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

2006－ICS－142 (16)
2006／1／13

Web検索の効率向上のための未知検索語の重要説明文抽出方式

辻 泰希 †, 渡部 広一 ††, 河岡 司 ††

† 同志社大学大学院工学研究科知識工学専攻 †† 同志社大学大学院工学研究科知識工学専攻

現在，WWW(World Wide Web) の急速な発展により膨大な量の情報が誰でもオンライン上で入手可能となり，Web検索サービスを利用することで目的とする情報が瞬時に得ることが出来る．このように情報がオンライン上で瞬時に入手可能な状況は大変便利ではあるが，その一方必要な情報と不必要な情報が混ざった状態で検索されてしまうなど「情報の洪水」が問題となっている．そのため，現在より簡単にユーザが必要としている情報を獲得する手法として重要文抽出や複数テキスト要約が脚光を浴びている．

本稿ではユーザ自身がよく知らない新語や固有名詞などの語を Web を用いて調べる場合に，その語を端的に表している情報文を複数の Web ページから自動的に選別を行い，ユーザに提供する手法を提案する．この手法を用いることでユーザは Web をある種の辞書的な使い方で用いる場合に，不必要な複数の Web ページを見ることなく目的とする情報にたどり着ける事が可能となる．そして更に詳しい情報が知りたい場合には，提供された説明文から具体的な絞り込みのためのキーワードを得ることが出来る．また，提案手法を用いることで，雑音の少ない情報獲得が可能となり，Web からより高品質なコーパスの自動作成，知識ベースの自動作成などが可能になることが期待される．

Important and Explanation Setntences Extraction of Unknown Search Words for Efficient Web Searching

Yasuki TSUZI † Hirokazu WATABE †† Tsukasa KAWAOKA ††

† Graduate Student,Department of Knowledge Engineering and Computer Sciences, Doshisha University, Tatara 1-3, Kyotanabe, Kyoto,610-0394 Japan

†† Department of Knowledge Engineering and Computer Sciences, Doshisha University

Now there is a huge amount of information on World Wide Web, and we can get necessary information by using Web search service. On the other hand, "Flood of Information", retrieving with necessary and needless information mixed, became a serious problem. Based on such a background, "Important Sentence Extraction and Multiple Text Summarization" is paid to attention as a technique for acquiring information.

In this paper, it proposes the technique, when the user examines the word that the user doesn't know well with Web , for automatically selecting the information sentences explaining the word on Web pages and offering it to the user. On using Web as a dictionary, this technique enables obtaining necessary information without seeing needless Web pages. And, to know more detailed information, the user can get the key words for narrowing from the explanation.

1 はじめに

現在，WWW (World Wide Web) の急速な普及と利用者の劇的な増加に伴い，膨大な電子化された文書がオンライン上に蓄積されている．その膨大な電子化された文書は，新聞記事に留まらず，電子辞書や Blog に代表されるような多種多様な文書が電子化され，すでに存在しており一般的に必要とされるような情報は Web 上に確実にあると言っても過言ではないような状況になってきている．しかしながら，既存の検索システムではユーザが必要とする情報だけを手に入れることは困難であり，このような「情報の洪水」を背景に，的確にユーザが必要としている情報のみを獲得するための技術が現在求めれている．

そのための主たる技術として検索，テキスト自動要約，情報抽出などがあげられる．特に Web のように複数の文書から情報を獲得するためには，複数文書にわたる検索と情報抽出が必要である．

テキスト自動要約では文書中から重要な箇所を必要な量だけ抜き出す重要文抽出が基本であるが，複数の文書を要約する場合にはテキスト集合の中に同じ内容の文（センテンス）が含まれている可能性があり，重要な箇所を含みつつも内容の重複した文を特定し，必要な重要文のみを抽出する必要がある．

本稿ではユーザ自身がよく知らない語を Web を用いて調べる場合に，その語を端的に表している情報文を複数の Web ページから自動的に選別を行い，ユーザに提供する手法を提案する．例えば検索語「イチロー」であれば「メジャーリーグシアトル・マリナーズの日本人外野手．84 年間破られることのなかったジョージ・シスラーの 257 安打を 5 本上回る 262 安打を記録.」の様なイチローを説明している情報文を提供することを目的とする．

