(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4967133号
(P4967133)
(45)発行日　平成24年7月4日(2012.7.4)　(24)登録日　平成24年4月13日(2012.4.13)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 350C
G06F 17/30 340Z
G06F 17/30 170A

請求項の数 4　(全 31 頁)

(21)出願番号　特願2007-85469(P2007-85469)
(22)出願日　平成19年3月28日(2007.3.28)
(65)公開番号　特開2008-243024(P2008-243024A)
(43)公開日　平成20年10月9日(2008.10.9)
審査請求日　平成21年9月25日(2009.9.25)

(73)特許権者　504174135
国立大学法人九州工業大学
福岡県北九州市戸畑区仙水町１番１号
(74)代理人　100099634
弁理士　平井　安雄
(72)発明者　野村　浩郷
福岡県飯塚市大字川津６８０－４　九州工業大学内

審査官　鈴木　和樹

最終頁に続く

(54)【発明の名称】情報取得装置、そのプログラム及び方法

(57)【特許請求の範囲】
【請求項１】
少なくとも１つの検索情報を取得している情報取得装置において、
重み付けされた検索情報の特徴ベクトルを作成する特徴ベクトル作成手段と、
全検索情報の特徴ベクトルの組み合わせの類似度を計算する情報間類似度計算手段と、
前記情報間類似度計算によって得られた数値の類似度行列を計算する類似度行列計算手段と、
前記類似度計算結果を数値解析し、特徴ベクトルの最大固有値の固有ベクトルを求める固有ベクトル作成手段と、
前記検索情報の問い合わせ内容の質問ベクトルを作成する質問ベクトル作成手段と、
前記特徴ベクトルと質問ベクトルの余弦の計算値に固有ベクトルの数値を乗じて求められる関連情報の検索順位を決定する検索順位決定手段と、
検索された情報の文中に含まれる品詞の係り受け関係を解析する係り受け解析手段と、
各文中の動詞を含む文節に係る文節中の名詞を抽出する名詞抽出手段と、
前記抽出された名詞の単体の名詞間の類似度S_1及び名詞集合の類似度S_2を計算する名詞集合間類似度比較計算手段と、
抽出された名詞の表示の一致する割合の類似度S_3を計算する名詞表示一致割合計算手段と、
前記類似度S_2に類似度S_3を加えて文類似度Ｓを計算する文類似度計算手段と、
検索情報の文タイプによる選定を行う文タイプ選定手段と、

前記文類似度計算及び文タイプ選定された関連情報の内容を統合したもの、並びに、前記検索順位決定手段によりスコアリングされた検索結果を出力すると共に、前記検索結果の適否及び／又はパラメータの重み付けの度合いを入力するための入力フォームを出力する出力手段と、

前記入力フォームに入力された情報に基づいて、前記特徴ベクトル及び質問ベクトルを修正する修正手段とを備え、

前記検索順位決定手段が、前記修正手段にて修正された前記特徴ベクトル及び質問ベクトルに基づいて、再度前記関連情報の検索順位を決定し、前記出力手段が、前記検索順位決定手段によりスコアリングされた検索結果を、前記検索情報間の経時的な関連性を含めて出力することを特徴とする情報取得装置。

【請求項２】

前記請求項１に記載された情報取得装置において、

前記特徴ベクトル作成手段は、

検索情報の文の形態素解析を行う形態素解析手段と、

情報毎に単語とその単語の出現回数ＴＦを計算するＴＦ計算手段と、

全単語について文書頻度ＤＦ及びそのＩＤＦを計算するＩＤＦ計算手段と、

各情報の各単語についてＴＦ－ＩＤＦ法を用いて単語重み付けを計算する単語重み計算手段と、

前記単語重み付けから各文書の特徴ベクトルを作成する特徴ベクトル作成手段とを備えることを特徴とする情報取得装置。

【請求項３】

少なくとも１つの検索情報を取得している情報取得装置としてコンピュータを機能させる情報取得プログラムにおいて、

重み付けされた検索情報の特徴ベクトルを作成する特徴ベクトル作成手段、

全検索情報の特徴ベクトルの組み合わせの類似度を計算する情報間類似度計算手段、

前記情報間類似度計算によって得られた数値の類似度行列を計算する類似度行列計算手段、

前記類似度計算結果を数値解析し、特徴ベクトルの最大固有値の固有ベクトルを求める固有ベクトル作成手段、

前記検索情報の問い合わせ内容の質問ベクトルを作成する質問ベクトル作成手段、

前記特徴ベクトルと質問ベクトルの余弦の計算値に固有ベクトルの数値を乗じて求められる関連情報の検索順位を決定する検索順位決定手段、

検索された情報の文中に含まれる品詞の係り受け関係を解析する係り受け解析手段、

各文中の動詞を含む文節に係る文節中の名詞を抽出する名詞抽出手段、

前記抽出された名詞の単体の名詞間の類似度S_1及び名詞集合の類似度S_2を計算する名詞集合間類似度比較計算手段、

抽出された名詞の表示の一致する割合の類似度S_3を計算する名詞表示一致割合計算手段、

前記類似度S_2に類似度S_3を加えて文類似度Ｓを計算する文類似度計算手段、

検索情報の文タイプによる選定を行う文タイプ選定手段、

前記文類似度計算及び文タイプ選定された関連情報の内容を統合したもの、並びに、前記検索順位決定手段によりスコアリングされた検索結果を出力すると共に、前記検索結果の適否及び／又はパラメータの重み付けの度合いを入力するための入力フォームを出力する出力手段、

前記入力フォームに入力された情報に基づいて、前記特徴ベクトル及び質問ベクトルを修正する修正手段としてコンピュータを機能させ、

前記検索順位決定手段が、前記修正手段にて修正された前記特徴ベクトル及び質問ベクトルに基づいて、再度前記関連情報の検索順位を決定し、前記出力手段が、前記検索順位決定手段によりスコアリングされた検索結果を、前記検索情報間の経時的な関連性を含めて出力することを特徴とする情報取得プログラム。

【請求項４】
少なくとも１つの検索情報を取得している情報取得装置のコンピュータが、
重み付けされた検索情報の特徴ベクトルを作成する特徴ベクトル作成ステップと、
全検索情報の特徴ベクトルの組み合わせの類似度を計算する情報間類似度計算ステップと、
前記情報間類似度計算によって得られた数値の類似度行列を計算する類似度行列計算ステップと、
前記類似度計算結果を数値解析し、特徴ベクトルの最大固有値の固有ベクトルを求める固有ベクトル作成ステップと、
前記検索情報の問い合わせ内容の質問ベクトルを作成する質問ベクトル作成ステップと、
前記特徴ベクトルと質問ベクトルの余弦の計算値に固有ベクトルの数値を乗じて求められる関連情報の検索順位を決定する検索順位決定ステップと、
検索された情報の文中に含まれる品詞の係り受け関係を解析する係り受け解析ステップと、
各文中の動詞を含む文節に係る文節中の名詞を抽出する名詞抽出ステップと、
前記抽出された名詞の単体の名詞間の類似度S_1及び名詞集合の類似度S_2を計算する名詞集合間類似度比較計算ステップと、
抽出された名詞の表示の一致する割合の類似度S_3を計算する名詞表示一致割合計算ステップと、
前記類似度S_2に類似度S_3を加えて文類似度Ｓを計算する文類似度計算ステップと、
検索情報の文タイプによる選定を行う文タイプ選定ステップと、
前記文類似度計算及び文タイプ選定された関連情報の内容を統合したもの、並びに、前記検索順位決定手段によりスコアリングされた検索結果を出力すると共に、前記検索結果の適否及び／又はパラメータの重み付けの度合いを入力するための入力フォームを出力する出力ステップと、
前記入力フォームに入力された情報に基づいて、前記特徴ベクトル及び質問ベクトルを修正する修正ステップとを実行し、
前記検索順位決定ステップが、前記修正手段にて修正された前記特徴ベクトル及び質問ベクトルに基づいて、再度前記関連情報の検索順位を決定し、前記出力ステップが、前記検索順位決定手段によりスコアリングされた検索結果を、前記検索情報間の経時的な関連性を含めて出力することを特徴とする情報取得方法。
【発明の詳細な説明】
【技術分野】
【０００１】
関連内容の情報の検索とそれらを集約する情報取得装置に関する。
【背景技術】
【０００２】
Ｗｅｂ検索を含めた多くの情報検索システムでは、キーワード検索を元にしており、ユーザは検索キーワードをシステムに与えることで検索結果を得る。単純な単語のマッチングのみを条件として検索を行うために、検索結果が膨大になることが多く、またノイズも多い。現状ではユーザは膨大な量で、しかも玉石混淆の検索結果から要求に合致したテキストを探さなければならない。そのため、ユーザが検索結果から合致した情報を得るためには多大な労力を必要とする。また、キーワード検索の途中で関連する情報を発見したいという状況も頻繁に発生することがある。さらに、近年情報機器の普及により様々な情報が電子化されており、大量の情報がいつでも閲覧できるようになった現在、その中から必要な情報を効率よく選ぶ作業は、情報の電子化が急速に進んでいる中、困難になっている。
そこで、複数のデータベースを検索して所望の情報を取得し、その情報をユーザの望む形式に編集・加工する情報編集・加工方法が、特開平９－１８５６３２号公報に開示され

ている。

【０００３】

背景技術の情報編集・加工方法は、遠隔にある少なくとも１つのデータベースが保有していると推定される目的情報の検索指示、検索した情報の編集加工指示、編集加工した情報の出力形態決定指示とを受け付ける第１の過程と、前記検索指示に基づいて前記少なくとも１つのデータベースを検索するコマンドを生成し、前記データベースを検索する第２の過程と、前記第２の過程により取得した少なくとも１つの目的情報に対して前記第１の過程の指示に従って、編集・加工を施す第３の過程と、前記第３の過程によって、編集・加工された目的情報に対して前記情報の出力形態決定指示に従って、所定の出力形態に変換し、視覚、聴覚または他の感覚に捉え得る方法によって出力する第４の過程よりなることを特徴とする。前記第３の過程は、前記第４の過程において表示する検索結果をユーザが指定した出力順にソートする。前記出力順として、情報の関連度順、情報発生時間順、または検索順のいずれかを用いるものである。

【特許文献１】特開平９－１８５６３２号公報

【発明の開示】

【発明が解決しようとする課題】

【０００４】

上述のように背景技術の情報検索・編集方法及び装置によれば、得られた情報間の関連を意識するので、関連のある情報同士を近接してユーザに提供することができる。また、複数のデータベースから得られた情報をユニフォームに扱うため、異なるデータベースから得られた情報の提供を時間順であっても関連度順であっても適切に行うことができる。

【０００５】

しかしながら、提供された情報がトピックスについては条件を満たしていても、情報内容についての質あるいは量が、ユーザの要求を必ずしも充分に満たしているとは限らない場合がある。そのため、ユーザの希望する情報が不足している場合に改めて情報検索を行わなければならいという課題を有する。

