合を得るかを述べた. 組合せクラスタリングから Web 文書集合を抽出できたと仮定し, それぞれの集合の解釈の手順を述べる. Web 文書は文字列部とタグ部から多重に構成されている. そこで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定義し, これを *Semantic Textual Unit* (STU) と呼ぶ. 本稿は 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て階層的要約を得る.

html ではタグで囲われた文書の一部を要素と呼び, 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, STU はタグを考慮して構成された (完結した) 文章であると言える. 本稿で対象とするタグは <P> <UL> <OL> <DL> <TITLE> <TABLE> <BLOCKQUOTE>である. 以下では，タグの入れ子構造とリンクを利用し STU 同士の関連を表現してクラスタ階層を形成できることを示す.

html ではタグが多段階の入れ子構造になることを許すため, 通常タグを同時に指定する場合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 (**文脈**を持つ) とみなし，リンク先 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．階層型クラスタ構成方式では，単連結法 (single linkage method) 完全連結法 (complete linkage method) 群平均法 (average linkage method) が知られる．前者は実データに適さないため，本稿では後者 2 つの構成方式を用いる．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 (center) とする．このとき各クラスタは重心によって表現する．最終的に，Web 文書集合から STU による階層表現を得る．

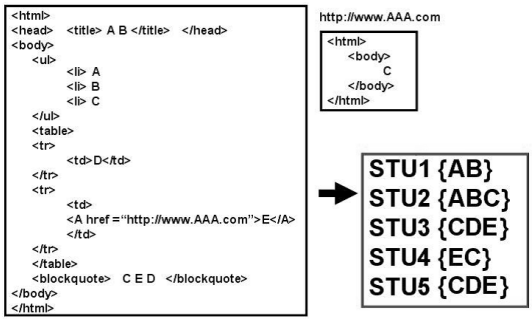


図 3.4: Taking STUs from Web Pages

本提案手法を用いて例を示す．図3.4のように,2つのWebページと単語 A,B,C,D,E に対して 5 つの STU が生成され. これらの STU を群平均法と完全連結法で階層型クラスタリングした結果を図 3.5 に示す. さらに，群平均法の結果を表によってまとめ直したものを図 3.6 で示す.

図 3.4 より， 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. 図 3.6 における C07 は図 3.5 の群平均法における STU1 と対応し，C06 が STU2，C05 が STU3，C09 が STU4，C08 が STU5 と対応している.C04 は類似度 [1.000] で C08 と C09 が合併し要素数が (2) 個となったクラスタを示す.

3.6　実験

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 件を用いて要約の対象とする．組

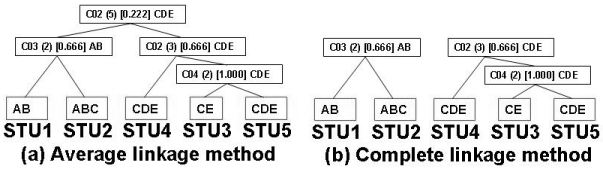


図 3.5: Hierarchy using STUs

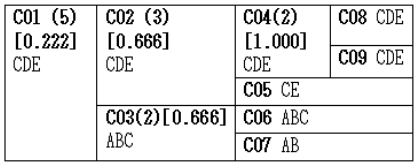


図 3.6: Average Linkage Method

合せクラスタリングによって 6 つのクラスタが得られ, この結果に対する人手による解釈と，クラスタの要素の URL を表 3.1 に示す.

クラスタ 1 を群平均法と完全連結法で階層型クラスタリングした結果を図 3.7 に示す. 以降, 各クラスタの要素数を (..) で, 合併したときの類似度を [..] で示す.

各クラスタの群平均法による結果を示す. 図 3.8 よりクラスタ 1 の要素である実験設備ページが C00/ C01/ C03/ C07/ C14/ C17/ C18 と対応し，訃報ページは C08，無料掲示板は C16 と対応していることが確認できる．旭川天文学部に関する STU が確認することができるが, クラスタ 1 の要素には旭川天文学部ではなく旭川医大が含まれていることからトピックドリフトが起こったと考えることができる .

同様に図 3.9 より, クラスタ 2 のミサワホーム / 大学：化学専攻 / 横浜線と対応する STU が確認できる. またミサワホーム内の相互リンクやタグの入れ子構造過多により，ミサワホームと対応するクラスタ階層が多く形成されていることが確認できる. しかし,C03 の下位のクラスタはすべてミサワホームに関するものにもかかわらず C03 の STU はミサワホームとは関連がない．これは C03 の要素数が多いことから平均ベクトルが相応しくない重心の STU が選択するほどずれてしまったと考えられる.

図 3.10 より，クラスタ 3 では土木工学科/ 北関西情報 /職業能力開発センター / 長岡技大 /フリーウェアのページと対応する STU が確認できる．図 3.11 より, クラスタ 4 で機械宇宙システム研究室 /法政大学 /岩手大学 /公務員情報のページと対応する STU を確認できる．図 3.12 より，クラスタ 5 でイラスト・写真に関する個人ページやアイドル写真集のページと対応する STU を確認できる. また, 個人ペー

表 3.1: Test Pages by 組合せクラスタリング

	クラスタの要素	解釈
1	momiji.i.ishikawa-nct.ac.jp（大学：通信研究室） hlweb.rri.kyoto-u.ac.jp/（大学：実験設備管理） cent-scorpio.asahikawa-med.ac.jp/（大学：旭川大医学部） ace.wisnet.ne.jp/（個人：全国訃報ネットワーク） cs.pst.jp/（個人：無料掲示板）	大学
2	www.misawa.co.jp/（ミサワホーム） fphy.hep.okayama-u.ac.jp/（大学：研究室） kanows1.ms.kagu.STU.ac.jp/（大学：品質管理） barato.sci.hokudai.ac.jp/（大学：化学専攻） hamasen.vis.ne.jp/（個人：横浜線）	大学
3	cive.gifu-u.ac.jp/（大学：土木工学科） cad7.nagaokaut.ac.jp/（大学：研究室） fmv-nt.winpal.co.jp/（個人：職業能力開発センター） likeonline.tripod.co.jp/（個人：フリーウェア） ke-tai.nkansai.ne.jp/（個人：北関西情報）	大学
4	horse.mes.titech.ac.jp/（大学：機械宇宙システム研究室） orion.mt.tama.hosei.ac.jp/大学：サーバー） jinsha.iwate-u.ac.jp/（大学：岩手大学人文社会科学部） great.pobox.ne.jp/accusation/akinbo/（個人：公務員情報）	大学
5	groovy_5.tripod.co.jp/（個人：デザイナー） moemoe.lowtech.ne.jp/（個人：イラスト） grandbleu.hoops.ne.jp/（個人：写真） bauhaus.co.jp/（個人：アイドル写真）	個人ページ
6	prize.crafteriaux.co.jp/（個人：クラフテリオ工作大賞） juujou.co.jp/100nin/2001/01super/（個人：子育て写真） furusatomura.pref.niigata.jp（個人：新潟ふるさと村） hironee.tripod.co.jp/（個人：日記） interface.tripod.co.jp/（個人：日記）	個人ページ

ジ内に福岡関連のページへのリンクがある為, 福岡に関するページへトピックドリフトが起きている.

各実験の結果より STU の内容が各クラスタの要素と対応していることが確認できる. 完全連結法の場合，互いに独立な要素を持つために併合されないクラスタが生じたが，それらからクラスタ内のトピックを報知的 (informative) に解釈することができる.(報知的とは原文の情報を極力落とさない要約を意味する). クラスタ 1・3・4・5 に群平均法を適応した場合，分割されても重心 STU が変わらないクラスタ階層を確認できる. これらのクラスタ階層は，要素数が非常に多い傾向にあるため，クラスタ 1・3・4・5 の主なトピックであると解釈できる. クラスタ 2・6 のように要素数の少ないクラスタ階層が多い場合, 分割で重心 STU が変わりやすくなる. これはクラスタに幅広いトピックが存在すると解釈することができ, クラスタ構造の解釈は複雑なものとなる.

以上のことから,STU の内容とそれを重心とするクラスタ階層は，Web 文書集合の内容を要約していることが確認できる. 同時に，クラスタ 1・2・5 ではトピックドリフトが発生していることも確認できた. 上位クラスタでこの影響を消去しているため大勢を変えるわけではないが，改善策としてリンク先の STU の重みの工夫等の方法が考えられる.