提案手法は，自動的にWebから複数の文書を取得し，その中から重要な説明情報文を獲得する．提案手法では「重複している部分が重要である」という考えと「その文書内容を特徴づける語句が多数含まれている部分が重要である」という考えに基づいて，複数テキストからの情報抽出を行うことで重要な説明文の獲得を実現した．この手法により，ユーザが検索時に不必要なWebページを見ることなく目的とする情報にたどり着ける効率的なWeb検索の実現や，ノイズが少ない情報獲得が可能になることからWebからより高品質なコーパスの自動作成，知識ベースの自動作成などが可能になることが期待される．

2 複数テキストからの情報抽出の概要

2.1 複数テキストからの情報抽出

複数テキストからの情報抽出の研究は以下の3つの技術がポイントに挙げられている[1)]

- 関連するテキストの自動収集
- テキスト間の文体の違い考慮した要約文書の作成
- 関連する複数テキストからの情報抽出

2.2 関連する複数テキスト要約・情報抽出研究

柴田[2)]らは，複数の情報源から得られた特定の話題に関する複数の記事を解析対象として，「複数記事の共通箇所を抽出することが関連記事の重要箇所を抽出すること」であるという考えに基づいている．手法としては，形態素の出現頻度を利用して重複文(テキスト間の共通点)の同定を行い，特定した重複文の1文のみを用いて要約作成を行っている．本手法も柴田らが提案した「記事間の共通箇所を抽出することが関連記事の重要箇所を抽出すること」と言う考え方はほぼ同じではあるが，柴田らが形態素の表層的な面からしか重複文の同定を行っていないのに対し，提案手法では意味内容を考慮に入れて重複文の同定を行っている点に特徴がある．また，柴田らは情報空間を限定して行っているのに対し，提案手法ではWeb情報空間全体を対象としている点も大きく異なっている．

本手法の特徴は，形態素解析の情報のみの表層的な手法で重複箇所を探索するのではなく，概念ベースと関連度計算を用いることで意味内容を考慮に入れ重複箇所を探索することである．また，事前情報無しに情報の選別を行う点も本手法の特徴である．

3 情報抽出に用いた手法

3.1 TF・IDF

TF・IDFによる重み付け[3)]とは，語の頻度と網羅性に基づいた重み付け手法である．TFは頻度情報を用いて適切な情報を集めたものであり，IDFは特定性情報を用いて特徴付けた情報を集めたものである．

属性 a_j の文書 D_i における重み $w_i(a_j)$ は，以下の式で定義される．

$$w_i(a_j) = tf_i(a_j) \times idf(a_j) \quad (1)$$

$$tf_i(a_j) = f_{ij} \quad (2)$$

$$idf(a_j) = 1 + \log(N/n_j) \quad (3)$$

f_{ij} は a_j の文書 D_i における出現頻度，N は検索対象となる総文書数，n_j は a_j が現れる文書数である．

3.2 Web-IDF

Web-IDF[6)]とは3.1で説明したIDFをWeb上での特定性情報に基づいて算出する重み付けである．検索対象文書数 N をGoogleが保有している日本語のページ数とする．ただし，Googleが全言語において保有しているページ数は公開されているが，日本語のページとして保有している数は公開されていない．そこで日本語の文書として最も使われている「は」で検索を行ったヒット件数(416,000,000)をGoogleが保有している日本語の全ページ数[7)]とし，検索対象文書数 N とする．索引語 t が出現する文書の数 n_j は，索引語 t をGoogleで検索を行った時のヒット件数とする．

3.3 概念ベース

概念ベース[4)]とは，複数の国語辞書や新聞等から機械的に構築した，語(概念)とその意味特徴を表す単語(属性)の集合からなる知識ベースである．概念と属性のセットにはその重要性を表す重みが付与されている(式4)．概念ベースには，現在約9万語の概念が収録されており，1つの概念あたり平均30個の属性が存在する．しかしながら，概念ベースにも登録されていない語も存在しており，その語を以下本稿では未定義語と定義する．

$$A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_n, w_n)\} \quad (4)$$

各概念に付与されている属性は，概念ベースに概念表記として登録されている語で構成されるため，各属性を一つの概念表記としてみなした場合，さらにそれを表す属性を導くことができる(Fig.1)．このように，概念は概念ベースにより n 次の属性連鎖集合として定義する．また，n 次の属性集合を n 次属性と呼ぶ．