【０００６】

また、あるトピックスに対して複数の観点から作成された文書等の情報に関して、それらの情報を比較し、理解を深めるということも可能になってきてはいるが、その作業もまた情報量の増大につれて困難になってきている。しかも、関連情報に関しては情報の重複個所を何度も繰り返し取得するためにユーザの負担が大きくなるという課題もある。

【０００７】

本発明は、前記課題を解決するためになされたものであり、続報情報を発見する検索装置において関連した情報を比較、整理して効率よく必要な情報を取得し、利用者の要望に沿った形式で出力を行うことができる情報取得装置の提供を目的とする。

【課題を解決するための手段】

【０００８】

本発明に係る情報取得装置は、少なくとも１つの検索情報を取得している情報取得装置において、重み付けされた検索情報の特徴ベクトルを作成する特徴ベクトル作成手段と、全検索情報の特徴ベクトルの組み合わせの類似度を計算する情報間類似度計算手段と、前記情報間類似度計算によって得られた数値の類似度行列を計算する類似度行列計算手段と、前記類似度計算結果を数値解析し、特徴ベクトルの最大固有値の固有ベクトルを求める固有ベクトル作成手段と、前記検索情報の問い合わせ内容の質問ベクトルを作成する質問ベクトル作成手段と、前記特徴ベクトルと質問ベクトルの余弦の計算値に固有ベクトルの数値を乗じて求められる関連情報の検索順位を決定する検索順位決定手段と、検索された情報の文中に含まれる品詞の係り受け関係を解析する係り受け解析手段と、各文中の動詞を含む文節に係る文節中の名詞を抽出する名詞抽出手段と、前記抽出された名詞の単体の名詞間の類似度S_1及び名詞集合の類似度S_2を計算する名詞集合間類似度比較計算手段と、抽出された名詞の表示の一致する割合の類似度S_3を計算する名詞表示一致割合計算手段と、前記類似度S_2に類似度S_3を加えて文類似度Ｓを計算する文類似度計算手段と、検

索情報の文タイプによる選定を行う文タイプ選定手段と、前記文類似度計算及び文タイプ選定された関連情報の内容を統合したもの、並びに、前記検索順位決定手段によりスコアリングされた検索結果を出力すると共に、前記検索結果の適否及び／又はパラメータの重み付けの度合いを入力するための入力フォームを出力する出力手段と、前記入力フォームに入力された情報に基づいて、前記特徴ベクトル及び質問ベクトルを修正する修正手段とを備え、前記検索順位決定手段が、前記修正手段にて修正された前記特徴ベクトル及び質問ベクトルに基づいて、再度前記関連情報の検索順位を決定し、前記出力手段が、前記検索順位決定手段によりスコアリングされた検索結果を、前記検索情報間の経時的な関連性を含めて出力するものである。ここで、「情報」には、例えば、文、文書、記事、画像、音声等を含む。また、「文タイプ」とは、例えば、重複個所、固有個所、補足説明等である。さらに、要旨、予定、理由、分析、補足説明、様態・伝聞、比況・推量等を含むものとする。

【０００９】

これにより、重み付けされた検索情報の特徴ベクトルを作成し、全検索情報の特徴ベクトルの組み合わせの類似度を計算し、前記情報間類似度計算によって得られた数値の類似度行列を計算し、前記類似度計算結果を数値解析し、特徴ベクトルの最大固有値の固有ベクトルを求め、前記検索情報の問い合わせ内容の質問ベクトルを作成し、前記特徴ベクトルと質問ベクトルの余弦の計算値に固有ベクトルの数値を乗じて求められる関連情報の検索順位を決定し、前記検索順位の決定された関連情報の内容を統合して出力するので、すでに検索された情報から、その情報と類似度が高い情報を優先的に選択し、類似度に応じた確率で読み進めていくという仮想的なユーザを考えるとき、無限時間後に定常状態になった時点で、どの情報に行きつくかという確率に相当することを判断しながら関連情報を検索し、ユーザにとって必要な情報を取得することができる。

また、検索された情報の文中に含まれる品詞の係り受け関係を解析し、各文中の動詞を含む文節に係る文節中の名詞を抽出し、前記抽出された名詞の単体の名詞間の類似度S_1及び名詞集合の類似度S_2計算し、抽出された名詞の表示の一致する割合の類似度S_3を計算し、前記類似度S_2に類似度S_3を加えて文類似度Ｓを計算し、検索情報の文タイプによる選定を行い、前記文類似度計算及び文タイプ選定された関連情報の内容を統合して出力するので、検索された関連情報について、重複個所、固有個所、補足説明等の情報内容を整理した状態で、関連情報を取得することができる。また、膨大な量の関連情報に含まれる重複情報が何度も繰り返し表示されることによるユーザの負担を軽減でき、整理された固有個所や補足説明の情報を効率よく利用者の要望に沿った形式により取得することができる。

さらに、検索順位決定によりスコアリングされた検索結果を開示し、前記検索結果の適否を入力し、特徴及び質問ベクトルを修正して、検索情報の内容を出力するので、特徴ベクトル及び質問ベクトルをユーザが適合していると判断した情報に近づけ、不適合であると判断した情報から遠ざけるように特徴ベクトル及び質問ベクトルを生成していくことができる。これを繰り返し適用することにより、確実にユーザの望む検索結果を得ることができる。

【００１４】

本発明に係る情報取得装置は必要に応じて、前記特徴ベクトル作成手段は、検索情報の文の形態素解析を行う形態素解析手段と、情報毎に単語とその単語の出現回数ＴＦを計算するＴＦ計算手段と、全単語について文書頻度ＤＦ及びそのＩＤＦを計算するＩＤＦ計算手段と、各情報の各単語についてＴＦ－ＩＤＦ法を用いて単語重み付けを計算する単語重み計算手段と、前記単語重み付けから各文書の特徴ベクトルを作成する特徴ベクトル作成手段とを備えるものである。

【００１５】

これにより、重み付けされた検索情報の特徴ベクトルを作成する特徴ベクトル作成手段は、検索情報の文の形態素解析を行い、情報毎に単語とその単語の出現回数ＴＦを計算し、全単語について文書頻度ＤＦ及びそのＩＤＦを計算し、各情報の各単語についてＴＦ－

ＩＤＦ法を用いて単語重み付けを計算し、前記単語重み付けから各文書の特徴ベクトルを作成するので、特徴ベクトル作成手段情報検索において目的の情報を探すために，関連情報についての重要度を判断するためのひとつの指標とすることができ、文書と単語の関連性の数値演算を行い、その値の高いものを特徴ベクトルに反映することにより、よりユーザにとって重要な関連情報を取得することができる。

【００１６】

本発明に係る情報取得プログラムは、少なくとも１つの検索情報を取得している情報取得装置としてコンピュータを機能させる情報取得プログラムにおいて、重み付けされた検索情報の特徴ベクトルを作成する特徴ベクトル作成手段、全検索情報の特徴ベクトルの組み合わせの類似度を計算する情報間類似度計算手段、前記情報間類似度計算によって得られた数値の類似度行列を計算する類似度行列計算手段、前記類似度計算結果を数値解析し、特徴ベクトルの最大固有値の固有ベクトルを求める固有ベクトル作成手段、前記検索情報の問い合わせ内容の質問ベクトルを作成する質問ベクトル作成手段、前記特徴ベクトルと質問ベクトルの余弦の計算値に固有ベクトルの数値を乗じて求められる関連情報の検索順位を決定する検索順位決定手段、検索された情報の文中に含まれる品詞の係り受け関係を解析する係り受け解析手段、各文中の動詞を含む文節に係る文節中の名詞を抽出する名詞抽出手段、前記抽出された名詞の単体の名詞間の類似度S_1及び名詞集合の類似度S_2を計算する名詞集合間類似度比較計算手段、抽出された名詞の表示の一致する割合の類似度S_3を計算する名詞表示一致割合計算手段、前記類似度S_2に類似度S_3を加えて文類似度Ｓを計算する文類似度計算手段、検索情報の文タイプによる選定を行う文タイプ選定手段、前記文類似度計算及び文タイプ選定された関連情報の内容を統合したもの、並びに、前記検索順位決定手段によりスコアリングされた検索結果を出力すると共に、前記検索結果の適否及び／又はパラメータの重み付けの度合いを入力するための入力フォームを出力する出力手段、前記入力フォームに入力された情報に基づいて、前記特徴ベクトル及び質問ベクトルを修正する修正手段としてコンピュータを機能させ、前記検索順位決定手段が、前記修正手段にて修正された前記特徴ベクトル及び質問ベクトルに基づいて、再度前記関連情報の検索順位を決定し、前記出力手段が、前記検索順位決定手段によりスコアリングされた検索結果を、前記検索情報間の経時的な関連性を含めて出力するものである。

【００１８】

本発明に係る情報取得方法は、少なくとも１つの検索情報を取得している情報取得装置のコンピュータが、重み付けされた検索情報の特徴ベクトルを作成する特徴ベクトル作成ステップと、全検索情報の特徴ベクトルの組み合わせの類似度を計算する情報間類似度計算ステップと、前記情報間類似度計算によって得られた数値の類似度行列を計算する類似度行列計算ステップと、前記類似度計算結果を数値解析し、特徴ベクトルの最大固有値の固有ベクトルを求める固有ベクトル作成ステップと、前記検索情報の問い合わせ内容の質問ベクトルを作成する質問ベクトル作成ステップと、前記特徴ベクトルと質問ベクトルの余弦の計算値に固有ベクトルの数値を乗じて求められる関連情報の検索順位を決定する検索順位決定ステップと、検索された情報の文中に含まれる品詞の係り受け関係を解析する係り受け解析ステップと、各文中の動詞を含む文節に係る文節中の名詞を抽出する名詞抽出ステップと、前記抽出された名詞の単体の名詞間の類似度S_1及び名詞集合の類似度S_2を計算する名詞集合間類似度比較計算ステップと、抽出された名詞の表示の一致する割合の類似度S_3を計算する名詞表示一致割合計算ステップと、前記類似度S_2に類似度S_3を加えて文類似度Ｓを計算する文類似度計算ステップと、検索情報の文タイプによる選定を行う文タイプ選定ステップと、前記文類似度計算及び文タイプ選定された関連情報の内容を統合したもの、並びに、前記検索順位決定手段によりスコアリングされた検索結果を出力すると共に、前記検索結果の適否及び／又はパラメータの重み付けの度合いを入力するための入力フォームを出力する出力ステップと、前記入力フォームに入力された情報に基づいて、前記特徴ベクトル及び質問ベクトルを修正する修正ステップとを実行し、前記検索順位決定ステップが、前記修正手段にて修正された前記特徴ベクトル及び質問ベクトルに基づいて、再度前記関連情報の検索順位を決定し、前記出力ステップが、前記検索順位