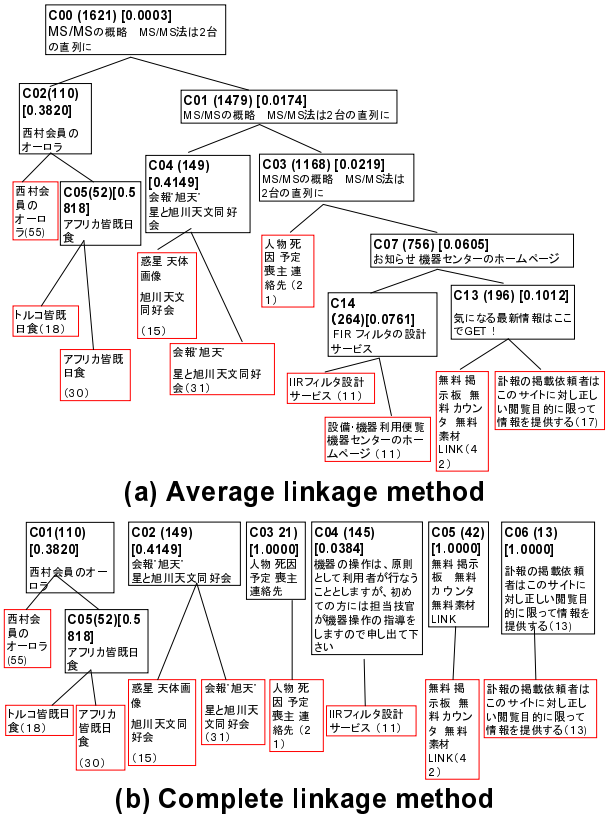


図 3.7: Hierarchical Summarization

3.7　結び

本稿では Web 文書集合を階層表現により要約する手法を提案した．Web 文書集合を STU に分割し，階層型クラスタリングを用いることで容易に実現可能なことを実験により示した．しかし，リンクやタグ構造を利用することでトピックドリフトが発生してしまうことも確認できた．今後の展開としては Web 文書集合の文脈となるページを利用した要約との比較が考えられる．

<table>
<tr><td rowspan="8">C00
(1620)
[0.0003]
MS/MSの概略
MS/MS法は2台の直列に</td><td rowspan="6">C01
(1479)
[0.0174]
MS/MSの概略
MS/MS法は2台の直列に</td><td rowspan="5">C03
(1168)
[0.0219]
MS/MSの概略　MS/MS法は2台の直列に</td><td rowspan="3">C07
(756)
[0.0605]
お知らせ機器センターのホームページ</td><td rowspan="2">C13
(196)
[0.1012]
気になる最新情報はここでGET！</td><td>C15 (17)
訃報の掲載依頼者はこのページに対し正しい</td></tr>
<tr><td>C16 (42)
無料 掲示板　無料 カウンタ 無料素材</td></tr>
<tr><td rowspan="2">C14
(264)
[0.0761]
FIR フィルタの設計</td><td>C17 (11)
設備・機器利用便覧機器センター</td></tr>
<tr><td>C18 (11)
FIR フィルタの設計</td></tr>
<tr><td colspan="3">C08 (21)　人物 死因 予定 喪主 連絡先</td></tr>
<tr><td rowspan="2">C04 (149)
[0.4149]
会報"旭天"　星と旭川天文同好会</td><td colspan="3">C09 (31)
会報"旭天"　星と旭川天文同好会</td></tr>
<tr><td colspan="3">C10 (15)
惑星 天体画像 旭川天文同好会</td></tr>
<tr><td rowspan="3">C02
(110)
[0.3820]
西村会員のオーロラ</td><td rowspan="2">C05 (52)
[0.5818]
アフリカ皆既日食</td><td colspan="3">C11 (30)
アフリカ皆既日食</td></tr>
<tr><td colspan="3">C12(18)
トルコ皆既日食</td></tr>
<tr><td colspan="4">C06 (55)　西村会員のオーロラ</td></tr>
</table>

図 3.8: Cluster 1 by Average Linkage Method

<table>
<tr><td rowspan="10">C00
(5513)
[0.0044]
Okayama University HEP Lab.</td><td rowspan="9">C01
(4771)
[0.0376]
釧路ミサワホーム本店</td><td rowspan="8">C03
(3027)
[0.0594]
これまでに制作した「偉人の筆跡カレンダー」</td><td rowspan="3">C05
(1520)
[0.1793]
ミサワホーム，ミサワホーム株式会社 住宅メーカー</td><td rowspan="3">C07
(1092)
[0.2150]
ミサワホーム株式会社 住宅メーカー</td><td rowspan="2">C11(967)
[0.2464]
ミサワホーム株式会社 住宅メーカー</td><td>C15 (298)
ミサワホーム株式会社 住宅メーカ</td></tr>
<tr><td>C16(55)
MISAWA兵庫</td></tr>
<tr><td rowspan="2">C12 (120)
[0.4285]
丈夫な木の住まい</td><td>C17 (63)
丈夫な木の住まい</td></tr>
<tr><td>C18 (25)
MISAWA千葉</td></tr>
<tr><td colspan="3">C08(32) お問い合わせはミサワホーム</td></tr>
<tr><td rowspan="3">C06
(412)
[0.3698]
ミサワホーム中国岡山支店</td><td colspan="3">C09(50) MISAWA岩手</td></tr>
<tr><td rowspan="2">C10
(356)
[0.4135]
ミサワホーム中国岡山支店</td><td>C13(45)
MISAWA愛知</td></tr>
<tr><td>C14 (45)
ミサワホーム中国岡山支店</td></tr>
<tr><td colspan="4">C04 675) 横浜市　停車列車</td></tr>
<tr><td colspan="5">C02 (44) Okayama University HEP Lab.</td></tr>
</table>

図 3.9: Cluster 2 by Average Linkage Method

<table>
<tr><td rowspan="6">C00
(3847)
[0.0325]
水田のビオトープについて
山田　利彦
(B4)</td><td rowspan="3">C01
(2470)
[0.0494]
水田のビオトープについて
山田
利彦
（B4）</td><td rowspan="2">C03 (518)
[0.0689]
土木学会第55回年次学術講演会講演概要集</td><td colspan="2">C07 (320)
土木学会第55回年次学術講演会講演概要集</td></tr>
<tr><td colspan="2">C08 (198)
岐阜大学・Gifu University 大学の午前の風景</td></tr>
<tr><td>C04 (1952)
[0.0953]
水田のビオトープについて山田
利彦（B4）</td><td>C09 (600)[0.1106]
ライブラリー情報 アビリティガーデン・ライブラリー内に所蔵している</td><td>C13 (34)
北関西ＮＡＶＩ
C14 (378)
学術雑誌目次速報データベース</td></tr>
<tr><td colspan="3">C10 (41)
長岡技術科学大学長谷川研究室</td></tr>
<tr><td rowspan="2">C02
(664)
[0.0830]
このホームページに掲載の文章・イラスト・写真等の無断転載を禁じます</td><td rowspan="2">C05 (566)
[0.1120]
このホームページに掲載の文章・イラスト・写真等の無断転載を禁じます</td><td>C11 (108)
[0.9317]一般情報 技法一覧 教育ゲーム 伝言ゲーム 会社インタビュー</td><td>C15 (36)
一般情報 技法一覧 教育ゲーム伝言ゲーム 会社インタビュー
C16 (34)
伝言ゲーム 会社インタビュー</td></tr>
<tr><td colspan="2">C12 (219)
このホームページに掲載の文章・イラスト・写真等の無断転載を禁じます</td></tr>
<tr><td colspan="3">C06 (40)
フリー素材 作者の方々のご厚意により，音楽関連のフリー素材</td></tr>
</table>

