情報学基礎 64－9
自然言語処理 145－9
（2001．9．11）

展望台システムによる複数文書の要約とWebページ集合への適用

砂山 渡[†] 谷内田 正彦[†]

情報の電子化と多様化が進む現在の世の中において，誰もが必要な情報を素早く獲得することを望んでいる．我々の限られた時間と肉体を用いる中で，多くの良質な情報に目を通すための環境が不可欠となってきた．そこで本論文では，重要文抽出法である展望台システムを用いて複数の文書の要約を試みる．展望台システムは，文章の主題とのみ関わる文章を特徴づけるキーワードを発見して重要文を抽出するシステムである．しかし，文の出現位置や出現順序を文の評価に用いない手法であるため，複数の文章を併せて一つの文章と見立て，複数文書の要約を行なう．

Extracting Key Sentences from Multiple Documents by Panoramic View System and An Application to Web pages

WATARU SUNAYAMA[†] and MASAHIKO YACHIDA[†]

We'd like to see many documents in order to get useful one from plenty of electronic information. Now, the environment for information retrieval is necessary because of our restricted time and power. In this paper, we suggest a system which summarize multiple documents by Panoramic View System(PVS) a method for key sentences extraction. PVS is a system which discovers feature keywords that only related to the topic of a document, and output key sentences. Since PVS doesn't use position information and order information of sentences, the system regards multiple documents as a single document and summarize those documents.

1. はじめに

情報の電子化と多様化が進む現在の世の中において，誰もが必要な情報を素早く獲得することを望んでいる．我々の限られた時間と肉体を用いる中で，多くの良質な情報に目を通すための環境は不可欠である．そのため，Webページ，電子掲示板，ネットニュース，電子図書館など電子的に存在する多くの情報の中から，必要な情報のみを取り出す様々な手段が考えられてきた．その最も代表的な情報抽出のためのシステムがサーチエンジンである．

サーチエンジンはユーザが入力した検索語を含むWebページを出力するため，全世界のWebページ全体を一つの文章と見なした時に，検索語を観点としたWebの要約を与えるシステムとしても捉えられる．すなわち，検索結果は一つの要約であり，ユーザは再検索によって，自身の望む要約を得るための試行錯誤を繰り返している．この再検索支援は現在最も盛んな研究分野の一つとなっており，検索結果のクラスタリング，関連フィードバックによる検索語の自動補完，検索語の関連語の提供とそれに伴う情報の視覚化などが例として挙げられる．しかし，最終的に提示される情報を読むのはユーザ自身であり，その読まれる情報とは文章に他ならない．したがって，検索の性能向上と同様に検索結果を説明する文章の改善が必要であり，ユーザが入力した検索語に関連して重要文を抽出する展望台システム[10]はその流れに沿ったシステムである．本論文では，この展望台システムを用いて複数文書[☆]の要約を試みる．

複数文書を要約する目的としては，

1. 複数文書の共通点（大勢を占める話題）の把握
2. 複数文書の異なり（少数意見）の把握

の2点が挙げられる．1.の文書間の共通点の把握という点に関していえば，個々の文書から高頻度語を取り出し[4]，多くの文書に共通のキーワードを取り出す[9]ことで実現できる．しかし2.においては，キーワードの選び方が多種多様であり，いまだ決定的な手法は存在しない．例えば文書中に現れる全単語のうちで高頻

† 大阪大学大学院基礎工学研究科
Graduate School of Engineering Science, Osaka University

☆ 本稿において「文書」は文字によって書かれた「もの」を表し，「文」とは文章中の「1文」を表す用語として，逆に「文章」は「文の集合」を表す用語として用いる．

度の単語はひと握りであり，およそ８割の単語の出現頻度は１回のみである．すなわち高頻度の単語を含む話題以外にも，低頻度の単語を効果的に含めた文書の要約として，文書集合全体の主旨から大きく逸れない重要文を取り出すことが望まれる．展望台システムはこれを実現するための次の２つの要素をもっている．

1. 文書の観点となるキーワードに関わる重要文を取り出す．
2. 文書の観点となるキーワードとのみ関連して文章を特徴付けるキーワードを取り出す．

この 1. によって，出力される重要文全体に統一感をもたせることができると共に，2. によって頻度の低い単語をも積極的に評価して，文独自の内容が述べられている重要文を抽出することができる．