決定手段によりスコアリングされた検索結果を、前記検索情報間の経時的な関連性を含めて出力するものである。

【発明を実施するための最良の形態】

【００２０】

ここで、本発明は多くの異なる形態で実施可能である。したがって、下記の実施形態の記載内容のみで解釈すべきではない。実施形態では、主に装置について説明するが、所謂当業者であれば明らかな通り、本発明は、コンピュータで使用可能なプログラムとしても実施できる。また、本発明では、ハードウェア、ソフトウェア、または、ソフトウェア及びハードウェアの実施形態で実施可能である。プログラムは、ハードディスク、ＣＤ－ＲＯＭ、ＤＶＤ－ＲＯＭ、光記憶装置または磁気記憶装置等の任意のコンピュータ可読媒体に記録できる。さらに、プログラムはネットワークを介した他のコンピュータに記録することが出来る。

【００２１】

［１．ハードウェア構成］

図１に本発明の実施形態における情報取得装置のハードウェア構成図を示す。コンピュータ１は、例えば、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）２、メインメモリ３、ＨＤＤ（Ｈａｒｄ　Ｄｉｓｋ　Ｄｒｉｖｅ）４、ビデオカード５、マウス６、キーボード７、光学ディスク８等を含む。なお、必要に応じて、データベース等を接続することもできる。

【００２２】

［２．ブロック構成］

図２に本発明の実施形態に係る情報取得装置のブロック構成図を示す。本発明は、主として、入力部１０、続報情報検索部２０、情報内容統合部３０、出力部４０を含む。ここで、続報情報検索部２０は、ベクトル作成部２１、検索順位決定部２２、適合・非適合判定部２３を含む。さらに、ベクトル作成部２１は、形態素解析部２１１、ＴＦ計算部２１２、ＩＤＦ計算部２１３、単語重み計算部２１４、特徴ベクトル作成部２１５、情報間類似度計算部２１６、類似度行列計算部２１７、固有ベクトル作成部２１８、質問ベクトル作成部２１９を含む。また、情報内容統合部３０は、係り受け解析部３１、名詞抽出部３２、名詞集合間類似度比較計算部３３、名詞表示一致割合計算部３４、文類似度計算部３５、文タイプ選定部３６、要約文作成部３７、記事集約部３８を含む。

まず、入力部１０により入力された記事データは、続報情報検索部２０におけるベクトル作成部２１送られて処理されることになる。ここで、本発明の実施形態の例として、検索対象に新聞記事を一例に挙げて、各構成の内容について以下に詳説する。

【００２３】

［２．１　続報情報検索］

［２．１．１ ベクトル空間モデル］

ベクトル作成部２１において、まず、記事データは形態素解析部２１１、ＴＦ計算部２１２、ＩＤＦ計算部２１３、単語重み計算部２１４、特徴ベクトル作成部２１５で処理される。ここで、ベクトル空間モデル（ｖｅｃｔｏｒ－ｓｐａｃｅ　ｍｏｄｅｌ）は検索対象となる個々のデータの性質を表現するための特徴量として、多次元ベクトルを個々のデータに対応づける。この間に類似度（Ｓｉｍｉｌａｒｉｔｙ）を定義することにより、問い合わせ（質問）と類似したものを探し出す方法である。いま、検索対象の特徴としてｎ個の属性が備わっており、ｉ番目の属性をw_iとする。そしてｊ番目のデータに（数式１）のベクトルを対応させることを考える。これらのベクトルが線形独立であれば、ｎ次元のベクトル空間が定義される。このように定義されたベクトル空間において、ｊ番目データの特徴ベクトルは

【００２４】

【数１】

$$\overrightarrow{D_j} = (w_{dj1}, w_{dj2}, w_{dj3}, \Lambda\ , w_{djn})$$

のように表すことができる（ｄ$_{ji}$はのｗ$_i$に対する値）。
ベクトル空間モデルにおける検索システムへの問い合わせ（質問）もベクトルで表される。ｎ次元のベクトル空間に対するその質問ベクトルは
【００２５】
【数２】

$$\vec{Q} = (w_{q1}, w_{q2}, w_{q3}, \Lambda\ , w_{qn})$$

のように表すことができる（ｑ$_i$は質問ベクトルのｗ$_i$に対する値）。
【００２６】
検索は検索対象の（数式１）の特徴ベクトルと（数式２）の質問ベクトルの類似度を計算することにより行われる。この特徴ベクトル（ｆｅａｔｕｒｅ　ｖｅｃｔｏｒ）を得る方法は、検索の目的や、その対象であるデータの種類などによって異なる。例えば検索対象が文献データならば単語の出現頻度を基にベクトルの各要素の重み付けを行い、画像であれば画素ごとの濃淡や色のデータなどを用いることができる。なお、新聞記事では、１記事を１つのベクトルに割り当て、記事中の単語のＴＦ－ＩＤＦ をベクトルの重みづけに利用している。
【００２７】
［２．１．２ 単語の重みづけ］
情報検索において目的の文書を探すために、文書と単語の関連性の数値演算を行い、その値の高いものを候補とする。そこで用いられる評価値は文書中には重要な単語がどれくらい多く含まれているかを表している。文書中の単語がどの程度重要であるか重み付けに用いられているのが以下に述べるＴＦ－ＩＤＦ 法である。この手法は次の２つのキーワードの性質に注目している。
（１）文書に数多く、高い頻度で現れる単語は重要である
（２）少ない数の文書にしか現れない単語は重要である
単語出現頻度（Ｔｅｒｍ　Ｆｒｅｑｕｅｎｃｙ：ＴＦ）単語ｔが文書ｄに高い頻度で現れるなら、ｔはｄを良く特徴付ける。この考えによる尺度が単語出現頻度、ｔｆ（Ｔｅｒｍ　Ｆｒｅｑｕｅｎｃｙ）である。ある文書ｄにおける単語ｔの出現頻度ｔｆ（ｄ，ｔ）は次式で定義され、ＴＦ計算部２１２において計算が行われる。
【００２８】
【数３】

$$tf(d,t) = freq(d,t)$$

ｆｒｅｑ（ｄ，ｔ）：文書ｄにおける単語ｔの出現頻度。
【００２９】
文書出現頻度（Ｄｏｃｕｍｅｎｔ　Ｆｒｅｑｕｅｎｃｙ：ｄｆ）ｔｆが大きいというのは重要な性質だが、それだけでは十分に文書を特徴付けることはできない。例えば、日本語文書で「は」という助詞はどんな文書でも高い頻度で現れるが、特定の文書を特徴付けないことは明白である。そこで、単語ｔが検索対象となる文書集合のうちの少数の文書にしか現れないという性質が重要である。単語ｔの出現する文書数を文書出現頻度ｄｆ（ｄｏｃｕｍｅｎｔ　ｆｒｅｑｕｅｎｃｙ）は、次式で定義される。
【００３０】

【数４】

$$df(t) = dfreq(t)$$

ｄｆｒｅｑ（ｔ）：単語ｔが出現する文書数
ｄｆが小さいことが単語ｔの文書を特徴付ける能力が高いことを表すので、実際にはこの逆数をｌｏｇと文書集合中の文書総数Ｎにより正規化したｉｄｆ（ｉｎｖｅｒｓｅ　ｄｏｃｕｍｅｎｔ　ｆｒｅｑｕｅｎｃｙ）を用いる。

【００３１】

【数５】

$$idf(t) = \log \frac{N}{dfreq(t)} + 1$$

Ｎ：文書の数
ｆｒｅｑ（ｄ，ｔ）：文書ｄにおける単語ｔの出現頻度

【００３２】

なお、ＩＤＦ計算部２１３では、まず、ｄｆを求めた後に、ｉｄｆを計算することになる。ＴＦ－ＩＤＦよる重み付け単語ｔについて、その単語が文書内に出てくる回数とそれが全文書内に占める割合の積を計算することで、その単語の重要性と、その出現頻度によって文書の重要性を表すことが目的である。単語ｔがｔｆとｉｄｆの両者の性質を併せ持つ、すなわちｔｆが大きく、ｄｆが小さいならば、単語ｔは文書ｄを真に特徴付けるといえる。この考え方を数値の尺度として表現したのがＴＦ－ＩＤＦによる重み付けである。文書ｄにおけるキーワードｔの重みｗ（ｔ，ｄ）は次のように定義され、単語重み計算部２１４で計算される。

【００３３】

【数６】

$$w(t,d) = tf(d,t) \cdot idf(t)$$

ｄｆｒｅｑ（ｔ）：単語ｔが出現する文書数
そして、特徴ベクトル作成部２１５において、これらの求められた数値を利用して特徴ベクトルを作成する。

【００３４】

［２．１．３　類似度］

特徴ベクトル作成部２１５、質問ベクトル作成部２１９で処理されたデータは検索順位決定部２２に送られる。ここで、ベクトル空間モデルにおいて、検索を行うためにはベクトル間の類似度を定義しなければならない。類似度の尺度としては様々なものがあるが、ここではベクトル間の余弦を用いる。

類似度として２つのベクトル間の余弦の値を利用する方法である。特徴ベクトルＤと質問ベクトルＱの類似度ｓｉｍ（Ｄ，Ｑ）は以下のようになる。

【００３５】

【数７】

$$sim(D,Q) = \frac{\vec{D}\vec{Q}}{\left|\vec{D}\right|\left|\vec{Q}\right|} = \frac{\sum_{i=1}^{n}(w_i q_i)}{\sqrt{\sum_{i=1}^{n}(w_i)^2 \sum_{i=1}^{n}(q_i)^2}}$$

ｓｉｍ（Ｄ，Ｑ）の値は０以上１以下であり、１に近づくほど類似度が高くなる。検索順位決定部２２では、余弦、いわゆるコサイン相関値を用いた類似度評価を行う。

【００３６】

［２．１．４　ベクトル空間モデルにおける関連性フィードバック］

検索順位決定部２２において検索結果が得られた場合、出力部４０で処理される。一度の検索で最終的な結果を得るのではなく、結果に対するユーザのフィードバックを元に新たな質問を生成し、繰り返し検索を行い、徐々に検索結果をユーザの求める結果に近づけていくフィードバック検索を行う。つまり、改めてユーザの検索結果に対する適否データを入力部１０において入力する。

【００３７】

ベクトル空間モデルにおける、関連性フィードバック（Ｒｅｌｅｖａｎｃｅ　Ｆｅｅｄｂａｃｋ）では、装置への質問式も質問ベクトル作成部２１９により作成された多次元ベクトルで表現される。質問の結果については、質問ベクトル作成部２１９で作成された質問ベクトルと特徴ベクトル作成部２１５で作成されたデータの特徴ベクトルの類似度を計算した結果のデータの集合として求める。この類似度が高いデータほど、質問の答えとしてふさわしいものであると考え、検索結果に含まれるデータに、それがどれだけ質問に適合していたかという順位をつけてユーザに提示する。ユーザは提示された検索結果からフィードバックを返す。