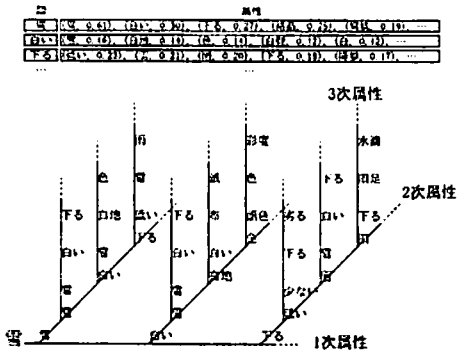

Fig. 1 概念ベース

3.4 関連度計算

2つの概念がある時，概念ベースを利用することによってそれらの概念間の関連度[4]を求めることができる．関連度は概念間の意味的な関連の深さを0〜1の実数値で表すものである．2つの概念の関連度は，一致度の和が最大になるような一次属性の組み合わせを作ることによって計算する．一致度は，それぞれの概念を二次属性まで展開し，一致する属性とその重みによって求める実数値である．

概念ベースに定義された任意の概念 A と概念 B を

$$A = \{(a_i, u_i)|i = 1 \sim L\}$$
$$B = \{(b_j, v_j)|j = 1 \sim M\}$$

とすると，概念 A と概念 B の一致度：$MatchW(A, B)$ は次式で表される．

$$MatchW(A, B) = \frac{\left(\frac{s_A}{n_A} + \frac{s_B}{n_B}\right)}{2} \tag{5}$$

$$s_A = \sum_{a_i = b_j} u_i \ , \ s_B = \sum_{a_i = b_j} v_j \tag{6}$$

$$n_A = \sum_{i=1}^{L} u_j \ , \ n_B = \sum_{j=1}^{M} v_j \tag{7}$$

さらに，一致度から算出される概念 A と概念 B の関連度 $ChainW(A, B)$ は，

1. 属性の少ない方の概念を A とし $(L \leqq M)$，概念 A の属性を基準とする．

$$A = \{(a_1, u_1), (a_2, u_2), \cdots, (a_L, u_L)\}$$

2. 概念 B の属性を，概念 A の各属性との一致度 $MatchW(a_i, b_{xi})$ の和が最大になるように並び替える．

$$B = \{(b_{x1}, v_{x1}), (b_{x2}, v_{x2}), \ldots, (b_{xL}, v_{xL})\}$$

ただし，対応にあふれた概念 B の属性 $\{b_{xj}|j = L+1, \cdots, M\}$ は無視する．

3. 概念 A と概念 B の関連度を以下に示す式で定義する．

$$ChainW(A, B) = \frac{\left(\frac{s_A}{n_A} + \frac{s_B}{n_B}\right)}{2} \tag{8}$$

$$s_A = \sum_{i=1}^{L} u_i MatchW(a_i, b_{xi}) \tag{9}$$

$$s_B = \sum_{i=1}^{L} v_{xi} MatchW(a_i, b_{xi}) \tag{10}$$

$$n_A = \sum_{i=1}^{L} u_i \ , \ n_B = \sum_{j=1}^{M} v_j \tag{11}$$

4 提案手法

4.1 目標と概要

本稿ではWeb情報空間から検索語1語で検索を行い，複数のページを獲得し，その検索語を説明するに過不足のない情報文の抽出を行うことを目的としている．

提案手法の概要は，自動的に多数のテキストページをWebから収集することで網羅的に情報を集める(4.2)．集めてきたテキスト情報にはもちろん重要な情報が含まれている可能性も高いが，全く関係のない雑音と言えるような情報も含まれている可能性も非常に高い．そのため，集めてきたテキスト情報の中から検索語に関連する情報文の選別を行いノイズの少ない情報文のみを残す(4.5)．重要箇所を「重複している部分が重要である」という考えに基づき集めてきたテキスト集合内において，文（センテンス）単位の意味内容距離情報を元に文のクラスタリング(4.7)を行い，まとまりが大きなクラスタの内容ほど重要であるとする(4.8)．つまり意味内容が同じ文が多数出現するほど重要であると考える．それぞれのクラスタ内全ての文を採用すると冗長となるため，各クラスタ内から代表文を選び出し，それを重要説明文と定める(4.10)．

4.2 テキスト自動収集方法

以上のような目的を達成するためには，前提としてまず重要な情報が含まれているテキストページを獲得しなければならない．その為には，情報を網羅的かつ偏りなく集める必要がある．提案手法では「一般性」「正確性」「時事性」の観点から検索を行いテキストの自動収集を行う．その後，各文書を改行・句読点・空白情報を元に文に分割を行う．