文章や，文独自の内容を評価する手法として tfidf 法[1] がある．しかし，この方法によるキーワードは文章の主題に関わらず取り出されるため，この手法のみを用いた場合では，最終的に出力される要約が無意味な文の羅列になる可能性が高い．

本論文では，以下の 2で複数文書の要約に関する仮定について述べ，3で提案するシステムの概要とアルゴリズムの詳細について述べる．4で文章から重要文を抽出する実験・評価を行い，5で本論文を締めくくる．

2. 複数文書の要約について

本研究で用いる展望台システムは，単語の意味情報や文脈を全く扱わず，文章を記号のデータ列と見なして処理を行なうために，単一文章から重要文抽出を行なうのと同様の手続きによって，複数文書からの重要文抽出を行なうことができる．同様に文章を記号のデータ列として扱う研究として，KeyGraph[6] があるが，KeyGraph で最終的な出力である文章の主張となるキーワードを展望台システムにおいては入力として扱える点や，KeyGraph では文章全体が主題を導くという構造を仮定していることから，雑多な文書集合に対しての適用は困難と考えられる．従来の複数文書の要約手法には，同じ主題をもつニュース記事を扱う研究[8] や，同じ分野の論文をまとめる研究[5] などがある．しかし，対象とする文書間に１つの主題があると仮定し，その主題に関して要約を生成する手法であったり，文書内の他文書への参照情報を利用する方法であるため，これらの仮定が成り立たない文書集合へは適用はできない．これに対し本論文では，雑多な Web ページ集合を要約する手法を提案する．Web ページ集合を要約する先行研究[3] もあり，Web ページ集合をクラスタリングした上で，統計手法によって各クラスタにラベルとしての要約を付与する手法を試みている．しかし，クラスタリングによって生成される各クラスタは，ユーザの興味とは無関係に生成されるため，ユーザが望む情報を得るための確かな手がかりになるとは限らない．

単一文書の要約に比べ，複数文書の要約では以下の３点に関する議論が必要となる．

1. 入力文書集合に対する仮定
2. 重複する内容の取り扱い
3. 要約文の提示方法

まず，1. の点に関していえば，文書集合にはラベル（文書間の共通点，文書を集める際に用いた知識）が存在している．それらは例えば，

a. 同じ主題をもつ文章集合
b. 同じキーワードを含む文書集合
c. 同じ分野における様々な主題をもつ文書集合
d. 同じ作者による様々な主題を含む文書集合

などである．Web 上のデータをもとに具体例をあげると，a. は一つのニュースを報じる様々なサイトからの記事の集合，b. は検索エンジンの検索結果として得られる Web ページの集合，c. は電子掲示板やニュース，d. は個人サイトの日記ページが該当する．展望台システムにおいては文書の主題を表す文を重要文として抽出するため，文書集合に共通の主題があるかないかによってシステム構成が２つに大別される．

すなわち 2. の点に関して，重複する内容の有無は，文書集合における共通の主題の有無と密接に関連する．共通の主題がある場合には文書間での重複部分も多くなるため，最終出力となる要約においては重複部分を削除する必要性が生じる[7]．展望台システムでは，各文は文に含まれるキーワードの評価値合計で評価されるため，評価値の高い文が一つあった時に，それと全く同じ別の文にも同等の評価値が与えられてしまう．そのため，必要に応じてこれを削除する操作が必要となる．

最後に３の出力する重要文の提示方法に関して，出力する重要文の数とその提示順序が重要となる．展望台システムにおいては，文間のつながりを考慮していないため，出力される要約は重要項目の列挙という形式になる．同一文書内においてはその出現順に表示することが可能であるが，複数の文書から重要文を出力する場合には，目的に応じて重要度の高い文章順（検索結果のランキング順）や，各文書の作成日時の情報をもとにした時系列順などの方法による工夫が必要となる．

3. 展望台システムによる複数文書要約

重要文抽出法である展望台システムへの入力として複数の文章を与え，それら複数の文章から重要文を抽出する方法について本章で述べる．文章の重要文としては次の2種類がある．

1. 文章の構成（物語）を理解する上で重要な文
2. 文章の観点に基づいた主題を理解する上で重要な文

このうち展望台システムが各文章から出力する要約は上記 2. の重要文である．特に，複数文書からの要約にはもとからストーリーがない場合がほとんどであると考えられるため，ひとつの話としてすべてをまとめた要約を得ることは困難である．だからといって，全く異なる内容の重要文を数多く列挙しても，文書集合全体の把握に有効とはいい難い．本論文で提案する手法においては，文書集合全体の観点となるキーワードに基づいて重要文が抽出されるため，意味のつながりがありかつ，特徴的なキーワードを含む重要文が抽出される．