図 3.10: Cluster 3 by Average Linkage Method

<table>
<tr><td rowspan="6">C00
(5091)
[0.0044]
【温度計に息吹いてます！】ジョンブル西
暑い～暑いよ～（ＴＴ）</td><td rowspan="6">C01
(5031)
[0.0044]
【温度計に息吹いてます！】ジョンブル西
暑い～暑いよ～（ＴＴ）</td><td rowspan="4">C03
(2715)
[0.0044]
【温度計に息吹いてます！】ジョンブル西 暑い～暑いよ～（ＴＴ）</td><td rowspan="2">C05
(494)
[0.0044]
法政大学 社会学部 助教授 白田 多摩キャンパス 社会学部棟 917号室</td><td>C09
（408）
[0.0044]
法政大学 社会学部 助教授 白田 多摩キャンパス</td><td>C13 (323)
法政大学 社会学部 助教授 白田 多摩キャンパス 社会学部棟 917号室
C14 (85)
情報革命と著作権法の問題点 Chapter 1</td></tr>
<tr><td colspan="2">C10 (50)
講義の方針・目標
情報化社会の様態を把握し，</td></tr>
<tr><td rowspan="2">C06
(1798)
[0.0044]
タレコミと内部告発が日本を救う</td><td>C11 (1163)
タレコミと内部告発が日本を救う，総合アングラページ</td></tr>
<tr><td>C12 (129) 岩手大学のホームページ</td></tr>
<tr><td rowspan="2">C04
(2316)
[0.0044]
日本産アリ類カラー画像データベース</td><td colspan="3">C07 (1593)
日本産アリ類カラー画像データベース</td></tr>
<tr><td colspan="3">C08 (410)
ヒラセヨツバアリ
イツツバアリ
ミツバアリ
ヨツバア</td></tr>
<tr><td colspan="5">C02 (60) 識別記号 NASA識別記号 打ち上げ 軌道要素 軌道タイプ : 円軌道</td></tr>
</table>

図 3.11: Cluster 4 by Average Linkage Method

<table>
<tr><td rowspan="5">C00
(4407)
[0.0007]
このページ内の記事及び画像の転載，二次使用を堅く禁じます.</td><td rowspan="4">C01
(3345)
[0.0033]
このページ内の記事及び画像の転載，二次使用を堅く禁じます.</td><td rowspan="3">C03
(626)
[0.0358]
福岡グルメリンク 地域別</td><td rowspan="2">C07
(206)
[0.0788]
福岡グルメリンク 地域別</td><td>C09 (206)
[0.1854]
登録するお店の該当する地域 福岡市内北九州市内</td><td>C11 (114)
登録するお店の該当する地域
福岡市内 北九州市内</td></tr>
<tr><td></td><td>C12 (95)
北九州市周辺 下関市周辺 九州らーめん</td></tr>
<tr><td colspan="2"></td><td colspan="2">C10 (126)
福岡グルメリンク 地域別</td></tr>
<tr><td colspan="3">C08 (73)
桜井淳子 著者(撮影者)： 平地勲 サイズ：菊上製 ページ数：96 発行年月：2001年1月 価格：3500円</td></tr>
<tr><td colspan="3">C04 (151)
このページ内の記事及び画像の転載，二次使用を堅く禁じます.</td></tr>
<tr><td rowspan="2">C02 (60)
[0.6169]
表は誰がいるのかしら 9/13(木)0:21) 大丈夫か (9/13(木)0:21)</td><td colspan="4">C05 (30)
表は誰がいるのかしら9/13(木)0:21) 大丈夫か (9/13(木)0:21)</td></tr>
<tr><td colspan="4">C06 (30)
＼＼　おにぎりワッショイ！！　／／　+
+　＼＼　おにぎりワッショイ！！／+</td></tr>
</table>

図 3.12: Cluster 5 by Average Linkage Method

<table>
<tr><td rowspan="6">C00
(2351)
[0.0160]
バックナンバー：‘98 ニーフル春号 バックナンバー：‘97 ニーフル秋・冬号 バックナンバー：‘97 ニーフル夏号</td><td rowspan="5">C01
(840)
[0.0530]
ふるさと村フリーペーパー「ニーフル」のオンライン版です.</td><td rowspan="4">C03
(272)
[0.6877]
七五三和装男の子 七五三和装女の子</td><td rowspan="2">C07 (88)
[0.0044]
百日衣装コレクション（メルヘン） / m3十条写真スタジオ 百日祝着男</td><td>C11 (48)
百日衣装コレクション（メルヘン） / m3十条写真スタジオ 百日祝着男</td></tr>
<tr><td>C12 (40)
百日衣装コレクション（祝着女の子）十条写真スタジオ 百日ドレス</td></tr>
<tr><td rowspan="2">C08 (94)
[0.9735]
七五三和装男の子 七五三和装女の子</td><td>C13 (8)
七五三和装男の子 七五三和装女の子</td></tr>
<tr><td>C14 (8)
百日祝着男 百日祝着女 百日メルヘン</td></tr>
<tr><td colspan="3">C04 (8) うるおいの新潟（社）新潟県観光協会</td></tr>
<tr><td rowspan="2">C02
(346)
[0.0313]
蹴球七日ギャラリ この国どんな味？ キャラクター原論</td><td colspan="2">C06 (62)
[0.7881]
作品タイトル 暗き庭で鳴るは霹靂 作者名 有田佳貴 職業中学生</td><td>C09 (20)
作品タイトル 暗き庭で鳴るは霹靂 作者名 有田佳貴 職業 小学生</td></tr>
<tr><td colspan="2"></td><td>C10 (34)
作品タイトル 暗き庭で鳴るは霹靂 作者名 有田佳貴 職業 中学生</td></tr>
<tr><td colspan="3">C05 (29) 小池氏原作「マッド・ブル2000」（画・井上紀良）</td></tr>
</table>

図 3.13: Cluster 6 by Average Linkage Method

第4章　Web文書集合の階層的要約と評価

4.1　前書き

近年，インターネットの普及により World Wide Web(WWW) の世界は急激に拡大し膨大なテキスト情報をもたらした. そのため，個別の Web 文書要約だけでなく Web 文書集合全体から内容をすばやく把握する方法，即ち Web 文書集合の自動要約に対するニーズが高まっている.

この問題に取り組むためのアプローチとして *Web* **クラスタリング**が挙げられる. これまで数多くの Web クラスタリング手法が提案されているが [5]. その目的は様々であり,Web 上でのクラスタリング, Web ログ・セッション入手,Web セッションクラスタリング, Web コミュニティ検出 (*Authority* や *Hub*),Web 文書クラスタリング, 検索エンジン結果の集約など多岐に渡っている.

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 [13]. つまり, クラスリング技法は”類似性”の定義とその実行方法に依存して, 隠れたパターンをどれだけ見出せるかを競い合っているといえる. これまで知られたクラスタリング技法は, 大きく分割方式 (オブジェクト集合を分割し, ある基準で評価する), 階層化 (オブジェクト集合をある基準で階層的に分解する), 密度に基づく手法 (結合度・密度関数による評価) などに大別され, 類似性の定義は距離の定義として考察されることが多い.

ほとんどの Web 文書が扱うデータはカテゴリカル (例えば *ComputerScience* ,*Biology*,*athematics* など）であり, 数値データではない. このため距離や順序概念が考えにくく, 過去に提案された多くのアプローチがなじみにくいの理由のひとつになっている.

要約はクラスタリングとは少々異なり，情報ソースにおける重要な内容を簡約な形で提示することである. ここでは一見するだけで内容が把握できることが望ましい. 人手による要約は，大意と要旨の二つに区別することができる [2]. 大意とは原文の表現をできるだけ用いて順序を変えずにまとめたものであり，要旨と

は原文の主題や結びに焦点を絞り原文の表現形式にとらわれずにまとめたものである.

これに対応して，自動要約の研究では**抜粋** *(Extraction)* と**抽象化** *(Abstraction)* という手法が提案され研究されている. 抜粋とは対象文書 (あるいは集合) から文章を抜き出す手法を意味する. 例えば，重要文抽出法では，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である. 新たに文章を作る必要がなくまた言語知識をほとんど必要としないため実現が容易である. 反面，代名詞や接続詞などの関連が整合性を有さないことがあり，語句間の関連性を壊す可能性があること，箇条書きや表などの構造情報の取り扱いに統一性がなく，この結果長大な題材をうまく取り扱えない. これに対し，対象文書 (集合) の内容を把握し，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を抽象化と呼ぶ. 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. 過去には，新聞記事のように出現内容を規定できるものや，教科書などのように語句が整形された文書に限定して適用されている.