【００３８】

［２．１．５ ユーザからのフィードバック］

入力された検索結果の適否データは、適合・非適合判定部２３に送られ、質問ベクトル作成部２１９及び単語重み計算部２１４に送られる。

具体的なユーザからのフィードバックとして、最も多いのは結果の正例（ｐｏｓｉｔｉｖｅ　ｅｘａｍｐｌｅ）、負例（ｎｅｇａｔｉｖｅ　ｅｘａｍｐｌｅ）の提示である。また、正例のみをフィードバックするもの、それぞれの妥当性の度合いをランクづけてフィードバックするものなど、様々なものがある。また、ユーザからのフィードバックを検索に反映させる方法としては、大きく以下の二つに分けることができる。

（１）質問ベクトル修正（Ｑｕｅｒｙ　Ｖｅｃｔｏｒ　Ｍｏｖｅｍｅｎｔ）は、検索質問のベクトルを修正・変換して、正例の特徴に近づけ、負例から遠ざける。

（２）再重みづけ（Ｆｅａｔｕｒｅ　Ｒｅ－ｗｅｉｇｈｔｉｎｇ）は、特徴ベクトルに対応するための重みをユーザのニーズにあわせて調節する。すなわち、正例を検索するのに好都合な次元を強調し、負な例のものの影響を減らすように重みづけを動的に変更する。本発明では、この両方のフィードバックを利用する。

【００３９】

［２．１．６　質問ベクトルの修正］

ユーザのフィードバックした結果から、検索結果をユーザの求めるものに近づける手法として、質問ベクトルをユーザが適合していると判断した記事に近づけ、不適合であると判断した記事から遠ざけるように質問ベクトルを生成していく。これを繰り返し適用することにより、徐々にユーザの望む検索結果を得ることができる。このために良く利用されるのはＲｏｃｃｈｉｏフィードバック手法であり、Ｒｏｃｃｈｉｏの式は以下のように与えられる。

【００４０】

【数８】

$$Q_{i+1} = Q_i + \alpha \frac{1}{Rn} \sum_{j \in R} \vec{D}_j - \beta \frac{1}{Nn} \sum_{j \in N} \vec{D}_j$$

【００４１】
Q_iは前回の検索時に用いられた質問ベクトルであり、Q_{i+1}が新しく生成された質問ベクトルである。Ｒは適合だと判断された文書D_jに対する特徴ベクトルであり、Ｎは不適合であると判断された文書に対する特徴ベクトルである。R_n、N_n はそれぞれ適合文献数、不適合文献数である。α、βはそれぞれ適合文献、不適合文献に対する変数であり、αの値が高いと適合文献による変更が重要視され、βの値が高いと不適合文献による変更が重要視される。適合フィードバックの結果として、問合せ位置はQ_iからQ_{i+1}に移動するととらえることができる。ここで、特徴ベクトル作成部２１５におけるデータは、質問ベクトル作成部においても処理される。

【００４２】
［２．１．７　状態遷移確率を考慮に入れた重要度評価］
情報間類似度計算部２１６、類似度行列計算部２１７、固有ベクトル作成部２１８は、特徴ベクトル作成部２１５からのデータを以下の内容で処理する。
図３は本発明の実施形態に係る情報取得装置の記事間の類似度による記事の重要度評価の説明図である。ＰａｇｅＲａｎｋは、ｗｗｗ上のハイパーリンクによって結ばれたＷｅｂページ群において、「多くの良質なページからリンクされているページは、やはり良質なページである」、という再帰的な関係をもとに、Ｗｅｂページの重要度を評価する理論、およびそれによって求められるページの重要度である。ＰａｇｅＲａｎｋを用いることで、ハイパーリンク構造のような相互参照関係があるときに、どのページがもっとも重要であるかを定量的に求めることができる。

【００４３】
図３（ａ）はＰａｇｅＲａｎｋの概念図を示す。この図を例に基本的なＰａｇｅＲａｎｋの計算方法を説明すると、まず全てのＷｅｂページはそれぞれＰａｇｅＲａｎｋの値を持っている。そしてこの値はそのページがリンクしている先のページへ均等に分配されることになる。図３（ａ）を例に取ると、図中にある１００の値を持ったページは２つのページへのリンクを持っているので、このページの持つ１００の値は２つに分割されてリンク先へ与えられる。つまり、リンク先のページはそれぞれ５０ずつの値を得ることになる。

【００４４】
［２．１．８ 記事間の類似度による記事の重要度評価］
ＰａｇｅＲａｎｋがページ間のリンクの重みを平等に扱っているのに対し、本発明では各記事との類似度で重み付けを行う。これによって新聞記事群を関連度の強さに応じたリンクによって結ばれたグラフ構造と考える。そのなかから、より関連性が高いとしてリンクされている記事を、ＰａｇｅＲａｎｋ同様、遷移確率の最大固有値における固有ベクトルを算出することで求める。図３（ｂ）は 新聞記事間の類似度を示すものであり、その算出方法を以下に説示する。
まず、記事数をＮとするとき、情報間類似度計算部２１６が、Ｎ×ＮのＮ次正方行列、要素に各記事間の類似度をそれぞれ計算し、類似度行列計算部２１７が類似度行列を作成する。図３（ｂ）について、類似度行列を求めた結果である行列Ａを以下に示す。

【００４５】
【数９】

$$A=\begin{pmatrix}0&4&1&3\\4&0&2&5\\1&2&0&6\\3&5&6&0\end{pmatrix}$$

次に、各記事、すなわち各列について合計が１になるように正規化し行列Ａを状態遷移

確率行列Ｍとする。図３（Ｃ）は、新聞記事間の類似度による重み付けを行った遷移確率を示す。このときの記事間の関係は図３（Ｃ）のように示される。

【００４６】

【数１０】

$$M=\begin{pmatrix} 0 & 4/11 & 1/9 & 3/14 \\ 4/8 & 0 & 2/9 & 5/14 \\ 1/8 & 2/11 & 0 & 6/14 \\ 3/8 & 5/11 & 6/9 & 0 \end{pmatrix}$$

行列Ｍの状態遷移確率行列から、固有ベクトル作成部２１８が、最大固有値の固有ベクトルを計算した結果を図４に示す。

【００４７】

図４は本発明の実施形態に係る情報取得装置の記事の重要度計算例である。図４は、より多くの記事から高い重みで参照されている記事ほどスコアが高くなっていることを示している。この図４のスコアは、現在見ている記事から、その記事と類似度が高い記事を優先的に選択し、類似度に応じた確率で読み進めていくという仮想的なユーザを考えるとき、無限時間後に定常状態になった時点で、どの記事に行きつくかという確率に相当する。すなわち、その記事が類似性があるとしてユーザが興味を持ち、辿り着きやすいかというスコアであり、また、記事群の中でどの記事が多くの記事から類似性を持っているとして高い重みでリンクされているか、というのを示すスコアであるともいえる。

【００４８】

以上によって求められたスコアを、ベクトル空間モデル上の類似度を計算したスコアに併用することにより、質問ベクトルとの類似度でユーザの興味を考慮に入れつつ、そのなかで代表らしい記事を結果として示すことができる。これによりユーザが検索結果の判断に用いるのに適している記事を得て、効率良くフィードバック検索を行おうとするものである。

【００４９】

［２．２　情報内容統合システム］

続報情報検索部２０で得られた続報記事の情報データは、情報内容統合部３０において情報データの整理・分類処理される。その際に行われる重複箇所、固有箇所、補足説明の各カテゴリの設定、及びカテゴリ分けを行う類似度、文タイプによる判定について以下に説示する。

【００５０】

［２．２．１ カテゴリ設定及び分類］

内容統合において、 複数新聞記事を文カテゴリに分類し、それらの組み合わせにより、利用者の要望に沿った形式の出力を目指す。よって、その際の各カテゴリは、ユーザの情報取得の選択肢を広げ、複数新聞記事を比較する際の利点に沿ったものでなければならない。そこで、各記事に共通の箇所である重複箇所、各記事に固有の箇所である固有箇所と、記事中における補足的な内容である補足説明という合計３つのカテゴリを設定する。なお、重複箇所中の文の対応の定義として、一方の文に比較対象の文の話題が、完全にまたは部分的に含まれていることとする。

対象記事を「重複箇所」、「固有箇所」、「補足説明」の３ つのカテゴリに分類するために、その判定基準として、「文単位の類似度」、「文タイプ」の２つを用いる。

【００５１】

「文単位の類似度」では、記事データが、係り受け解析部３１で解析処理される。そして、名詞抽出部３２、名詞集合間類似度比較計算部３３により各文中の動詞をキーとした

名詞集合中の名詞単体の概念間の距離と表記を利用して求めた値と、名詞表示一致割合計算部３４により求めたそれらの結果を利用した名詞単語中の表記が同じ名詞の割合の合計を、文類似度計算部３５により算出された結果の値とする。

また「文タイプ」では、文タイプ選定部３６が、各文に対して文のタイプ付けを行う。以下に、それぞれの判定に関する詳細な説明を述べる。

【００５２】

［２．２．２ 複数新聞記事間における文単位の類似度］

重複箇所、固有箇所の選定の一基準として、文単位の類似度を採用している。以下に類似度の算出方法について述べる。一般に、文の類似度の指標には、構文構造の類似度と意味的な類似度が考えられる。類似文検索では、構文構造の類似度を求めるために「動詞への係り受け」を使用する。また、意味的な類似度を求めるために「動詞に直接係る文節中の名詞の意味属性」、「名詞表記の一致の割合」を利用する。類似文の検索は、次の４つのステップで行われる。

（１）動詞を含む文節に係る文節中の名詞の検出
（２）（１）で抽出した動詞をキーとする名詞集合毎の類似度の比較
（３）（２） の結果を利用した名詞表記の一致の割合
（４）（２）と（３）の結果を利用した類似度の算出

【００５３】

［２．２．２ 動詞を含む文節に係る文節中の名詞の抽出］

図５は、本発明の実施形態に係る情報取得装置の動詞を含む文節に係る文節中の名詞の抽出の例である。図５の場合には網掛け部分の名詞Ａ，Ｂ，Ｃ，Ｄ，Ｇ，Ｈ，Ｉ，７，８，９を抽出する。かかる処理は、文中に含まれる動詞に関する係り受けを利用することから、例えば日本語係り受け解析器ｃａｂｏｃｈａを用いて行うことができる。