展望台システムでは，以下の3種類のキーワードを扱い，それぞれのキーワードの役割に応じて文章から重要文を抽出する．

1. 背景キーワード:
   文章が書かれるもととなる背景知識を表す単語
2. 観点キーワード:
   人が主観的にもっている文章の主題を表す単語
3. 特徴キーワード:
   文章の主題とのみ関わって主題を特徴付ける単語

これら3つのキーワードを組み合わせて用いることで，各キーワードの特性を生かした重要文を文章から抽出している．以下ではまず，展望台システムのアルゴリズムについて説明する．

3.1 展望台システムのアルゴリズム

ステップ0：キーワード候補の抽出

文章のキーワードの候補として，形態素解析[2]によって，名詞，動詞，形容詞を取り出す．

ステップ1：背景キーワードの抽出

まず，各単語 w の評価値 $key1(w)$ を文章中の出現頻度 $frequency(w)$ として式1で与える．

$$key1(w) = frequency(w) \tag{1}$$

この評価値による上位の単語を背景キーワードとし，文章に含まれるキーワード候補の種類数の 20 %を目安として最大で 20 個を取る．

ステップ2：観点語の入力 (展望台の建築)

ユーザから与えられる入力として，文章をどのような観点から要約したいかを表した観点語を受けとる．この観点語を文章の中心を表す観点キーワードの初期集合とする．また仮に，観点キーワードの入力がなかった場合には，背景キーワード集合 G を元に，次式による評価値の高い単語 w を観点キーワードの初期集合として与える．

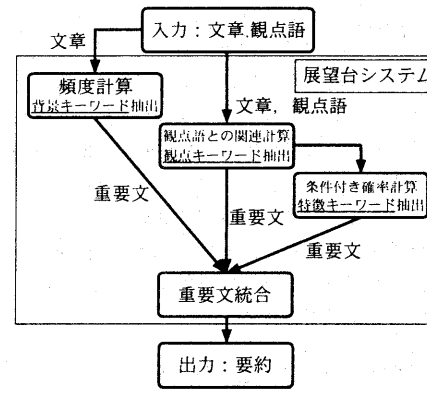

図1　展望台システムのアルゴリズム

$$key2'(w) = \sum_{g \subset G} \frac{n(w \cap g)}{n(g)} \tag{2}$$

初期集合の観点キーワードの数は単語種類数の 4 %で最大 5 個とする．

ステップ3：観点キーワードの補完

観点キーワードの初期集合をもとに，文章中で観点キーワードと強く関連する単語を新たに観点キーワードに加える．そこで，観点キーワード集合 S が現れたときに，同時に現れやすい単語を求める．すなわち，全ての単語 w に次の式による評価値 $key2$ を与え，その評価値による上位（単語種類数の 8 %で最大 10 個）に含まれかつ，まだ S に属さない単語を新たな観点キーワードとして S に加える．

$$key2(w) = \sum_{s \subset S} \frac{n(w \cap s)^2}{n(w)n(s)} \tag{3}$$

この式は単語 s が現れた時に単語 w が現れる条件付確率と，単語 w が現れた時に単語 s が現れる確率との積を表している．

ステップ4：特徴キーワードの発見

最後に文章の観点となっている観点キーワードの流れに沿って，文章を特徴付けているキーワードを抽出する．そのための単語 w の評価値 $key3$ を式4で与える．

$$key3(w) = \sum_{s \subset S} \frac{n(w \cap s)}{n(w)} \qquad (4)$$

すなわち，観点キーワードが出現する時（文）にのみ出現する単語を探して，どれだけ多くの観点キーワードと密接に関係しているかを表す評価値を各単語に与える．

ステップ5：文章中の各文を評価

文章の各文に文中に含まれる単語の評価値の総和を各文の重要度として与える．各文Tには式5から式7の3つの評価値を与える．

$$sentence1(T) = \sum_{w \subset T} key1(w) \qquad (5)$$

$$sentence2(T) = \sum_{w \subset T} key2(w) \qquad (6)$$

$$sentence3(T) = \sum_{w \subset T} key3(w) \qquad (7)$$

ステップ6：重要文の決定

式1から式4による各文の評価値を元に，重要文を総合的に決定する．まず各文に式5から式7のそれぞれの値に基づいて，評価値の高い順に順位をつける．すなわち，各文Tは式5による順位$rank1(T)$，式6による順位$rank2(T)$と式7による順位rank3(T)の3種類の順位をもつその順位付けの後に，式8の値を文の評価値として与える．