一般性 Web上の一般サイトと検索にはGoogle[11]を用いて検索結果の上位10ページから情報を収集する．通常のWeb検索エンジンの特徴として，新旧にとらわれない情報を集めることが出来る．

正確性 オンライン辞書サイト内で検索することで情報を集める．本稿ではオンライン上のフリー辞書サイトのウィキペディア[12]を利用した．ウィキペディアは誰でも編集可能なユーザ参加型のフリー辞書であり，2005年現在約1万の記事が投稿されている．

時事性 複数の新聞社がオンラインで提供している記事を検索することで最新の情報を集める．本稿ではGoogle News BETA[13]を利用し，検索結果の上位30ページの情報を獲得することとした．

4.3 未定義語の属性獲得手法

未定義語 X の意味的特徴を表す属性（単語）とその重要性を表す重みの組をWebを用いて50組自動的に構成[8]する．まず4.2で獲得したテキスト情報から形態素解析[10]を行い自立語を出現単語とする．その後，獲得したテキスト情報空間内での出現単語の出現頻度とWeb-IDFの算出を行い，TF・Web-IDF重み付けを行う．重み上位50の自立語とその重みの対の集合を X の属性とする．この手法を用いて検索語の属性とその重みの組を構成する．検索語 X の属性は式12のように構成される．

$$X = \{(x_1, w_1), \cdots, (x_{50}, w_{50})\} \quad (12)$$

4.4 文構成語の重み付け手法

本稿では文構成語 Si の重み付け手法として2種類の重み付けを行った．次に文（センテンス）Sen は，以下の式で定義される．

$$Sen = \{(s_1, w_1), \cdots, (s_M, w_M)\} \quad (13)$$

文を形態素解析を行って得られた自立語を構成語とする．文構成語 Si の重み付け手法の1つ目としてTF・Web-IDF重み付け手法を用いた．

$$w_i = TF_{(Si)} \times Web\text{-}IDF_{(Si)} \quad (14)$$

普段使っている"野球"と言う語の重みと，検索語「イチロー」からみた"野球"の重みは大きく違うが，TF・Web-IDFでは一般的観点からの重み付け手法で，検索語の観点は反映されていない．そこで，検索語の観点が反映されるような重み付け手法を提案する．

$$w_i = TF_{(Si)} \times Web\text{-}IDF_{(Si)} \times ChainW(X, Si) \quad (15)$$

式15はTF・Web-IDF重みと文構成語 Si と検索語 X の関連度積を表しており，検索語の特徴を重みに反映することが出来る．以下この重み付け手法をTF・Web-IDF・Chain重み付け手法とする．

4.5 一次選別

4.3で構成した検索語の属性群を用いて，獲得した複数の文から無関係な文を取り除く処理を本稿では一次選別と呼ぶ．

入力キーワード X と獲得したテキストから得られた文 Sen との関連性を先ほど定義した式12，13と関連度計算式（式8）を用いて算出し，閾値以上の文のみを取り出す．

4.6 文同士の意味的距離

文同士の意味的距離とは，文 A と文 B の意味内容がどれだけ近いかを示す指標である．一般的には2つの文に同一単語が多数出現する場合，2つの文の意味的距離が近いと考えることが出来る．一般的に用いられる計算方式がベクトル空間モデルと用ばれる方式で4.6.1で詳しく説明を行う．しかしながら，ベクトル空間モデルでは表記が完全に一致した部分でしか計算することが出来ない．そこで，提案手法では概念ベースを用いることで，表記が異なった場合でも意味内容を考慮する文間意味関連度計算方式を用いた．

4.6.1 ベクトル空間モデル

ベクトル空間モデル（Vector Space Model）はSalton[5]らにより提案され，現在情報検索分野において幅広く利用されている技法である．出現単語に基づいて文書あるいは文を1つのベクトルで表現し，ベクトルの向きによって内容を判断する点が特徴である．文を形態素解析を行って得られた自立語を出現単語とし，重み付けにはTF・Web-IDF，TF・Web-IDF・Chain重み付け手法をそれぞれ用いた．