【００５４】

［２．２．３ 動詞を含む文節に係る文節中の名詞の概念関係を利用した比較］

前述の日本語形態素解析器ｃａｂｏｃｈａにより部分的な重複を文間の類似度の情報に入れるため、各文中の動詞を含む文節に係る文節中の名詞を抽出し、各動詞に対する名詞集合を作成する。その際の集合中の各名詞間の類似度は、表記が異なるものはＥＤＲ電子化辞書により、概念間の距離からその値を求め、表記が同一のものはその値を最大値にする。そして、各名詞単体同士の類似度から名詞集合同士の類似度を算出し、その中で最も類似度が高い値をとる。名詞の類似度を測る方法としては、意味属性体系上での共通親属性の位置や、両意味属性間のパスの長さから類似度を求める方法が考えられる。しかし、一般に名詞には複数の意味属性を割り当てることができる。そのため、名詞の類似度を求めるために、その名詞がどの意味属性の名詞として使われているのかを、文脈情報などから一意に決定しなければならない。本発明においては、この多義性の問題には立ち入らずに、「ＥＤＲ電子化辞書を用いた単語類似度計算法」［参考文献：崔ら，情報処理学会報告ＮＬ－９３－１，ｐｐ１－６］で提案されている手法である名詞に割り当てられた複数の意味属性から総合的に名詞の類似度を求める。また、動詞に係る文節中の名詞を類似度の指標として特に取り上げているのは、一文に含まれる話題の数の違いを考慮したことによる。

【００５５】

［２．２．４ ＥＤＲ電子化辞書について］

図６は、本発明の実施形態に係る情報取得装置のＥＤＲ辞書の構造図である。ＥＤＲ電子化辞書は、コンピュータによる先進的な言語処理のために開発され、単語辞書などのいくつかの大規模な個別辞書から構成されている。辞書は、単語辞書中で定義した概念の類義を記述する概念体系（シソーラス）、辞書記述の典拠としてのコーパスＤＢ（例文集）を統合した日本語と英語の語彙知識総目録と呼ぶにふさわしい機械処理用の電子化辞書である。言語学的偏向を極力排除し、各種応用へのチューンアップの容易さを保持することを開発方針として採用してあるものである。ＥＤＲ電子化辞書は単語辞書、対訳辞書、概念辞書、共起辞書、専門用語辞書とＥＤＲコーパスから構成されている。

【００５６】
本発明では、名詞の概念間の距離を調べるために、概念辞書、及び日本語単語辞書を利用する。日本語単語辞書は約２６万語の語彙を持つ単語辞書である。基本的役割は、単語と概念（意味）との対応関係を記述し、この対応関係が成り立つときの文法的特性を与えることである。概念辞書は、単語辞書に語義として導入された約４１万の概念についての知識が記述され、情報の種類によって、概念体系辞書と概念記述辞書に分けられる。概念体系辞書は約４１万の概念に対して、それらの間の上位下位関係を記述したものである。上位下位関係とは概念間の包含関係であり、一種のシソーラスと見なすことができる。概念記述辞書は文中に共起する概念間（２項）の意味的関係（動作主、道具、場所、等）を整理したものを記述したものである。

【００５７】
［２．２．５ 名詞集合間の類似度を算出する処理］
名詞集合間の類似度を算出するための処理について以下に述べる。名詞同士の比較を行い、表記が同じものは類似度を最大値の１として算出する。それ以外の表記が異なるものがある場合には、概念辞書を利用した比較を行う。概念辞書による名詞の比較の手順を以下に示す。まず、名詞の概念を表す概念識別子を日本語単語辞書からとりだし、それを利用して概念辞書から意味属性のリストを得る。次に両名詞の持つ意味属性から名詞間の関係を「類似文の比較による省略可能な格要素の認定」［参考文献：篠原ら，情報処理学会研究報告，ＮＬ－１３９－１４，ｐｐ１０１－１０８］の提案による同義関係と類似関係とに分類する。

【００５８】
図７は、本発明の実施形態に係る情報取得装置の名詞間の同義・類似関係図である。図７（ａ）は、同義関係を示す。また、図７（ｂ）は類似関係を示す。この２つの関係に基づき、概念関係を利用した表記が異なる名詞間の類似度を求める。同義関係の類似度ａと類似関係の類似度ｂはそれぞれ次式により求める。各式については、篠原らの名詞間の概念関係の式を採用する。また篠原らは、他に同一関係という概念識別子が同一であるという関係を定義しているが、ＥＤＲ電子化辞書においてはかなり詳細に概念が定義されているので、同一関係というものは採用していない。
同義関係の類似度ａ
【００５９】
【数１１】

$$a = \frac{2D_a}{A_1 + A_2} \qquad (0 \le a \le 1)$$

A_n ： 名詞ｎの意味属性数（ｎ＝１，２）
D_a ： 重複する意味属性数
類似関係の類似度b
【００６０】
【数１２】

$$b = \frac{1}{N_1 N_2} \sum_{i,j=0}^{N_1,N_2} \frac{2D_{ij}}{N_{1i} + N_{2j}} \qquad (0 \le a \le 1)$$

N_n ： 名詞ｎの意味属性数
N_{ni}： 名詞ｎの意味属性ｉの上位概念数
D_{ij}：意味属性ｉ，ｊの上位概念の重複数
求めた類似度ａ、ｂを使用し、次式により概念間の距離による名詞同士の類似度S_1を求める。

概念間の距離による名詞同士の類似度S_1
【００６１】
【数１３】

$$S_1 = 1 - e^{-(a+b)} \qquad (0 \le S_1 \le 1)$$

以上より、単体の名詞間の類似度を求める。そして、以下にそれらを利用した動詞をキーとした名詞集合間の類似度の算出方法を述べる。
【００６２】
図８は、本発明の実施形態に係る情報取得装置の名詞集合間の類似度算出の例である。ｃａｂｏｃｈａにより得られた係り受け情報から、動詞が含まれる文節に係る文節の中の名詞句を動詞をキーとした組として取り出す。図８中の名詞集合１と名詞集合３との類似度を算出する際には、記事１を主体と考えた場合に、名詞Ａ と名詞Ｆ、Ｇ間で類似度が高い方を名詞Ａ に対する類似した名詞とし、ここでは名詞Ｆとする。同様に名詞Ｂも名詞Ｆ、Ｇ間で類似度が高い方を名詞Ｂに対する類似した名詞とし、ここでは名詞Ｇとする。そして、主体側の名詞の数をｎ、名詞Ａ と名詞Ｆの類似度をS_{AF}、名詞と名詞Ｇ の類似度をS_{BG} した場合に、名詞集合間の類似度をS_2 とすると、S_2は以下のようになる。
【００６３】
【数１４】

$$S_2 = \frac{S_{AF} + S_{BG}}{n} \qquad (0 \le S_2 \le 1)$$

同様に、名詞集合２と名詞集合３を比較し、集合間の類似度を求める。そこで名詞集合１と名詞集合３、名詞集合２と名詞集合３の類似度をそれぞれ比較し、値が高い方を動詞をキーとする名詞集合間の類似度とする。ここでは名詞集合１と名詞集合３の類似度S_2とする。
【００６４】
［２．２．６ 名詞表記の一致］
前工程では名詞の概念間の距離を利用して最も類似度が高い動詞をキーとした名詞集合を各文で選んだ。ここでは、そこで選んだ名詞集合以外の文中の名詞単語中の表記が同じ名詞の割合を算出する。以下にその類似度S_3 を示す。
表記の一致の割合によるS_3
【００６５】
【数１５】

$$S_3 = \frac{2D_{ij}}{A_i + A_j} \qquad (0 \le S_3 \le 1)$$

D_{ij}：文ｉ と文ｊ の動詞に係る文節以外の部分の名詞の内の一致した数
A_i： 文ｉ中の動詞に係る文節以外の部分の名詞の数
A_j：文ｊ中の動詞に係る文節以外の部分の名詞の数
【００６６】
［２．２．７ 類似度の算出］
文の類似度Ｓは前述のS_2 とS_3 により以下のようになる。
文の類似度Ｓ

【００６７】
【数１６】

$$S = S_2 + S_3 \qquad (0 \le S \le 1)$$

【００６８】
［２．２．８ 文タイプによる選定］
より新聞記事の特色を利用した重複箇所の選定方法として、各文に新聞記事の特徴を考慮した文タイプを設定し、それに基づいた重複文・固有文・補足説明の選定を行う新しい手法を提案する。この手法により、新たに新聞記事特有の言い回し、表現というものを選定の指標として採り入れることが可能となる。
【００６９】
［２．２．９ 文タイプの種類］
また、各文タイプは従来の要約処理において定義されていた多くの文タイプの中から、新聞記事の特徴から要旨、予定、理由、分析、補足説明の５つの文タイプを、また「日本語のシンクタンスと意味２」［参考文献：寺村秀夫：くろしお出版］の記載による概言のムードと上記データ解析から様態・伝聞、比況・推量の２つの文タイプを本発明の実施の一例とする。以下に要旨、予定、理由、分析、補足説明の５つの文タイプと様態・伝聞、比況・推量の２つの文タイプの特徴、判断基準等について述べる。
【００７０】
［２．２．１０ 文タイプ：要旨、予定、理由、分析、補足説明について］
要旨、予定、理由、分析、補足説明の５種類の文タイプについて、判断基準と特徴に付いて述べる。また判断基準に際し、断定的表現、日時を表す表現に関しては判断基準中においてはそれぞれ#[ｄａｎｔｅｉ], #[ｎｉｔｉｚｉ?]（?＝１ｏｒ２ｏｒ３）としている。断定的表現に関しては実験データ 記事を解析した結果以下のように設定する。なお、以下の判断基準の表記形式はｒｕｂｙの正規表現の表現形式に準ずる。
【００７１】
【表１】

である|だとしている|となっている|だ|です

また、日時を表す表現については、時間を表す部分（ｎｉｔｉｚｉ１）と季節や日付を表す語句（ｎｉｔｉｚｉ２）、そしてそれらに付随する語（ｎｉｔｉｚｉ３）に大きく分けて設定する。
【００７２】
【表２】

●ｎｉｔｉｚｉ１
((１|２|３|４|５|６|７|８|９|０|1|2|3|4|5|6|7|8|9|0|一|二|三|四|五|六|七|八|九|〇)＋(時|日|分|秒))
●ｎｉｔｉｚｉ２
(昨|今|来)(春|夏|秋|冬|日|月|年)
●ｎｉｔｉｚｉ３
(において|にて|に当たって|にかけて|(に際し|に際して)|(のうち|のうちに)|(の際|の際に)|(の点|の点で)|(の時|の時に)|までに|頃|中|から|，|まで|後|を目指し|に)

【００７３】
これらのうち時間を表す部分（ｎｉｔｉｚｉ１）と季節や日付を表す語句（ｎｉｔｉｚ

ｉ２）は一般的に考えられるものと実験データ２００記事から設定したものである。また、それらに付随する語（ｎｉｔｉｚｉ３）は実験データの解析と、「自然言語処理の基礎技術」［参考文献：野村浩郷，社団法人電子情報通信学会，ｐｐ２４６］に記載されている図７．７の格助詞総当語における時空関係群を参考に設定する。