$$value(T) = \alpha rank1(T) + \beta rank2(T) + \gamma rank3(T) \qquad (8)$$

ただし，α, β, γは3種類のキーワードによる重要文のうちで，対象とする文章集合に応じて設定される重みである☆．この，式8の値に基づいて，用途に応じて必要な数☆☆の重要文を出力とする．

3.2 展望台システムの複数文書への適用

複数文書を入力として展望台システムに与え，複数文書の要約を得る方法としては次のものが考えられる．

1) 複数の文書をつなげて一つの文書とみなして要約する．

2) 各文書からの重要文をつなげて一つの要約とする．

1)は，複数の文書をつなげた一つの文書を展望台システムへの入力とし，展望台システムからの出力がそのまま文書集合の要約となる．2)は，各文書を別々に展望台システムへの入力とし，それぞれの結果をひとつにまとめる方法である．また，要約に用いる観点キーワードの取り方にも以下のものが考えられる．

☆ デフォルト値は$\alpha = \beta = \gamma = 1$である．

☆☆ 最重要文が必要な場合は1つ，全文数の10％に当たる数の重要文を取り出す要約（10％要約と呼ぶ）を行なう時には，その相当数の重要文を取り出す．

表1 検索用キーワード

広末	安倍	深田	優香	高橋	後藤	野村	郷
新庄	松井	伊藤	岡田	前田	華原	宇多田	
倉木	瞳	浜崎	平山	近藤	中村	今井	
篠原	藤本	福井	森田	大西	中谷	ゴルフ	
長嶋	鈴木	交際	研究	巨人	女優	メジャー	
投手	結婚	映画	番組	大学	野球	サッカー	
新聞	出演	写真	時代	代表	発売	アルバム	

a) 複数の文書をつなげた全文書からキーワードを取り出す．

b) 各文書ごとにキーワードを取り出す．

c) 人間が手入力で与える．

そこでまず手入力を除く，1)とa)，2)とa)，2)とb)の3通りの組合せのシステムを実装しその比較実験を行なった．実験に用いたデータベースは，平均30文の芸能スポーツ関係のニュース記事約15000件であり，要約する文章集合は1に示すキーワードを，それぞれ多く含む10件の文章として50回の試行を繰り返した．

要約文における文の数は，要約を行なう複数文書の数に一致させ，１０の文章を要約する際には１０文を重要文として抜き出す．すなわち，2)の方法で得られる最終的な要約は，各文書ごとに重要文を1つ抜きだし，それらをつなぎ合わせた１０文が出力となる．

その結果，1)と2)を比較すると1)は全文から要約を作成しているため，全体としての一貫性は2)を上回っているが，特定の文章から偏って重要文が選ばれる可能性がある．実際に1)とa)の組合せによって取り出された重要文が，もとのいくつの文書から抽出されているか（出典文書数と呼ぶ）を示したのが図2の系列1であり，平均で6.2文となった．

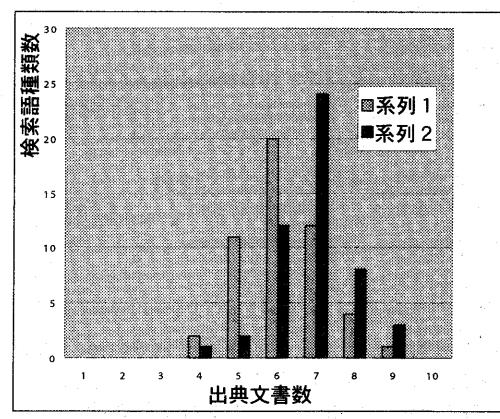

図2 出典文書数の分布

これはランダムに10文を選んだときの平均値6.6文よりも少ない値である．これは，各文章の長さが一定でないために，特定の文章に瀕出する単語がキーワードとなり，そのキーワードを含む文章からの重要文が数多く出力されたためである．これに対して，2) の要約は各文章から均等に重要文が得られる反面☆，各文書を別々に扱っているため全体に一貫性がなく，要約全体の理解や文章全体の内容を把握することが困難となる．そこで，1) と 2) を組み合わせた次の方法が考えられる．