これらの技術でテキストを対象として要約するため，リンクやタグなどの半構造を有する Web 文書に適用することは容易でない. そこで, 本稿では Web 文書クラスタリングと Web 文書要約という二つの考えに基づき, 階層型クラスタリングを用いた Web 文書集合の要約手法を提案する. 我々はこの階層をコストというペナルティを与えるタイプの尺度で評価する. 一般的に, クラスタリングと要約の評価は困難である. なぜならば, 正しい要約の出力を定義することができないからである. 本稿では 3 つの尺度, クラスタの可読性, 階層の可読性, そして読解 (reading comprehension) という評価方法を提案する。この評価方法は事前に人手による正解がなくても定量的に評価することができる。第 2 章では自動要約について現状を述べ，第 3 章では階層的抽象化手法について論じる. 第 4 章, 第 5 章で提案手法を述べ，第 6 章で実験結果を挙げ本手法の有用性を示し，第 7 章は結びである.

4.2　階層的要約手法

我々は新しい自動要約手法として構造化を提案した [28]. **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*[21] 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合せクラスタリングについて述べる [27]. 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.

4.2.1　組合せクラスタリング

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合せた Web 文書クラスタリング手法を**組合せクラスタリング**と呼び, その有用性を示した [27]. 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**組合せクラスタリング**の概要を要約する, 詳細は文献 [27] 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 [13]. 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”組合せる”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.

ここで組合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 4.1 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 4.2 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 4.3 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, これを組合せクラスタと呼ぶ. 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.

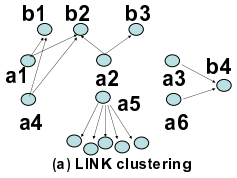


図 4.1: LINK Clustering

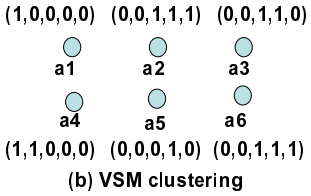


図 4.2: VSM Clustering

4.2.2　階層的要約

前節で我々はどのように Web 文書集合を得るかを述べた. 組合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 文書集合の階層的要約を生成する手法について述べる [28].

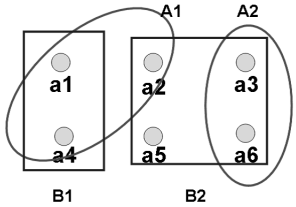


図 4.3: Combination Clusters

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 4.4 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的要約を得る.

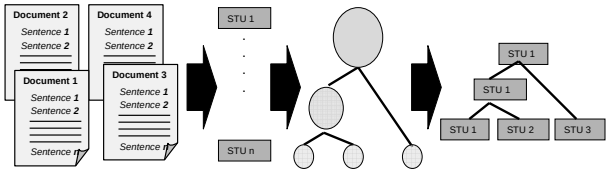


図 4.4: overview

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.

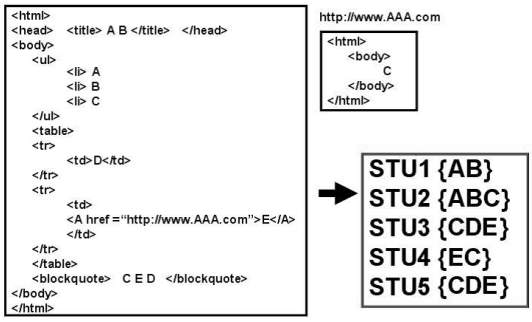


図 4.5: Taking STUs from Web Pages

図 4.5 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 4.6 に示す.

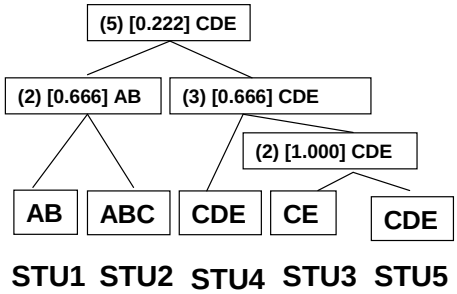


図 4.6: Hierarchy using STUs

4.3　評価手法

要約をどのように評価するかは難しい問題である [15]。問い合わせに対して答えを出力する場合は正答は存在するだろう, しかし要約や機械翻訳といった技術では正しい出力という概念にたどり着くのは非常に困難である。正しい出力の近似として人手の要約を用いる場合, 作成する人による要約の差異から相互の判定の同意をとり正解を作る手法がある [23], 一般に人間の要約は正しい出力とはならないかもしれないことには注意を払うべきである。こうしたことから自動要約研究者にとって評価方法は長い間関心をもたれている。

4.3.1　自動要約における既存の評価手法

Mani らが紹介した自動要約のいくつかの評価尺度がある [15]。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する。高い圧縮率の要約では, 当然原文の内容を全て包括しているわけではない。原文の内容をどの程度有しているかという尺度を情報量 (informativeness) と定義する。informativeness のアイデアは, 原文の内容を過不足なく包括する要約とは原文の複写であることを示唆している。一般に圧縮率と informativeness はトレードオフの関係にある。抜粋やクラスタリングにおいて適合率と再現率による F 尺度が informativeness の評価に対応する。これには事前に抜粋する正解の文章集合が必要である。抽象化において informativeness の評価方法は難しい。Dice 係数やコサイン類似度は原文と要約に含まれる単語の重なりを計算できるが, 同義語・同音異句語や構文を考慮していない。シソーラスや LSI, 構文解析, 対話理解といった技術

が必要となるだろう。次に可読性 (readability) と読解 (reading comprehension) の評価に着目する。可読性とは要約によって用意に内容をお把握可能かの尺度である。特にいくつかの文章で出力する場合, 孤立した照応詞 (接続詞や代名詞) や文脈を人手により採点する方法などが採用されている。読解タスクでは要約を読んだ人間の理解度について評価する。全文か要約を人間が読み, 内容についてテストする方法などがある。

4.3.2　TDT における既存の評価手法

階層構造の要約を評価するときにはどのような点に注意を払うべきかが問題となる. ここで我々は階層的要約と非常に似た例として階層トピック検出を挙げる。The Topic Detection and Tracking (TDT) project は the National Institute of Standards and Technology(NIST) が端を発した研究分野である。[1] この分野では大きく分けて二つタスクがある一つはニュース記事などを論じられているイベント (or topic) ごとに対応するクラスタに組織化することを要求する, 検出タスク。そしてもう一つは関連するクラスタを追跡するタスクである。検出タスクはニュース記事に関する単純な仮定に基づいて行われる。仮定（1）, 一つのニュース記事は一つのイベントについてのみ記述する。つまりニュース記事は明確に一つのクラスタにのみ配置されるとする。仮定（2）, イベント（トピック）間のいかなる階層関係も無視する。この結果, トピック検出システムは非階層的なグループに記事を配置することとなる。これは明らかに現実のイベント（トピック）の性質を反映していない。

2003 年からＴＤＴでは, より現実的なモデルとしてクラスタ間の関係を階層構造にする階層的トピック検出を奨励し始めた [32]。このとき従来のトピック検出タスクの評価方法では適切に階層構造を評価することが出来ない。Allan ら [2] は階層構造を評価するいくつかの方法について議論した。

我々は Allan らの階層構造を評価手法の内, 特に 3 つの尺度に着目する。最初に, トピック検出コストは従来のトピック検出と同様に false alarm と miss detection にペナルティを課す評価方法である。検出コストを評価するために, ニュース記事の理想的な排他的分類を正解クラスタと呼ぶ. システムの出力と正解クラスタの比較のとき四つの組合せを表に示す.