要旨は新聞記事中で第一文としてある全体の要約が述べられていると考えられる文である。判断基準は記事中の第一文を要旨の文タイプとしている。

【００７４】

予定はその文がこれから行われる出来事の日時等を述べられているなど、その文が出来事の予定を表す際につけられる文タイプである。判断基準は前述のトレーニングデータを人手で分類した結果から以下のようになっている。

【００７５】

【表３】

．＊今後．＊|．＊(方針|見込み|計画|見通し|予定)＃{ｄａｎｔｅｉ}．
|．＊(＃{ｎｉｔｉｚｉ１}|＃{ｎｉｔｉｚｉ２})＃{ｎｉｔｉｚｉ３}．＊

理由はその文が理由を述べている場合に付けられる文タイプである。判断基準は前述のトレーニングデータを人手で分類した結果から以下のように作成する。

【００７６】

【表４】

．＊(という|が|を)(計画|狙い|目的|理由|動機|考え)
(＃{ｄａｎｔｅｉ}|．)．＊|．＊(と|という) もの。＊

分析はその文が記者の観点からみた出来事に対する分析、意見の場合につけられる文タイプである。判断基準は前述のトレーニングデータを人手で分類した結果から以下のように作成する。

【００７７】

【表５】

＊見通し＃{ｄａｎｔｅｉ}．|．＊見込(み|んでいる|まれている)．|
．＊(期待|予想|と)(している|されている|みている|みられている)．|
．＊(未知数|程度|確実|初めて|初|とみられるから|形|実情)
＃{ｄａｎｔｅｉ}．|．＊と(みられる|いえそうだ)．

【００７８】

補足説明はその文が記事の補足的な説明の場合に付けられる文タイプである。捕足説明には前述のトレーニングデータ記事を分析した結果、以下のような種類がある。

（１）記事における登場人物の素性の紹介
（２）記事内の出来事に関しての識者、関係者の話
（３）記事内容に対する捕足的説明(専門用語等)

補足説明の際には記事中においてその箇所に特殊な記号が記事中で使用されていること

に着目した。そこで、その記号が出現した後の部分を補足説明として文タイプを設定する。補足説明の判断に用いた特殊記号は、「○」、「＜」、「＞」、「＊」、「◇」、「◆」である。

【００７９】

［２．２．１１ 文タイプ:様態・伝聞、比況・推量について］

後述の様態・伝聞、比況・推量という２つの文タイプについてその採用理由と判断基準について述べる。様態・伝聞は様態と伝聞が合わさった文タイプである。様態とは物の存在や行動のありさまを伝える文タイプである。また、伝聞とは直接にではなく人から伝え聞いているような文タイプである。判断基準を以下に示す。

【００８０】

【表６】

．＊という．|．＊によれば．＊|．＊によると．＊|．＊と(話している|
述べている|している|説明している|いうこと#{ｄａｎｔｅｉ})．＊|．＊そうだ．＊

比況・推量は比況と推量が合わさった文タイプである。比況とは動作・状態などをほかのものにたとえて表すような文タイプである。また推量とはある根拠・理由や、確かな論理的要請などに基づいて、込み入った事情や人の心の中などをおしはかっているような文タイプである。判断基準を以下に示す。

【００８１】

【表７】

．＊らしい．|．＊(よう|みたい)#{ｄａｎｔｅｉ}．|．＊かもしれない．|．＊だろう。

これら２つの文タイプは、前述の「日本語のシンクタンスと意味２」に記載された二次的

【００８２】

これら２つの文タイプは、前述の「日本語のシンクタンスと意味２」に記載された二次的ムードの助動詞中の概言のムードより抜粋する。前述の「日本語のシンクタンスと意味２」によると、現実のいろいろな場で、話し手が、コトを相手の前にもち出すもち出し方、態度を表す部分を「ムード」という構文要素としている。前述の「日本語のシンクタンスと意味２」の記載ではムードとして確言のムード、概言のムード、説明のムードをあげている。ここでは、その中で新聞記事の文タイプとして話し手のいろいろな主観を表すという点で有効と考えられる概言のムード中の様態・伝聞、比況・推量を文タイプとする。様態、伝聞、比況、推量は従来の文法書において「助動詞」という項目の中で、他の形式、たとえば、ナイ、（ラ）レル、（サ）セル、タイ、タなどと並んで個別的にその用法が記述されてきたものである。すなわち、動詞連用形、形容詞語幹につく「ラシイ」は「（根拠のある）推量」、「ヨウダ」は「伝聞」を表す、とされる。文法書によっては、「ヨウダ」あるいは伝聞の「ソウダ」も、形式体言に形式体言「ダ」がついたものとし、助動詞とは認めないものもある。また様態の「ソウダ」を、接尾語に「ダ」がついたものとする見方もある。寺村は、これらを、一定の統語的特徴と、一定の（最大公約数的な）意味を共有するものとして統合的に扱っている。少数の形式については、統語的特徴から外れるが、意味的な特徴から見て、この中に入れる。また、ふつうの文法書では助動詞としては扱われない「カモシレナイ」「カモワカラナイ」「ニチガイナイ」「トイウ」なども、もともと助詞や動詞や助動詞であったものが結びつき、その結びつきが強くなって、一語化したものと見て、前述の統語的、意味的特徴から、やはり概言の助動詞の中に含める。このような理由から概言のムードである様態・伝聞、比況・推量を文タイプの例とする。

【００８３】

［２．２．１２ 文タイプの適用優先順位］

以上のように７種類の文タイプを設定したが、文によっては複数の文タイプを兼ねるものも多数存在する。その際に文タイプを設定する優先順位というものを考慮にいれる必要が出てくる。そこで、前述の実験データ２００記事に対して文タイプの優先順位が未実装である装置を試作し、各文に対し文タイプを設定した結果を示す。図９は、本発明の実施形態に係る情報取得装置の文タイプごとの割合である。このうち、要旨は記事全体の要約であるという性質上、各文に一文程度設定されていると考えられるので、優先順位は最上位とする。また補足説明も補足的な説明を表すという性質上優先順位を最上位とする。それ以外の文タイプ（予定、理由、分析、様態・伝聞、比況・推量）を出現数が少ないものから優先する。結果、優先順位については以下のようになる。

「要旨＝補足説明＞比況・推量＞理由＞分析＞予定＞様態・伝聞」

【００８４】

［２．２．１３ 文タイプによる重複文、固有文、補足説明の出力］

図１０は、本発明の実施形態に係る情報取得装置の重複箇所，固有箇所、補足説明のカテゴリ分けである。文タイプによる重複箇所、固有箇所、補足説明の出力までの流れである。図１０のように同じ文タイプがない場合には前の工程で算出した類似度を利用した重複文選定を行っている。また要旨、補足説明については先に述べた性質から前もって対象文タイプを抜き取る。

【００８５】

［２．２．１４ 複数記事間の集約］

文タイプ選定部３６から送られたデータは要約文作成部３７又は記事集約部３８に送られる。

【００８６】

［２．２．１４．１ ２記事間の集約］

２記事間の集約は、１つめの記事の重複箇所、固有箇所の文章に、２つめの記事の固有箇所を合わせた文章を２記事の集約としている。これにより、２記事間で重複しているものや、補足的なものをカットした集約ができる。

【００８７】

［２．２．１４．２ ３記事間の集約］

３つの記事がある場合、２記事同士の集約を３つ作成する。３つの記事を記事Ａ、記事Ｂ、記事Ｃとする。続報記事を分類するので、時系列順でみた初めの２記事Ａ、Ｂを集約したものを基本とする。この集約と、記事Ｂ、Ｃを集約したものを１文づつ比較し、含まれていない文を集約として追加する。これにより、記事Ａ、Ｂ、Ｃについての内容についての集約ができる。記事Ａ、Ｃの集約において、Ｃ、の重複箇所と判断されたものが、もし集約に含まれていた場合、それを削除することによって、新しい集約とする。もちろん、この手法を応用することで、３つ以上の記事を一度に集約することもできる。

【００８８】

［２．２．１４．３ 記事間の関連度を考慮に入れた集約］

上記の方法に、各記事間の関連度を考慮にいれて集約を行う。記事間の関連度は、各記事を１つのベクトルで表し、そのベクトルを比較して求める。検索を行う際に記事間の関連度を得るが、続報記事であるため、１つめの記事と２つめの記事、２つめの記事と３つ目の記事は繋がりがあっても、１つめの記事と３つめの記事の繋がりが弱い場合がある。記事同士の関連度が低い場合は、記事Ａ、Ｃの比較を行わないようにすることで、無駄な処理をしなくて済む。

【００８９】

［３．動作］

ユーザは結果を見ながら記事の適合，非適合の選択、あるいはパラメータを修正することで記事の重要度の変更を行う。システムはそれをフィードバックとして得て、質問ベクトルの修正を行い、順位付けをして再びユーザに検索結果を返す。

図１１は、本発明の実施形態に係る情報取得装置の続報記事検索のフローシートである。データが入力される（Ｓ１００）。形態素解析部２１１が形態素解析により品詞分解し

、名詞（普通、固有、サ変）のみを取り出す（Ｓ１１０）。ここで、形態素解析器茶筌を利用することもできる。ＴＦ計算部２１２が、各記事毎に単語とその単語の出現回数（ｔｆ）を記事データベースに登録し、ＴＦを計算する（Ｓ１２０）。ＩＤＦ計算部２１３が記事データベースに登録された全単語について文書頻度（ｄｆ）及びそのＩＤＦを計算する（Ｓ１３０）。単語重み計算部２１４が各記事の各単語についてＴＦ－ＩＤＦ法を用いて評価値を求め、単語重みを計算する（Ｓ１４０）。特徴ベクトル作成部２１５が評価値から各文書の特徴ベクトルを作成する（Ｓ１５０）。つまり、各記事はその記事に出現する全名詞のＴＦ－ＩＤＦ値を要素にもつベクトルである。情報間類似度計算部２１６が、全記事ベクトルの組み合わせについて類似度を計算する（Ｓ１６０）。類似度行列計算部２１７が類似度行列を計算する（Ｓ１７０）。固有ベクトル作成部２１８がその結果を数値解析プログラムＯｃｔａｖｅに渡し、そのベクトルの最大固有値の固有ベクトルを求めることで各記事の参照重要度を計算する（Ｓ１８０）。質問ベクトル作成部２１９が質問ベクトルを作成する（Ｓ１９０）。検索順位決定部２２が検索結果のための類似度計算により検索順位を決定する（Ｓ２００）。ここで、ある記事D_iのスコアは、ユーザからの質問ベクトルＱとの類似度によるスコアｓｉｍ（D_i，Ｑ）と固有ベクトルにより求められた重要度を掛け合わせることで求められる。