文（センテンス）は，4.4で定義した式eqn:senで定義される．

Sen は語 s_i とその重み w_i の対の集合であるが，計算の対象群に含まれるすべての語の数を M とすると，文 Sen は，M 次元の重みベクトルとして，以下の式で定義される．

$$Sen = ((w_1, w_2 \cdots, w_M) \quad (16)$$

ある文 $SenA$ とある文 $SenB$ の類似度 $Vector(SenA, SenB)$ は，以下の式で定義される．

$$Vector(SenA, SenB) = \cos\theta \quad (17)$$

$$= \frac{SenA \cdot SenB}{|SenA||SenB|} \quad (18)$$

4.6.2 文間意味関連度

提案する手法では，概念間の関連度を計算する手法を用いて文と文の関連性を計ることが出来る．文を概念ベース中の概念と同じように，属性とその重みの対の集合という形式で表す事により，提案されている関連度計算方式を利用することが出来る．文（センテンス）間の関連度のことを文間意味関連度と呼ぶ．文の属性には，文に出現する単語を用いる．まず，文を形態素解析することによって得られる単語群の中から自立語を抜き出す．これは文の意味内容を表す属性としては自立語が適切だと考えられるからである．属性に概念ベースに登録されていない語である未定義語が出現した場合には，その表記情報だけを利用し，一致する属性が他方に有れば一致度を 1 とすることでより正確な文間意味関連度を求めることが出来る．

文 $SenA$ と文 $SenB$ の文間意味関連度は式 8 に $SenA$ と $SenB$ の属性列を用いて計算を行う．

4.7 同一意味内容文の特定

クラスタリング手法を用いて，意味内容で獲得した文を分類する．クラスタリングの手法として単純クラスタリング手法を採用した．N 個の文全てが別のクラスタに属する初期状態からスタートし，あらかじめ定義されたクラスタ間の距離関数 (式 8) に基づいて，閾値以上の 2 つクラスタを併合してゆき，最終的に k 個のクラスタを得る方法である．

4.8 文の重要度の定量化

本稿では文 $SenA$ の重要度 $SenW(SenA)$ をその構成属性の重みの総和とし，式 19 のように定める．

$$SenW(SenA) = \sum_{i=1}^{M} w_i \qquad (19)$$

M は文 $SenA$ の構成属性数，w_i は 4.4 で定めた構成属性のそれぞれの重みを示している．

4.9 意味内容の重要度の定量化と判別

同一意味内容で構成された各クラスタの重要度は，所属する文の総数となる．このクラスタの重要度が情報の重要度となり重要箇所を判別することが出来るようになる．クラスタ Ca の重要度は以下の式で定義される．

$$ClusterW(Ca) = Sizeof(Cluster(Ca)) \qquad (20)$$

次に重要性の高い意味内容のクラスタだけに絞り込みを行う．絞り込みには実験的に求めた閾値以上の重要度を持つクラスタを採用することとした．

4.10 代表文抽出

4.9 で重要意味内容と判断したクラスタの全ての文を採用すると重複・冗長となるので，各クラスタからそのクラスタの意味内容を代表する文を抽出する．文 $SenA$ が所属するクラスタ Ca での $SenA$ の重要性 $Score(SenA, Ca)$ を以下の式で定義する．

$$Score(SenA, Ca) = SenW(SenA) \times AveRel(SenA, Ca) \qquad (21)$$

$$AveRel(SenA, Ca) = \frac{1}{N}\sum_{i=1}^{N} SenChainW(SenA, Sen(i)) \qquad (22)$$

式 22 は文 $SenA$ が所属するクラスタ Ca に含まれている全ての文それぞれとの文間意味関連度の平均を示し，式 21 は文 $SenA$ の重み $SenW(SenA)$ と平均関連度の積を示しこれを重要度 $Score(SenA)$ とする．各クラスタで $Score(SenA, Ca)$ が最大となる情報文をそのクラスタの代表とし重要説明文として出力する．

5 評価方法

提案手法で獲得した重要説明文の評価を行う．しかしながら自動要約・情報抽出研究分野において，システムの出力である重要文や要約をどのように評価するかという問題に関しては，分野内においても明確な基準がなく評価は難しい問題とされてきた．現在でも日本語テキストの要約に関する共通の評価方法や評価基準の明確化が議論されているが未だ明確な基準は定義しきれていない．

そこで本稿では次のような評価方法を提案し，評価を行った．

5.1 重要説明文の定義

本稿では重要説明文を以下のように定義する．

- 検索語と説明文の間に必要十分条件を満たす文

次に検索語として「イチロー」を例に挙げ具体的に説明を行う．

A：イチロー：日本人である
B：イチロー：野球選手である
C：イチロー：シアトル・マリナーズの日本人外野手である

A〜C までの各情報文はイチローに関する情報文である．しかしながら，A は間違ってはいないが，適切な説明文とは言えない．次の B は A よりも説明文として確実に正しいと言えるが，必ずしも「野球選手」であることから「イチロー」と導くことは出来ない．C は「シアトル・マリナーズの日本人外野手」と言えば一意的に「イチロー」と言う言葉を導くことが出来，最も端的に検索語を説明していると言える．