ここで R+,R-,N+,N-はそれぞれのカテゴリに属する要素の数である。全要素数を n, クラスタの要素数を r としたとき, クラスタの miss detection 率 P_{miss} と false alarm 率 P_{fa} の定義は次式の通り

$$P_{miss} = \frac{R_{-}}{r} \tag{4.1}$$

表 4.1: Four combinations in comparison

system output	relevant	non-relevant
in cluster	R_-	N_+
not in cluster	R_+	N_-
total	r	$n-r$

$$P_{fa} = \frac{N_+}{n-r} \tag{4.2}$$

そして検出コストは P_{miss} と P_{fa} の線形の組合せで表現される。

$$C_{det} = C_{miss}P_{miss}P(target) + C_{fa}P_{fa}(1 - P(target)) \tag{4.3}$$

このとき, Fiscus や Doddington らは miss detection は false alarm よりもより強くペナルティを課すべきであると言及している. そのため, トピック検出の事前確率や miss detection コストと false alarm コストの重み次のように与える. $P_{target} = 0.02$,$C_{miss} = 10$,$C_{fa} = 1$. これにより検出コストは

$$C_{det} = 0.2P_{miss} + 0.98P_{fa} \tag{4.4}$$

次に, トラベルコストに着目する. トラベルコストはルートノードから各トピックに最適なクラスタを見つけるコストである. このコストはクラスタの深さと検出コストから定義する. ルートからの深さ D としたときトラベルコストは次式の通り

$$depth = D/max(D) \tag{4.5}$$

$$C_{travel} = C_{det} + depth \tag{4.6}$$

最後に, ルートノードからトラベルコストの最も小さいクラスタを評価するために最小コストを定義する.

$$C_{minimal} = min(C_{travel}) \tag{4.7}$$

実際に例題の階層を評価してみよう。このとき $1, 2, 3, 4$ の二つのクラスタを正解とする。図に各コストの計算の詳細を示す。最初の合併で $4, 5$ が合併するとき正

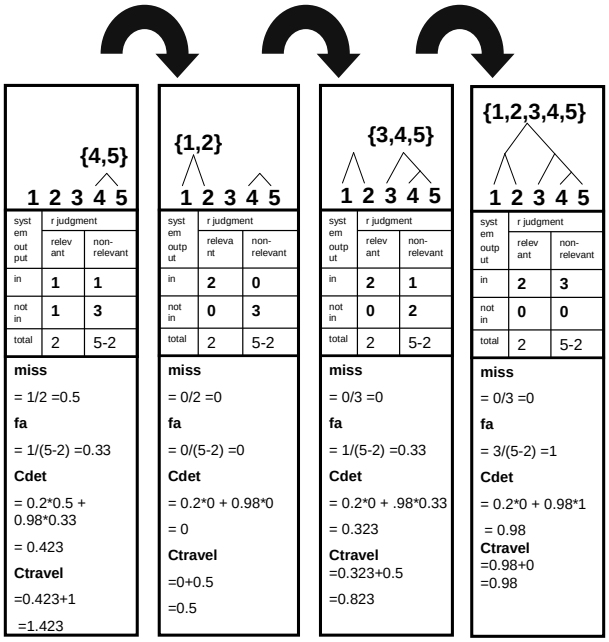


図 4.7: ex:allan

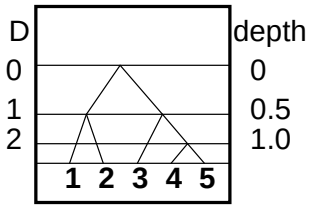


図 4.8: ex:depth

解クラスタ 3,4 と比較すると R_- は 3, N_+ には 5 が割り当てられ図の計算の通りとなる。このとき minimal cost は 1,2 クラスタの $C_{minimal} = 0.5$ となる。

allan の評価方法は階層構造を定量的に評価することができる．しかしながら, この評価方法にはいくつかの問題がある．一つは正解クラスタは階層を考慮した正解ではない点である．正解クラスタのように排他クラスタでは, ニュース記事を唯一つのカテゴリに分類するのは困難であるからである．例えばメジャーリーガーの結婚というニュース記事はスポーツカテゴリとゴシップカテゴリのどちらに分類するかは自明ではない． またそうした正解クラスタを事前に用意しておく必要があるという問題もある．もう一つは *Power set* の影響である．我々は複数のトピックを含むクラスタを Power set と呼ぶ．あるクラスタが Power set で

あるとき, クラスタの内容を把握することは困難である．しかしながら,Allan の手法は Power set にペナルティを寄与しない．例題の {1,2} と {3,4,5} の合併では $misscost = 0$ で $falsealarmcost = 1$ となり false alarm により $Power\ set$ にペナルティを与えることができている。しかし全要素数が巨大なとき (たとえば $n = 270,000$)$falsealarmcost = N_+/(270000 - r)$ となり, ルートノード付近のクラスタ要素数 r でなければペナルティを与えることができない。

4.3.3　提案評価手法

このセクションで我々の提案する評価方法について示そう。我々の提案する階層表現は共起性を用いることでクラスタの関係を階層で出力する。共起性と階層と両方の様相から評価する必要がある。これまで自動要約の評価方法と階層 TDT の評価方法をみてきた。我々はこれらの評価方法を元にクラスタと階層の両方の可読性を評価する方と, さらにルートノードから読解するコストという評価方法を提案する。

クラスタの可読性を評価するために一つの概念を導入する。クラスタには粒度という概念があり,「クラスタの粗さ細かさ」を意味する。粒度はクラスタのサイズとは違い, クラスタ要素が相互に類似していればクラスタの粒度は細かく, 要素が類似していなければ粒度は粗くなる。例えば,A〜Z の 26 個の要素からなるクラスタと要素 A が 26 個からなるクラスタではクラスタのサイズは同じだが, 明らかに後者のクラスタのほうがクラスタの内容を把握しやすいといえるだろう。つまり, 類似した内容の要素同士であれば容易に内容を把握することができるということだ。このクラスタの粒度はクラスタの可読性と強く関係している。ここで我々は allan の検出コストに再度注意を払う。検出コストは類似していない要素にペナルティを課す false alarm コストと類似しているがクラスタ内にない要素にペナルティを課す miss detection コストがある。われわれは可読性と Allan の検出コストを関連付けるアイデアを提案する。クラスタ内の相互非類似度は粒度の粗さ**内部粒度**とし, 次のように表す。

$$C_{in} = 1 - \frac{2}{m(m-1)} \sum_{i=1}^{m} \sum_{j=i}^{m} sim(x_i, x_j) \tag{4.8}$$

ここで, 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．また miss detection コストは他のクラスタに類似した要素があればペナルティを課すことから, クラスタ間の粒度をクラスタ間類似度によって定義し, **外部粒度**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,

$$C_{miss} = \sum_{i=1}^{s} \sum_{j=i}^{s} sim(Cl_i, Cl_j) \tag{4.9}$$

前者は Allan の false alarm コストと, 後者は miss detection コストと対応する. こうした二つのコストの線形和による式を, 可読性を評価する**粒度コスト**とする.

$$C_{det} = C_{miss} + C_{fa} \tag{4.10}$$

階層の評価方法としてコーフェン相関係数という階層クラスタリングの分野で使われている方法を導入する。まず次の例を見てほしい, ある類似度行列から二つの手法を用いて階層クラスタリングを生成している。

表 4.2: ex:similarity matrix

	1	2	3
1	0	1	0
2	0	0	0.8
3	0	0	0

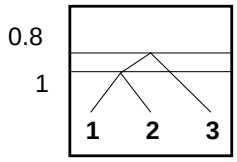


図 4.9: ex:hierarchical structure 1

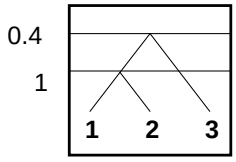


図 4.10: ex:hierarchical structure 2

前者の階層では要素 1 と要素 3 の類似度は行列では 0 であるのに対して, 階層では 0.8 と読むことが出来る。後者では階層では 0.4 であると読むことが出来る。後者のほうが類似度行列との差異は少く, より類似度行列を反映した階層を我々に提示している。こうした類似度行列と階層との差異を評価することでこの二つの行列の相関関係を評価することとなる。我々は階層の可読性とは, 類似度行列をより反映している階層を評価するための尺度であると考える。そこでコーフェン相関係数を用いる。階層を生成するときに類似度行列はアップデートを繰り返し行うため, もともとの類似度は保つことができない。このとき階層を行列の形で表した

ものをコーフェン行列と呼ぶ。類似度行列 x とコーフェン行列 y とのピアソン積率相関係数をとることでその歪みの量を評価することができる。

$$r_{x,y} = \frac{\sum xy - (1/n)(\sum x)(\sum y)}{\sqrt{\{\sum x^2 - (1/n)(\sum x)^2\}\{\sum y^2 - (1/n)(\sum y)^2\}}} \tag{4.11}$$