【００９０】

【数１７】

$$score(i) = \mu_i sim(D_i, Q)$$

μ_i：固有ベクトルにより求められた重要度

【００９１】

出力部４０が順位付け決定部でスコアリングされた結果をユーザに開示する（Ｓ２１０）。ユーザは検索結果をフィードバックするために検索結果の適否を入力する（Ｓ２２０）。適合・非適合判定部２３が検索結果は適合か否かを判定する（Ｓ２３０）。検索結果が適合でないと判定された場合に特徴及び質問ベクトルを修正する（Ｓ２４０）。検索が適合であると判定された場合に情報内容を統合する（Ｓ２５０）。なお、情報内容統合については以下に説示する。

【００９２】

図１２は、本発明の実施形態に係る情報取得装置の記事内容統合のフローシート（２）である。検索された記事が入力される（Ｓ２５１）。係り受け解析部３１が文中に含まれる動詞に関する係り受けを利用するために係り受け解析を行う（Ｓ２５２）。名詞抽出部３２が各文中の動詞を含む文節に係る文節中の名詞を抽出する（Ｓ２５３）。名詞集合間類似度比較計算部３３が単体の名詞間の類似度S_1を計算する（Ｓ２５４）。また、名詞集合間類似度比較計算部３３が名詞集合間の類似度S_2を計算する（Ｓ２５５）。名詞表示一致割合計算部３４が名詞表示一致割合の類似度S_3を計算する（Ｓ２５６）。文類似度計算部３５が類似度S_2及び類似度S_3から文類似度Ｓを計算する（Ｓ２５７）。文タイプ選定部３６が文タイプによる選定を行う（Ｓ２５８）。要約文作成部３７が記事内容を要約する（Ｓ２５９）。記事集約部３８が記事内容を集約する（Ｓ２６０）。ここで、記事内容は、要約されたかどうかに関わらず集約することができる。出力部４０が内容統合記事の出力を行う（Ｓ２６１）。なお、検索結果及び内容統合記事及び出力の具体的な内容を以下に説示する。

【００９３】

［４．可視化手法］

本発明の実施の形態に係る情報取得装置は、情報を分類する機能を用いて続報記事の情報をユーザに見やすいように整理し、その機能に応じたインターフェースを表示する。そこで、以下にその内容を詳説する。

図１３は、本発明の実施形態に係る情報取得装置の実行図である。情報検索においてイ

ンタラクションを促進する関連技術として、情報の可視化は欠くことのできない存在である。可視化によって、システムが提示するデータを効率的にユーザに伝えることができるだけでなく、ユーザのより柔軟なデータへのアクセスが可能になる。Ａは関連記事のタイトル表示、Ｂは記事内容の表示、Ｃはレーダーチャート、Ｄは各種コマンド、Ｅは記事間の関連表示である。

【００９４】

図１４は、本発明の実施形態に係る情報取得装置の検索式拡張・質問ベクトル選定のためのインターフェースである。図１４（ａ）は、図１３のＡの関連記事のタイトル表示であり、検索式拡張のためのインターフェースを示す。検索式拡張は、一旦検索した結果に対し利用者が適合文書であったか非適合文書であったかをフィードバックされた結果に基づいて検索式を拡張して、再度検索する。そのために利用者が検索された結果に対し、フィードバックを行うことができる。検索結果一覧表示をした画面に対して適合・非適合の入力を受け付けるように各々の記事の適合・非適合を選択可能とする。図１４（ａ）に示すように、文書検索結果としてユーザ画面上には検索結果の記事のタイトルがリスト表示される。ユーザがボタンを押すと、図１３のＢの本文表示用領域に本文が表示され、確認しながら適合・非適合を選択することができる。

【００９５】

図１４（ｂ）に図１３のＣのレーダーチャートである質問ベクトル選定のためのインターフェースを示す。質問ベクトルの選定は質問ベクトルとその元となった記事ベクトルとの類似度をレーダーチャートで示すことで行う。レーダーチャートの中心に向かう程、質問ベクトルとの類似性が低く、逆に外側に向かう程、類似性が高くなるように配置している。このように表示することで、質問ベクトルを発散させている記事は凹型になってあらわれる。そういった記事を質問ベクトルから外す、もしくはレーダーチャートの頂点をマウスでドラッグしてその記事に対する重みを補正することで、クエリベクトルの洗練を行う。なお、関連関係と関連度の関連計算は、ＴＦ／ＩＤＦ、ベクトル空間での類似度の判定、統計的手法による類似度の判定、ＰａｇｅＲａｎｋによる記事の重要度の判定、などを総合して行う。それぞれの関連計算をコントロールするためのものがレーダーチャートである。スクリーンの広さの制約から、レーダーチャートは二つのみ表示されている。これらのレーダーチャートは、上記の判定計算のいずれにも入れ替えることができる。また、サイズを小さくして、前記の四つを表示することもできる。関連計算のコントロールは、総合判定のとき、ＴＦ／ＩＤＦ、ベクトル空間での類似度の判定、統計的手法による類似度の判定、ＰａｇｅＲａｎｋによる記事の重要度の判定でそれぞれの重み付けを変えたり、それぞれの関連計算においてそれらの計算要素の重み付けを変えたりするものである。これらのコントロールは、スクリーン上でマウスなどの入力手段を使って行う。レーダーチャートの軸は、ＴＦ／ＩＤＦ、ベクトル空間での類似度の判定、統計的手法による類似度の判定、ＰａｇｅＲａｎｋによる記事の重要度の判定計算の計算要素である。軸上の値が円周に近いほど「重みの値が大きく」その計算要素が重要視される。軸上の値が円の中心に近いほどその計算要素の重要度が小さくされる。総合判定に関するレーダーチャートもある。このときの軸は、ＴＦ／ＩＤＦ、ベクトル空間での類似度の判定、統計的手法による類似度の判定、ＰａｇｅＲａｎｋによる記事の重要度の判定計算である。軸上の値は、それぞれの重要度を制御するための重みである。以上のインターフェースを提供することで絞りこみ等を行うことができる。

【００９６】

図１５は、本発明の実施形態に係る情報取得装置の検索結果表示、続報記事発見のためのインターフェースである。

図１３のＥには記事間の関連表示である検索結果をグラフィカルに表示する領域をそなえている。Ｘ軸に時間、Ｙ軸に検索ベクトルに対するスコアをとり、その空間上に記事を表す点を配置している。また、記事を表す点をクリックすることで、その記事本文を図１３のＢの本文表示用領域に表示する。各記事からはその記事に対する関連性の強さに応じたリンクが結ばれている。各リンクの関連性の強さを線の太さや種類等で識別することが

できる。図１５では関連性の強いものを実線で表し、関連性の弱いものを点線で表している。ここで、各リンクは強さに応じて色分けをすることもできる。また、ユーザの全体的な興味に対する指標をＹ軸の座標で、ユーザの局所的な興味に対する指標を記事間を結ぶリンクで表現しているため、例えば、ユーザはこの中から、できるだけＹ座標が大きく、また、現在見ている記事とのリンクが関連の強い色のリンクで結ばれた記事を読み進めることで、関連記事の中から続報性を持つ記事を発見することができる。

【００９７】

なお、記事内容の時間的経緯を考慮することもできる。図１３のスクリーンショットのグラフでは、新聞記事を例としているため、横軸が日付となっている。縦軸は、記事の重要度を示し、上部に表示されているほど重要度が高く、下部に表示されているほど重要度が低い。グラフ上の点は記事を表し、点と点を結ぶ線は「関連」があることを示している。点をクリックすると、記事内容が図１３のＢのウィンドに表示される。点と点を結ぶ線は実線は最も関連度が高いことを示し、点線は最も関連度が低いことを示す。なお、関連性が高い線を赤色に、関連性が低い線を黄色に着色することもできる。さらに、この線の彩色は、より顕示性を上げるために、赤色から青色へのスペクトル変化に対応させることもできる。

【００９８】

図１６（ａ）は本発明の実施形態に係る情報取得装置の複数の記事本文表示のためのインターフェース及び（ｂ）新たに記事の本文を表示する場合の表示方法の図である。図１３のＢの記事本文表示領域には指定した複数の記事本文を同時に表示するためのインターフェースがある。ユーザが本文を表示したい記事を左クリックすると、複数の表示領域の中で、記事本文が表示されていない領域に指定した記事の本文が表示される。もしも３つの表示領域が埋まっていた場合、左側の領域を初期化しその後右側の領域の記事本文を表示する。そして指定した記事本文を右側の領域に表示する。

【００９９】

図１７は、本発明の実施形態に係る情報取得装置の記事選択補助のための見出し表示領域である。Ａによってリストアップされた複数の新聞記事について、図１３のＤには記事の見出しを表示する領域と複数の記事を比較したり、関連度を計算させたり、要約させたり、集約させたり、再検索させたり、要約または集約した状態から元の記事に戻したりるための各種コマンドのボタンを用意する。なお、図１７では、一例として３つのウィンドの場合を示すが、ウィンドの数はいくつであってもよい。また、ユーザが記事を右クリックすると記事の見出しが表示される。これによって表示している記事の本文を変えずに見たい記事を探すことができる。

【０１００】

図１８は、本発明の実施形態に係る情報取得装置の記事を集約した結果を表示するインターフェースである。記事を集約する方法は、複数の記事同士を１つの文書とみなして、文章構文解析と要約文を作成する。文章構文解析では、意味段落を作成し、意味段落の連接関係を作成しながら、連接関係を崩さずに文章構成を再編し、意味段落の飛地構造解析も行う。文章構造解析された意味段落から、陳述形式による重要句を評価し、重み付けした句を抽出し、さらに、語の類似度を考慮した句の抽出を行い、語の補完を行うことで、集約した文章を作成する。なお、集約を行わず、各文書の要約のみを行い、必要ならば分類タグを付け、文書の分類整理に使うこともできる。分類タグは、例えば、ＴＦ／ＩＤＦの計算で得られた重要語から作成したり、ベクトル空間の計算のときのベクトルから作成したり、統計的手法による類似度計算に使った重みの値が高い重要要素から作成したりできる。

【０１０１】

図１９は、本発明の実施形態に係る情報取得装置の関連記事検索結果の一例である。Ａの記事リストの一番上のものを指定して関連記事を検索したものである。二番目以下の記事が検索された関連記事のリストである。関連記事の検索開始時に指定するものは、記事でもよいし（Ａの記事リストの一番上のもの）、自由に記述した文章でもよいし、キーワ

ードの組み合わせでもよい。関連関係と関連度の計算は、下に述べるユーザ・コントロールが行われていないときは、システムのデフォルト値を使って行われる。Ｅには、記事間の関連関係と関連度のグラフが表示されている。Ｂには、Ａで表示指定するか、またはＥのグラフ内の点をクリックしたときに、それらの記事内容が表示される。Ｂのサブウィンドの数はいくつでもよい。スクリーンショットでは、３つのサブウィンドが表示されている。Ｂの上部に選択ボタンが示されているように、現在は、サブウィンドの数は、１つ、２つ、３つの３種類を選択できるようになっている。サブウィンドの数をさらに増やすと、サブウィンドが細い縦長になり、読みづらくなる。レーダーチャートは、デフォルト値が表示されている。