このように検索語と説明文の間に必要十分条件を満たせる文が説明文として最も相応しい説明文と考えることが出来る．

5.2 目視評価

得られた重要説明文全てを人手で以下のポイントに着目し，評価を行う．手法から得られた情報文の中に適切な重要説明文が含まれている割合で評価する．

- 適切な重要説明文
  - 必要十分条件を満たすか？
  - ある範囲以内に特定するのに十分な情報を含んでいるか？
  - 不適切でないか？
- 重複箇所・冗長部の有無
  - 意味が同じ内容の文が出現していないか?
  - 端的な説明か？

この評価法の特徴としては人間が全てを評価するので，最も正確な抽出評価を行うことが出来るが，全ての作業が人手であるために非常に評価コストが高い．そのため，本稿では自動評価法を用いて評価を行った後に，目視評価を行うことにより低コストで正しい抽出評価が出来ると考える．

5.3 再検索適合ページ割合

5.1 で定義した重要説明文の定義を満たす説明文を Web 検索を利用して自動的に評価を行う．これは重要な情報ほど検索語との十分条件を満たすことに着目している．評価の流れを Fig.2 に示す．

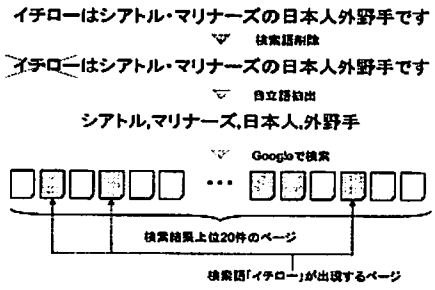

Fig. 2 再検索適合ページ割合

文 $SenA$ から検索語を取り除き，自立語を抽出する．抽出した自立語を用いて検索を行い，その検索結果の最大上位 20 ページを獲得する．その最大 20 ページ内で検索語が再び現れているページを適合ページとし，その割合を求める．この評価方法を本稿では再検索適合ページ割合とする．文 $SenA$ の再検索適合ページ割合 $AutoEst(SenA)$ を式 23 に定義する．

$$AutoEst(SenA) = \frac{ConformPageNum}{GetPageNum} \quad (23)$$

$ConformPageNum$ は検索語が含まれるページの数，$GetPageNum$ は検索結果上位順に獲得したページ数．この評価方法の特徴としては，検索時に絞り込むための自立語は多いほど優位であるが，逆に多すぎると検索結果が 0 ページとなり再検索適合ページ割合は 0 となる．このように再検索適合ページ割合では冗長な説明文も評価可能な点である．

6 評価実験

提案手法を実験を行い他種法と比較し評価を行う．実験のためにあらかじめ 13 分野・34 語の検索語からなるテストセットを作成した．以下にテストセットの一部を載せる．

[人名]:田中角栄，大谷吉継 / [病名]:PTSD，パーキンソン病 / [映像]:メガスター / [商品名]:リニモ / [国名・地名]:白神山地

Fig. 3 テストセット一例

提案手法のシステムは検索語から複数の重要情報文を出力場合もあれば，全く重要情報文が出現しない場合もある．そこで検索語 X に対するシステムの自動評価 $AutoEst(X)$ は出力重要情報文の再検索適合ページ割合の平均とし以下の式で定義する．

$$AutoEst(X) = \frac{\sum_{I=0}^{M} AutoEst(SenI)}{M} \quad (24)$$

M は得られた検索語から得られた重要説明文の総数，$AutoEst(SenI)$ は各重要説明文の再検索適合ページ割合を表している．また M=0 となる検索語の再検索適合ページ割合 $AutoEst(X)$ は 0 とする．

テストセットから導かれる提案手法の自動評価の精度は次の式で定義する．

$$AveAutoEst = \frac{\sum_{X=1}^{34} AutoEst(X)}{TestSetSize} \quad (25)$$

X はテストセットの語，$TestSetSize$ は語数の 34 となる．テストセットに対する自動評価を以下，平均再検索適合ページ割合とする．

実験の評価は自動評価を用いて評価を行い，適時目視評価を行う．目視評価はテストセットから得られた重要説明文に対し，適切な重要説明文が含まれている割合とする．

6.1 閾値による変化

6.1.1 獲得テキスト情報のノイズ文除去実験

4.5 で提案した手法を用いて無関係な文を判別する (検索語の属性を利用) 実験を行い，平均再検索適合ページ割合と出力文数の関係を調べた (Fig.4)．実験環境として，検索語に対して平均約 960 文の情報文を Web から得ることが出来，この中から重要説明文の抽出を行う．重み付けは TF・Web-IDF 重み付けを行った．