3) 複数の文書から重要文を抜き出し，それらをまとめた一つの文書を再び展望台システムへの入力として要約する．

この方法によれば，各文書をその長さによって正規化することができるとともに，一貫性のある要約を作成することが可能となると考えられる．

次に，a),b) の比較を行なう．これに関しても，1) と 2) の比較の時と同様に，a) のキーワードは瀕出単語を含む文章から偏って選ばれる傾向にある．しかし，b) のキーワードは各文書ごとに抽出されたキーワードであるために，それらのキーワードを全体としてまとめる機構が存在しないため，重要文をつなぎ合わせただけの，一貫性のない要約となってしまっている．そこで，観点キーワード抽出の方法として次の方法が考えられる．

d) 各文書ごとのキーワードから選択する．

この選択には次の式による評価値を各キーワード (w) に与え，その上位のキーワード（5個）を選んだ．ただし，$filenum$ は要約する文章の数，$rank_i(w)$ は，単語 w の文書 i における3の評価値の順位（1,2,3,...）である．

$$key(w) = \sum_{i=1}^{filenum} (101 - rank_i(w)) \qquad (9)$$

すなわち，複数文書に上位で現れるキーワードほど高い評価値が与えられる．そこで，各文章からまず3文ずつ d) の方法によるキーワード（選抜キーワードと呼ぶ）を観点として重要文を抽出し，それらをつなぎ合わせてできる文章を再び選抜キーワードを観点として要約するシステムを実装し，先のシステムと比較した．その結果，出典文書数は図3の系列2のようになり，平均で6.9の文章から重要文が抽出された．これによって，要約文の偏りが軽減されるとともに，要約の一貫性が保持される．

☆ ちなみに，出典文書数は10である．

以上の予備実験により，最終的に提案される複数文書要約システムのアルゴリズムは以下のようになる．

1. 各文書を展望台システムに入力して各文書の観点キーワードを抽出する
2. 1. で得られたキーワードに9の評価値を与え，上位のキーワードを抽出する
3. 2. のキーワードを観点キーワードとして，展望台システムによって各文書から重要文を抽出する
4. 3. の重要文をつないで一つの文章とみなし，展望台システムによって重要文を抽出する

また，検索エンジンなど手入力で観点キーワードの入力が与えられる場合のアルゴリズムは，ステップ1を次のように読み換えたものとなる．

1. 各文書と検索語（観点キーワードの初期集合）を展望台システムに入力して，各文書の観点キーワードを抽出する

ここで検索語を観点キーワードとして与えた場合と与えない場合とで，比較する実験を行なった．要約に用いる文章は,「広末」「安倍」「麗奈」「イチロー」「尚子」「山口」という単語を含む検索エンジンによって出力された文章であり，10文ずつ順に(10+2+3+10+5+10=)40組に対して実験した．出典文章数は，検索語を観点として与えた場合は7.3であり，与えない場合は前回と同じ6.9であった．また，各文書集合における検索語の平均頻度と出典文書数の関係を示したのが，図3と図4である．

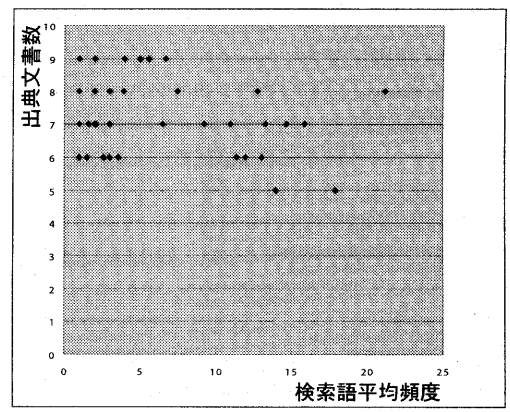

図3　検索語観点

出典文章数が8以上になるのは文書集合における観点キーワードの平均頻度が10以下の場合がほとんどである．これは，一つの文章に観点となる単語が多く現れていると，その文章から偏って重要文が抽出される可能性が高いことを裏付けており，観点キーワード