この $r_{x,y}$ をコーフェン相関係数と呼び,1 に近ければ正の相関,-1 に近ければ負の相関を持つ。

ルートから最適なクラスタへのパスは読解の評価と関連づけられる。ルートノードから近いクラスタに粒度の小さいまとまりのある内容のクラスタがあれば読解するのは容易になる。トラベルコストや最小コストでこれを評価することができるだろう。allan の方法と同様にルートからの深さ D としたとき読解コストは次のように定義する。

$$depth = D/max(D) \tag{4.12}$$

$$C_{travel} = C_{det} + depth \tag{4.13}$$

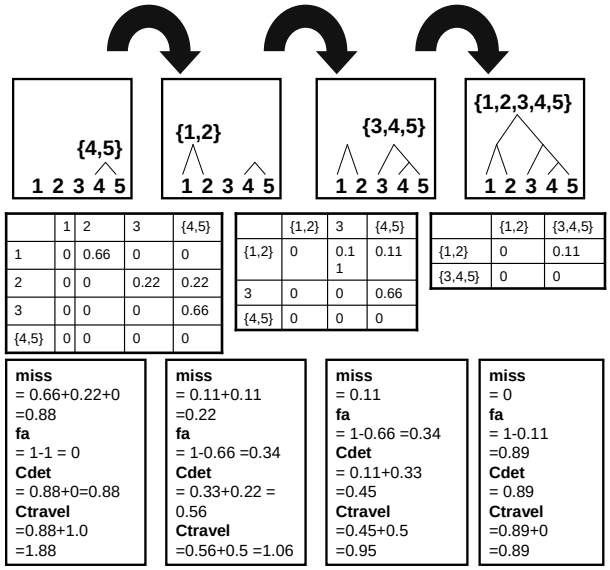
$$C_{minimal} = min(C_{travel}) \tag{4.14}$$

実際に提案手法による評価方法の例をみてみよう。例題の類似度行列は次の表に示す。クラスタの可読性と読解の評価について, 図 4.11 で合併により変化して

表 4.3: proposal similarity matrix

	1	2	3	4	5
1	0	0.66	0	0	0
2	0	0	0.22	0.22	0.22
3	0	0	0	0.66	0.66
4	0	0	0	0	1.0
5	0	0	0	0	0

いく類似度行列と各クラスタのコスト計算の詳細をみることができる。我々の提案する評価手法を用いることで, 階層構造を定量的に評価することができる. このとき正解クラスタを必要とせず,Power set にペナルティを課すことができるという利点がある.



	1	2	3	{4,5}
1	0	0.66	0	0
2	0	0	0.22	0.22
3	0	0	0	0.66
{4,5}	0	0	0	0

	{1,2}	3	{4,5}
{1,2}	0	0.11	0.11
3	0	0	0.66
{4,5}	0	0	0

	{1,2}	{3,4,5}
{1,2}	0	0.11
{3,4,5}	0	0

図 4.11: ex:proposal

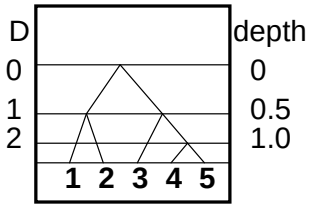


図 4.12: ex:proposal depth

4.4　実験

4.4.1　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 件を用いて要約の対象とする．

4.4.2　実験 1

提案評価方法と allan の方法を比較を行う。allan の方法を試すにあたり, 正解クラスタとして 1600 個の STU に対して 35 個クラスタを人手により割り当てたもの

を利用する. そして深さ毎の miss と fa のコストの平均のグラフと Cdet と Ctravel のグラフは以下のようになる.

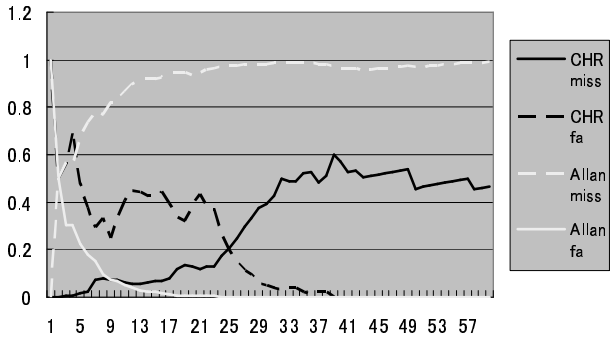


図 4.13: ex:miss cost fa cost

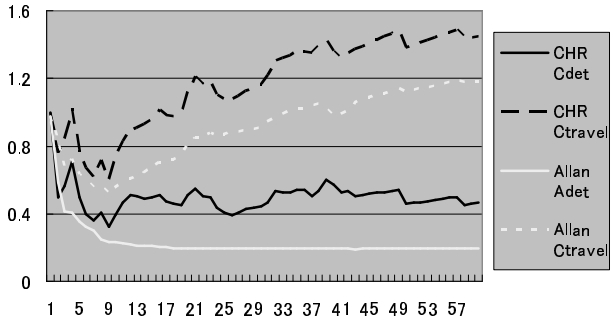


図 4.14: ex:detection cost travel cost

図 4.14 では detection cost と travel cost と両方で提案手法と allan のコストが最小となる深さが一致している. だが図 4.13 をみるとあきらかにコストが最小になる深さは提案手法と allan で異なっている. allan の手法は重み付けを行うことで理想的な detection cost と travel cost に近似させようとしたが, 提案手法は重み付けも正解も使わずとも同様の結果を得ることができた.

allan の評価方法による方法で,miss detection cost と false alarm cost の式の分母に着目する. 全 STU 数が正解クラスタのトピック数より十分大きい場合, ルートに近い深さ miss detection cost は減少していき,false alarm cost は増加していく傾向にある. つまり正解クラスタのトピック数と全 STU 数に強く依存しているといえる. 一方, 提案手法はトピック間の類似度も考慮していることから Power set に適切にペナルティを課すことができていることに起因すると考える.

4.4.3　実験 2

つぎに我々は STU を生成する際にリンクと入れ子を利用することで優位な要約となるかを評価する. 各 cost を最小にするような $\hat{y}$ を STU の生成方法の集合 $Y = STU, STU(Nest), STU(Link), STU(Nest + Link)$ から発見し, それを出力することを次式で表す.

$$\hat{y} = argmin_{y \in Y}(Cdet(y)) \tag{4.15}$$

これにより detection cost に優れた STU 生成方法をえる. 同様に travel cost は次式.

$$\hat{y} = argmin_{y \in Y}(Ctravel(y)) \tag{4.16}$$

またコーフェン相関係数においては次式をみたす $\hat{y}$ を出力する. こうして, クラスタの可読性, 読解, 階層の可読性にすぐれた STU 生成方法を調べる.

$$\hat{y} = argmax_{y \in Y}(cophenetic(y)) \tag{4.17}$$

detection cost,travel cost 両方とも STU(Nest+Link) が優位であることが図 4.15

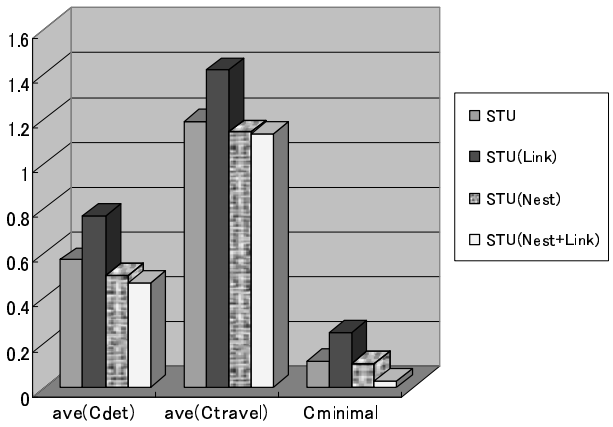


図 4.15: argmin

からわかる. 次に STU(Nest) が優位であることから, 入れ子の分 STU によりクラスタの粒度が細かくなり, コストを下げることができた. また, 表 4.4 より STU(Link) が優位であることが読み取れる. リンク先の内容を STU に反映させることにより, 各クラスタの重心に影響を及ぼしたことがコーフェン相関係数を向上させた原因であるだろう.

以上のことから, コーフェン相関係数が比較的優位で detection cost,travel cost で最も優位である STU(Nest+Link) は最も階層的要約に適した STU 生成方法であるといえる.

表 4.4: cophenetic correlation coefficient

	cophenetic correlation coefficient
STU	0.678701
STU (Nest)	0.667447
STU (Link)	0.775679
STU (Nest+Link)	0.717835

4.5　結び

本稿では Web 文書集合を階層表現により要約する手法を提案した. Web 文書集合を STU に分割し，階層型クラスタリングを用いることで容易に実現可能なことを実験により示した. しかし，リンクやタグ構造を利用することでトピックドリフトが発生してしまうことも確認できた. 今後の展開としては Web 文書集合の文脈となるページを利用した要約との比較が考えられる.