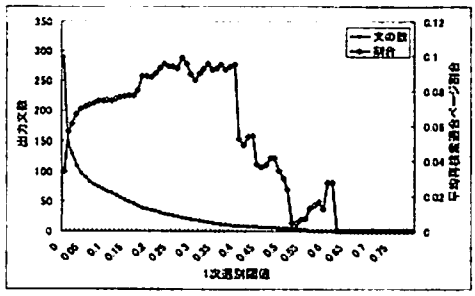

Fig. 4　一次選別における精度と文数の関係

閾値が 0.29 付近で平均再検索適合ページ割合はピークに達するが，残される文の数は約 32 文と非常に少ない．逆に 0.01 では平均再検索適合ページ割合が低い代わりに多数の文を得ることが出来る．0.1 付近は平均再検索適合ページ割合は多少良くなり，また文の数も約 220 文と比較的多い．つまり，「文数は少ないが良い文が残っている状態」，「文の数は多いがあまり良くない文も混ざっている状態」，「中間の状態」3 つの状態に分けることが出来る．

6.1.2　クラスタリング距離とクラスタ採用サイズ

次に 3 つのそれぞれでの状態からクラスタリングを行い代表文を抽出する実験を行った．(4.7-4.10)．クラスタ距離 (同一意味とみなす意味的距離の閾値) とクラスタ選別サイズ (クラスタの重要度 $ClusterW(Ca)$) は，代表文の再検索適合ページ割合が最高となる組を実験的に調べ用いた (Fig.5)．

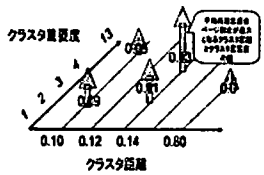

Fig. 5　2 つのパラメータ変化による精度

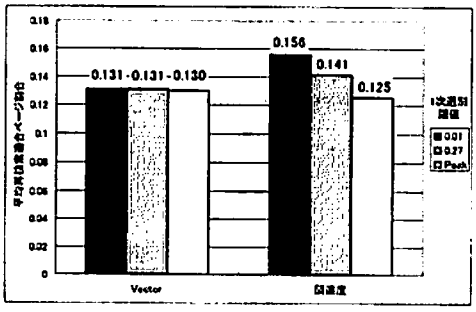

Fig. 6　1 次選別の状態とベクトル・関連度の評価

実験の結果 (Fig.6)，1 次選別の 3 種類のいずれの状態から同一意味内容を特定し重要説明文を獲得しても，文間意味関連度を用いた場合の方がベクトル空間モデルを用いた場合より良い結果を得られた．また，状態は文が多数ある状態から重要説明文を獲得するのが最も良かった．

6.2　重み変調ならびに他手法との比較実験

2 種類の重み付け手法を比較する．1 次選別の閾値は 0.01 を用いた．実験の結果 (Fig.7)，重み付け手法は TF・IDF・Chain 重み付け手法を利用し文間意味関連度を用いた場合が最も良かった．

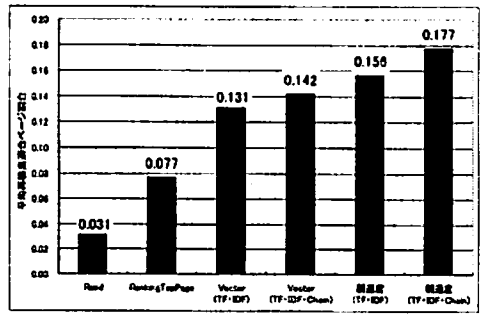

Fig. 7　重み付け比較と他手法の評価

Rand は各テスト語からランダムに重要説明文を 3 文出力し評価した．RankingTopPage は検索語で検索を行い，検索結果の最上位に出現したページ内全ての文を重要説明文として評価を行った．

6.3　目視評価

提案手法ならびに他手法の目視評価実験を行い，比較を行う (Table 1)．評価には 3 名の評価者を用意し，2 名以上が重要説明文だと判断した文を適切とした．提案手法から得ることが出来た重要説明文の一例を下に示す．