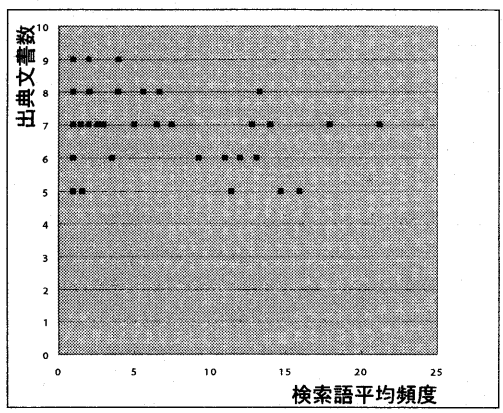

図4　自動補完

の各文章中での出現頻度が高いほど，重要文が偏って選ばれ出典文書の数が少なくなる傾向にある．検索語を用いた場合に出典文書数が少ない理由も同様に，選抜キーワードの方が各文章中における出現頻度が高くなっていたためと考えられる．

4. 重要文抽出実験

先に述べたデータベースに対するサーチエンジンを用いて，提示される要約のみから「NHKの朝の連続ドラマ「ちゅらさん」の主演女優は誰？」のようなクイズ形式の質問の解答を探す実験を行なった．質問はタレントの名前もしくは，番組，アルバム，映画などのタイトルを答える質問に限定した．要約はランキング上位の１０文書から作成された以下の各１０文による要約が，一回の検索に対して同時に提示される．

- システムA：展望台複数文書要約アルゴリズム
- システムB：各文書から検索語を含む最初の１文
- システムC：各文書から展望台による１文
- システムD：各文書の冒頭からの１文

もっとも早く解答が含まれる要約を提示したシステムを正答とし，その正答率（正答数／質問数）を計算したところ，システムAとCが83%，Bが25%，Dが14%となった．システムAとCの展望台システムによる要約の正答率が高いが，AとCの比較によって，本稿で提案する複数要約の効果を確かめるには至っていない．その原因としてはデータベースの大きさに対して検索語が多く使われたため，検索結果が10文以下になり，システムAとCで要約に差がでなかったものと考えられる．今後は質問の内容および用いる検索語の数に制限を加えるなどして再実験し，複数文書要約が有効に用いられる場面について検討していきたい．

5. 結　　論

文書要約システムである展望台システムを応用して，複数の文書を要約する手法を提案した．大量の情報が溢れる世の中において，その全体や一部分の概要を素早く理解して情報の取捨選択を支援するシステムは，今後ますます不可欠となる．本論文で提案する手法がこれに役立てば幸いである．

参 考 文 献

1) Salton, G. and Buckey, C.: "Term-Weighting Approaches in Automatic Text Retrieval", Readings in Information Retrieval, pp.323 – 328, (1997).
2) 松本裕治，北内啓，山下達雄，平野義隆，松田寛，浅原正幸：日本語形態素解析システム「茶筌」version2.0 使用説明書第二版，NAIST-IS-TR99012, (1999).
http://cl.aist-nara.ac.jp/lab/nlt/chasen/
3) 福原知宏，武田英明，西田豊明：統計情報を用いた話題特定と文脈の再構築による複数テキスト要約，人工知能学会全国大会 (第 13 回) 論文集, pp.555 – 558, (1999).
4) Luhn, H.P.: The automatic creation of literature abstracts, In IBM Journal for Research and Development, Vol.2, No.2, pp.59 – 165, (1958).
5) 難波英嗣，奥村学：論文間の参照情報を考慮したサーベイ論文作成支援システムの開発，自然言語処理, Vol6, No.5, pp.43 – 62, (1999).
6) 大澤幸生，ネルスベンソン，谷内田正彦: KeyGraph:単語共起グラフの分割・統合によるキーワード抽出，電子情報通信学会論文誌, Vol.J82-D-I, No.2, pp.391 – 400, (1999).
7) 大竹清敬，船坂貴浩，増山繁，山本和英: 冗長部削除による複数記事要約手法，自然言語処理, Vol.6, No.6, pp.45 – 64, (1999).
8) Radev, D., McKeown, K.: Generating Natural Language Summaries from Multiple On-Line Sources, Computational Linguistics, Vol.24, No.3, pp.469 – 500, (1998).
9) 砂山渡・大澤幸生・谷内田正彦：ユーザの興味の構造を用いて関連検索キーを提示する検索支援インターフェイス，人工知能学会誌，Vol.15. No.6, pp.1117 – 1124, (2000).
10) 砂山渡・谷内田正彦：観点に基づく文章要約のための重要文抽出法-展望台システム-，第 4 回人工知能学会ことば工学研究会資料, pp.15 – 22, (2000).