第5章　階層的要約を用いたWeb文書集合への問合せ

5.1　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 [14, 3]. 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 [14]. 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 [15]．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 [13]．

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 [27]．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 [28] 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, これを階層的要約と呼ぶ．

そこで, 本稿ではこの階層的要約に問合せをし, 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的要約手法の概要を要約する, 詳細は文献 [28] を参照. 第 4 章で本稿で提案する階層的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 章は結びである.

5.2　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 [14, 3]. 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 [30] 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 [25] 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 [20]. 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.

5.3　階層的要約手法

我々は新しい自動要約手法として構造化を提案した [28]. **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*[21] 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合せクラスタリングについて述べる [27]. 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.

5.3.1　組合せクラスタリング

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合せた Web 文書クラスタリング手法を**組合せクラスタリング**と呼び, その有用性を示した [27]. 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**組合せクラスタリング**の概要を要約する, 詳細は文献 [27] 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 [13]. 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”組合せる” 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.

ここで組合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 5.1 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 5.2 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 5.3 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, これを組合せクラスタと呼ぶ. 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.

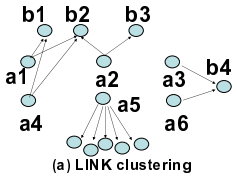


図 5.1: LINK Clustering

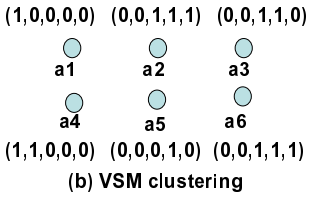


図 5.2: VSM Clustering

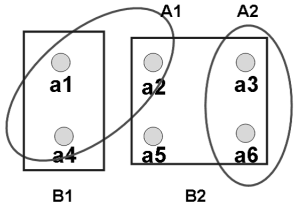


図 5.3: Combination Clusters

5.3.2　階層的要約

前節で我々はどのように Web 文書集合を得るかを述べた. 組合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 文書集合の階層的要約を生成する手法について述べる [28].

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 5.4 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的要約を得る.

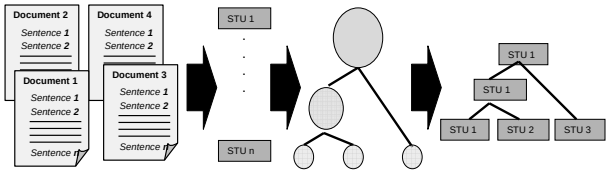


図 5.4: overview

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．

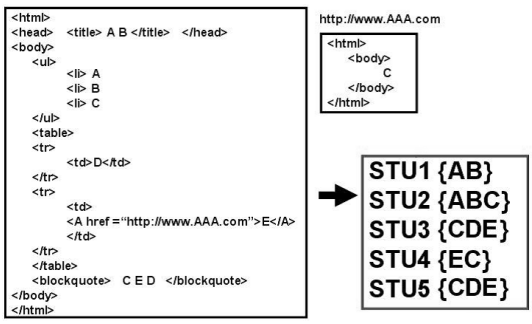


図 5.5: Taking STUs from Web Pages

図 5.5 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 5.6 に示す.

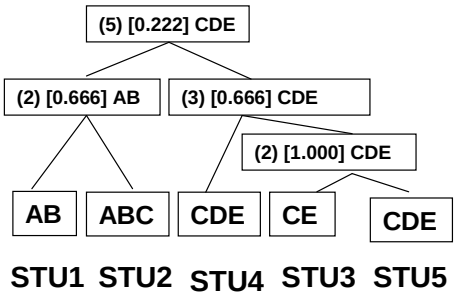


図 5.6: Hierarchy using STUs

5.3.3　階層的要約の評価方法

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 [15]. 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. 我々は階層的要約を定量的に評価する方法として CHR Method を提案した [29]. 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.

$$G_{in} = 1 - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i}^{n} sim(x_i, x_j) \tag{5.1}$$

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,

$$G_{out} = \sum_{i=1}^{s} \sum_{j=i}^{s} sim(Cl_i, Cl_j) \tag{5.2}$$

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.

$$C_{det} = G_{in} + G_{out} \tag{5.3}$$

5.4　階層的要約への問合せ

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的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 文書集合には階層的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. 階層的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 [12] 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.

$$cos(q\, N) = \frac{q \cdot N}{\|q\| \|N\|} \tag{5.4}$$

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 [12] は以下のように示される.

$$bGIOSS(Q\,N) = \|N\| \sum_{i=1}^{n} \frac{\|q_i\|}{\|N\|} \tag{5.5}$$

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.

ここで階層的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 5.7 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 5.8 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 5.9 で示す.

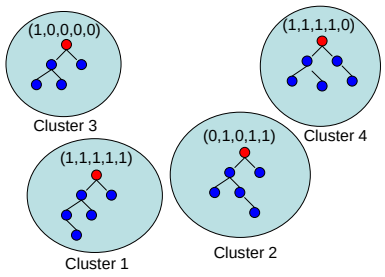


図 5.7: querying to root node

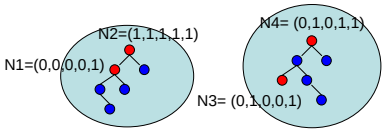


図 5.8: querying to nodes

5.5　実験

5.5.1　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 文書に組合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的要約を適用する．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．

- 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

5.5.2　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

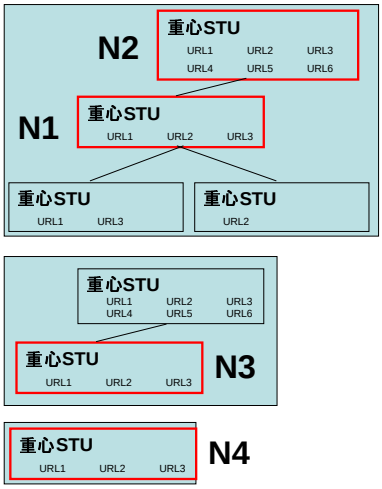


図 5.9: querying result

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.

$$\text{適合率} = \frac{w(URL_N \cap URL_{HITS})}{w(URL_N)} \tag{5.6}$$

$$\text{再現率} = \frac{w(URL_N \cap URL_{HITS})}{w(URL_{HITS})} \tag{5.7}$$

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 5.10 に示す.

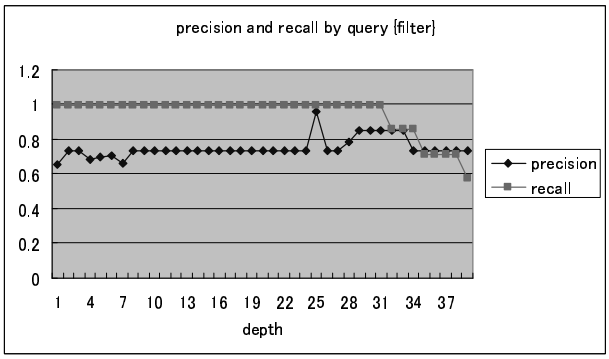


図 5.10: precision and recall by query {filter}

問合せ語 { フィルタ, 実験 } で問合せたときの関係を図 5.11 に示す.

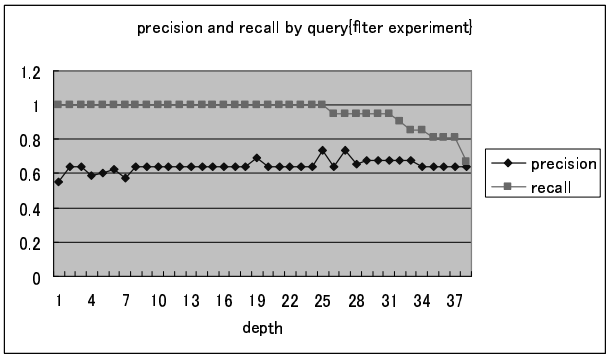


図 5.11: precision and recall by query {filter , experiments}

再現率は図 5.10,5.11 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.

5.5.3　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 5.12 に示す.