- 田中角栄
  - ロッキード事件は世界数カ国にまたがる、とてつもない広がりをもった前代未聞の航空機商戦の汚職疑惑なわけです。
  - 日中国交の道を再び開いた総理大臣。
  - 本県の政治風土にも詳しい高畠通敏・立教大法学部教授 (65) ＝政治学＝に聞いた。
  - 徳島県出身。
- 軍艦島
  - 軍艦島の模型１／２０００
  - 正式な島名は「端島」であるが、通称である「軍艦島」のほうが知名度が高い。
- ポリウォーター
  - 出力無し

Table 1　目視評価結果

手法	適切割合 (適切文数/総数)
提案手法	0.37(33/89)
Vector	0.29(28/96)
RankingTopPage	0.17(6/36)
Rand	0.07(7/102)

6.4　評価のまとめ

実験では自動評価においてベクトル空間モデルの1.35倍，ランダム出力の約5倍の再検索適合ページ割合が得られた．また，目視評価においてもベクトル空間モデルの1.28倍，ランダム出力の5.3倍の割合で適切な重要説明文が得られることが分かった (Table 2).

Table 2　評価のまとめ

手法	目視評価	自動評価
提案手法	0.37(33/89)	0.17
Vector	0.29(28/96)	0.13
RankingTopPage	0.17(6/36)	0.08
Rand	0.07(7/102)	0.03

7　考察

7.1　文間意味関連度の考察

提案手法では概念ベースを用いることにより，表記情報の一致だけのベクトル空間モデルを用いる手法よりも意味内容を考慮した文と文の間の意味的距離を求める手法を提案した．

A：イチローはヒットした。
B：イチローは安打した。
C：イチローは三振した。

この手法により文AとBの関連度を意味内容を考慮したことで，表記一致よりも正確に計算することが可能になった．しかし，従来の表記情報のみでは，文AとCの関係度は低い値だったが，概念ベースを用いることで"ヒット"と"三振"が関連があると判断されてしまい，誤った値を出してしまう可能性を考慮する必要が出た．

そのため，今後は語の意味内容だけではなく，文(センテンス)全体の意味を考慮に入れた方式を考案する必要がある．

8　おわりに

本稿では，Webから検索語の説明文を抽出する手法を提案した．提案手法によりWebの情報を用いるようなシステムにおいて1番の問題とされる，"無関係な情報の混在"の問題を解決する道しるべを示すことが出来た．今後は"Web情報を使った質問応答システム"の情報源として利用可能であると考える．

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行った．

参考文献

1) 奥村学, 難波英嗣："テキスト自動要約に関する最近の話題"自然言語処理", Vol.9, No.4, pp.97-116 (2002)

2) 柴田昇吾, 上田隆也, 池田裕治："複数文章の融合", 情報処理学会研究報告, 1997-NL-120, Vol.1997, No.069, pp.77-82 (1997)

3) Salton,G. and Buckley,C. : "Term-weighting approaches in automatic text retrieval", *Information Processing & Management*, Vol.41, No.4, pp.513-523, (1988)

4) 渡部広一, 河岡司："常識的判断のための概念間の関連度評価モデル", 自然言語処理, Vol.8, No.2, pp.39-54, (2001)

5) G.Salton and M.J.McGill. : "Introduction to Modern Information Retrieval", McGraw-Hill Advanced Computer Science Series(1983)

6) 辻泰希, 渡部広一, 河岡司："wwwを用いた概念ベースにない新概念およびその属性獲得手法", 人工知能学, 2D1-01, (2003)

7) 大森貴博, 笹塚清二, 近藤晶子, 水谷正大, 来住伸子, 小川貴英："統計的推定による日本語Webの調査", 情報処理学会第59回全国大会講演論文集3, pp.79-89 (1999)

8) Inderjeet Mani, 奥村学, 植田禎子, 難波英嗣："自動要約", 共立出版, 2003,ISBN: 4320120736

9) 奥村学, 難波英嗣："テキスト自動要約", オーム社,2005,ISBN:4274200426

10) 松本裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村友明："日本語形態素解析システム『茶筅』version1.0 使用説明書", NAIST TechnicalReport, NAIST-IS-TR97007 (1997)

11) Google http://www.google.co.jp/

12) ウィキペディア http://ja.wikipedia.org/

13) Google News BETA http://news.google.co.jp/