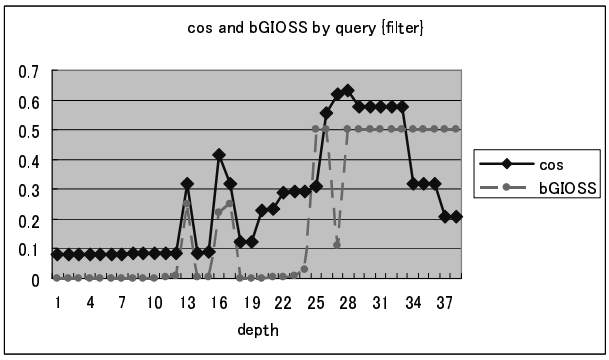


図 5.12: cos and bGIOSS by query {filter}

問合せ語 { フィルタ, 実験 } で問合せたときの関係を図 5.13 に示す.

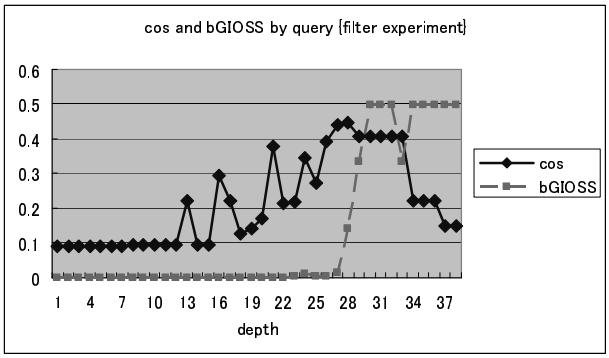


図 5.13: cos and bGIOSS by query {filter , experiments}

図 5.12 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 5.13 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.

5.5.4　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 5.14 に示す.

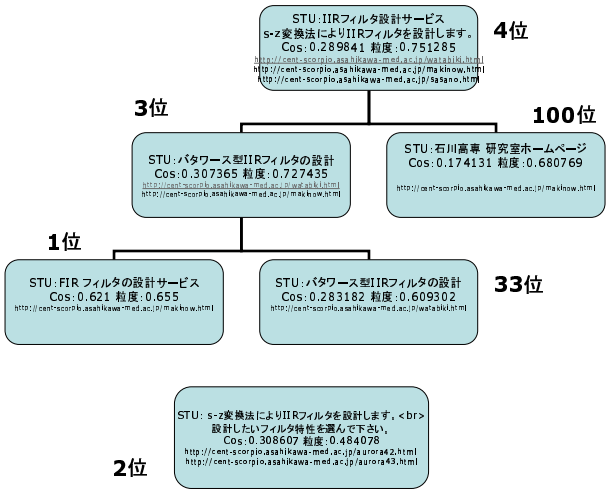


図 5.14: hierarchy by query {filter , experiments}

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.

5.6　結び

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. 今後の展開としては動的に Web 文書集合の階層的要約の生成する手法などが挙げられる.

第6章　結論

本研究では，Web 文書集合全体を閲覧することなく内容を素早く把握するために, 新しい自動要約手法と, それを定量的に評価する手法を提案し，実験によりその有効性を証明した.

まず類似した内容の Web 文書を同じグループにまとめる手法として, Web 文書クラスタリングについて論じた. Web 文書データのほとんどは非数値データで構成されていることから, 文書の内容をどのような単位で抽出するか, Web 文書ベクトルの類似度をどのように定義するのかが問題となっていた. そこで情報検索手法を用いて, 連続するカタカナや漢字を単語であると見なして Web 文書ベクトル生成した. Web 文書ベクトルにおける単語の分布の類似性からクラスタリングすることにより, 結果として口語体によるクラスタと文語体によるクラスタを得ることができた. 一方, ハイパーリンクの共起性に基づいてクラスタリングし, 同じリンク先を有する割合が高いほど Web ページ内容が類似しているという仮定に基づいて，二つのクラスタリングの結果を組合せることで, より類似した内容の Web 文書のクラスタを生成することができた.

次に, 階層的自動要約手法について論じた. 入れ子構造やリンク構造に着目することで, STU という意味的まとまりのある文章に Web 文書を分割した. そして STU を階層的に配置することで, 全体の内容を把握するならば上位階層から, より詳細な内容を求めるならば下位階層から内容を把握することが可能になった.

利用者がこの階層構造から求める情報を探すとき, 上位階層から下位の階層へと読み進めることとなる. こうした利用者の読解のしやすさという尺度を定量的に評価することのできるトラベルコストという評価尺度を提案した. また, 階層の各ノードの可読性や, 階層の可読性といった尺度の評価方法も提案し, 実験によりその有効性を示した.

また, 階層的自動要約手法の Web 情報検索への応用方法についても論じた. 従来の Web 情報検索の結果では困難になりがちな内容把握や, 問合せ語を含まない Web 文書を検索結果に含むことができない問題があった. 階層的自動要約手法を用いて検索結果を提示することで, 利用者が検索結果から合致する Web 文書への URL を探すときにこの階層構造は効果的に働くことを実験により示した. より詳細な内容は下位のノードの, 全体の内容を把握は上位ノードのラベルを手がかりと

してブラウズでき, 各ノード内の URL の内容を素早く把握することができる．そして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 URL も検索結果に含むことができた.

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

本研究を遂行し，まとめるにあたり，多くの方にお世話になりました．この場を借りて，感謝の意を述べさせていただきたいと思います．

指導教官である，法政大学工学部情報電気電子工学科　三浦孝夫教授には，日頃から数々のご指導，ご指示を頂きました．心からお礼申し上げます．

また，産能大学経営情報学部　塩谷勇教授には，本研究を進めるにあたり，格別の配慮を賜りました．心から感謝申し上げます．

データ工学研究室の先輩，同級生，後輩には，研究活動，学生生活の両方にわたり大変お世話になりました．

最後になりましたが，このような形で私の研究をまとめることができたのも，多くの皆様方のご支援ご協力の賜物であります．両親を始め，学生生活の中でお世話になったすべての方へ，この場をお借りしまして厚く御礼申し上げます．

参考文献

[1] J. Allan and J. Carbonell and G. Doddington and J. Yamron and Y. Yang.: Topic detection and tracking pilot study: Final report, In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998

[2] James Allan and Ao Feng and Alvaro Bolivar.: Flexible Intrinsic Evaluation of Hierarchical Clustering for TDT, Proceedings of the twelfth international conference on Information and knowledge management, 2003

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

[4] 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)

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

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

[7] 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).

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

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

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

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

[12] 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

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

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

[15] Mani, I.: Automatic Summarization, John Benjamins, 2001

[16] Mori, M., Miura, T. and Shioya, I.: Labeling Temporal Cluster of Web Pages, DBSJ Letters 3-2, 2004, pp.109-112

[17] Mori, M., Miura, T. and Shioya, I.: Extracting Events From Web Pages, proc. AISTA, 2004

[18] Mori, M., Miura, T. and Shioya, I.: Abstracting Temporal Clusters, proc. ITA, 2005

[19] Literature Kathleen Mckeown: Generating Patient-Specific Summaries of Online,1998

[20] Okumura, M., Mochizuki, H. and Nanba, H. : Query-biased Summarization Based on Lexical Chaining, In Proceedings of PACLING'99, pp.324-334, 1999.

[21] 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

[22] 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

[23] Radev, D., Jing, H. and M. Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, Information Processing and Management, 2004, pp.919-938

[24] Sakuma, M.:“ 要約文の表現類型 ”(1994).

[25] 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

[26] Stefanowski, J., Weiss, D.: Carrot2 and Language Properties in Web Search Results Clustering, Atlantic Web Intelligence Conference, 2003

[27] 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

[28] 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

[29] 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

[30] 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

[31] Sebastiani,F.: Machine Learning in Automated Text Categorization, proc.ACM Computing Surveys,Vol.34,No.1,2002 pp.1-47

[32] Trieschnigg, D. and Kraaij, W.: Scalable Hiearachical Topic Detection, SIGIR, 2005

[33] Takahiro Wakao, Terumasa Ehara, Katsuhiko Shirai.: Text summarisation for production of closed-caption TV programs in Japanese, Computer Processing of Oriental Languages Special issue on Information Retrieval on Oriental Languages (CPOL-IROL), No.4 1998.

[34] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen.: Meeting browser: Tracking and summarizing meetings, Proceedings of the Broadcast News Transcription and Understanding Workshop, p 281–286, 1998

[35] Zamir, O. and Etzioni, O.: Web Document Clustering – A feasibility Demonstration, SIGIR, 1998

[36] Zipf, G, K.: The human behavior and the principle of least effort, Addison Wesley, 1949