【０１０２】

図２０は、本発明の実施形態に係る情報取得装置の各記事を要約した一例である。Ｂの各記事を要約したものを表示している。関連関係と関連度を計算するとき、記事の長さが短いほうが計算速度が速くなるため、要約する。また、要約により記事の重要な情報に絞られているため、集約処理の品質がよくなる。

【０１０３】

図２１は、本発明の実施形態に係る情報取得装置の３つの要約文書を１つに集約した一例である。Ｂの３つの要約文書を１つの文書に集約したものを独立なウィンドに表示している。集約は、要約をしない記事についておこなうこともできる。

【０１０４】

図２２は、本発明の実施形態に係る情報取得装置の集約結果の検討の一例である。集約文書をＢの右端のウィンドに移し、集約結果を検討し、必要に応じて、真中や左端の記事を新しく選択し表示する。これらの記事について、関連関係と関連度を再計算し、その結果に基づいて再検索し、新しい関連関係と関連度をグラフ表示する。その結果として、Ｅのグラフが更新される。必要に応じて、このような操作を繰り返し、最終的な集約を得る。この集約が情報検索の結果である。すなわち、いわゆる情報検索の結果は、一つの文書として出力される。これは、現在の多くの情報検索システムがＵＲＬのリストを情報検索の出力としているのとはまったく異なるものである。

【０１０５】

以上の前記実施形態により本発明を説明したが、本発明の技術的範囲は実施形態に記載の範囲には限定されず、これら各実施形態に多様な変更又は改良を加えることが可能である。そして、かような変更又は改良を加えた実施の形態も本発明の技術的範囲に含まれる。このことは、特許請求の範囲及び課題を解決する手段からも明らかなことである。

【図面の簡単な説明】

【０１０６】

【図１】本発明の実施形態に係る情報取得装置のハードウェア構成図である。

【図２】本発明の実施形態に係る情報取得装置のブロック構成図である。

【図３】本発明の実施形態に係る情報取得装置の記事間の類似度による記事の重要度評価の説明図である。

【図４】本発明の実施形態に係る情報取得装置の記事の重要度計算例である。

【図５】本発明の実施形態に係る情報取得装置の動詞を含む文節に係る文節中の名詞の抽出の例である。

【図６】本発明の実施形態に係る情報取得装置のＥＤＲ辞書の構造図である。

【図７】本発明の実施形態に係る情報取得装置の名詞間の同義・類似関係図である。

【図８】本発明の実施形態に係る情報取得装置の名詞集合間の類似度算出の例である。

【図９】本発明の実施形態に係る情報取得装置の文タイプごとの割合である。

【図１０】本発明の実施形態に係る情報取得装置の重複箇所、固有箇所、補足説明のカテゴリ分けである。

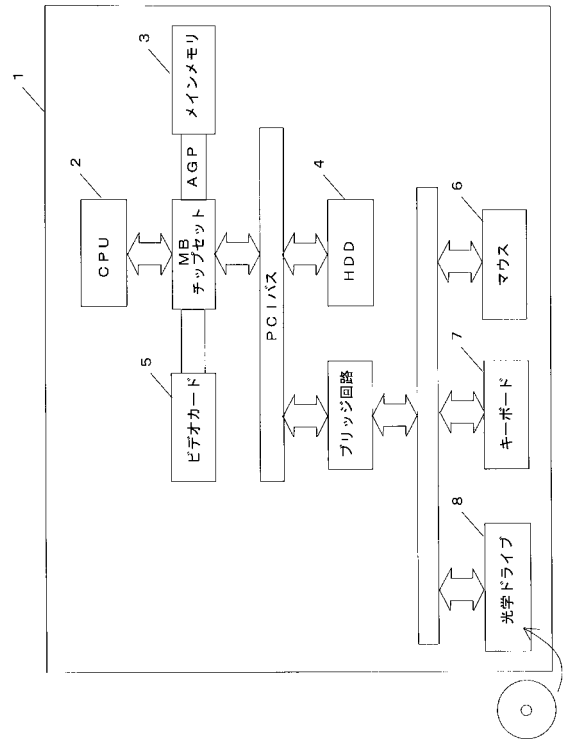
【図１１】本発明の実施形態に係る情報取得装置の続報記事検索のフローシートである。

【図１２】本発明の実施形態に係る情報取得装置の記事内容統合のフローシートである。

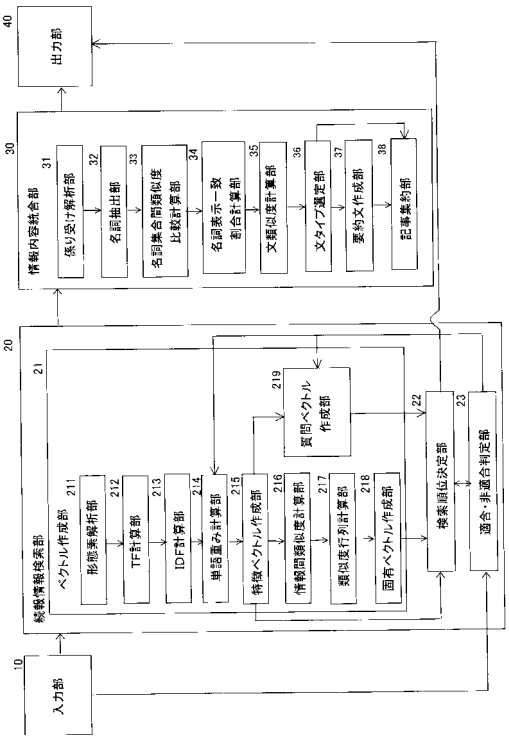
【図１３】本発明の実施形態に係る情報取得装置の実行図である。

【図１４】本発明の実施形態に係る情報取得装置の検索式拡張・質問ベクトル選定のためのインターフェースである。
【図１５】本発明の実施形態に係る情報取得装置の検索結果表示、続報記事発見のためのインターフェースである。
【図１６】本発明の実施形態に係る情報取得装置の複数の記事本文表示のためのインターフェース及び新たに記事の本文を表示する場合の表示方法の図である。
【図１７】本発明の実施形態に係る情報取得装置の記事選択補助のための見出し表示領域である。
【図１８】本発明の実施形態に係る情報取得装置の記事を集約した結果を表示するインターフェースである。
【図１９】本発明の実施形態に係る情報取得装置の関連記事検索結果の一例である。
【図２０】本発明の実施形態に係る情報取得装置の各記事を要約した一例である。
【図２１】本発明の実施形態に係る情報取得装置の３つの要約文書を１つに集約した一例である。
【図２２】本発明の実施形態に係る情報取得装置の集約結果の検討の一例である。
【符号の説明】
【０１０７】
１　コンピュータ
２　ＣＰＵ
３　メインメモリ
４　ＨＤＤ
５　ビデオカード
６　マウス
７　キーボード
８　光学ディスク
１０　入力部
２０　続報情報検索部
２１　ベクトル作成部
２２　検索順位決定部
２３　適合・非適合判定部
３０　情報内容統合部
３１　係り受け解析部
３２　名詞抽出部
３３　名詞集合間類似度比較計算部
３４　名詞表示一致割合計算部
３５　文類似度計算部
３６　文タイプ選定部
３７　要約文作成部
３８　記事集約部
４０　出力部
２１１　形態素解析部
２１２　ＴＦ計算部
２１３　ＩＤＦ計算部
２１４　単語重み計算部
２１５　特徴ベクトル作成部
２１６　情報間類似度計算部
２１７　類似度行列計算部
２１８　固有ベクトル作成部
２１９　質問ベクトル作成部

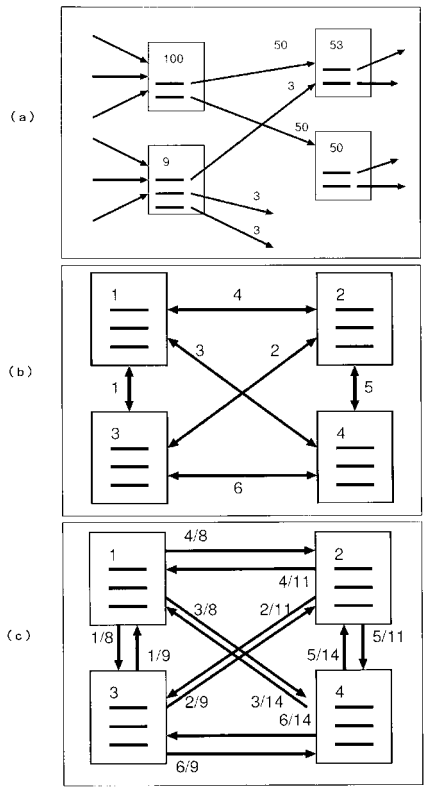
【図１】



【図２】



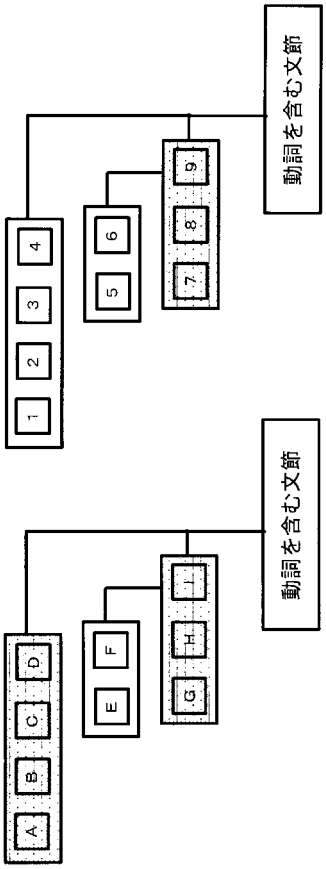
【図３】



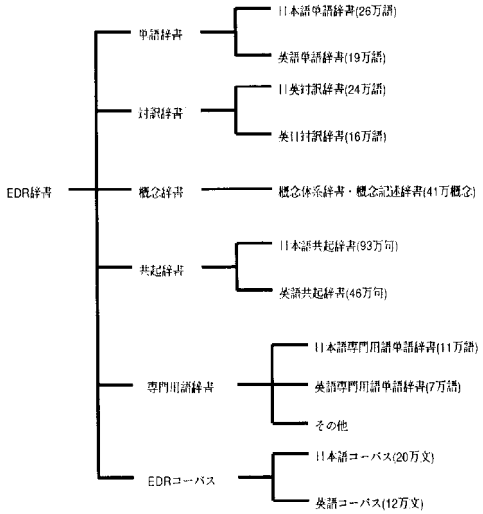
【図４】

文書番号	発リンク	被リンク（合計）	PageRank	順位
1	1/8, 1/8, 3/8	1/11, 1/9, 3/14 (0.6688)	0.19773	1
2	1/11, 2/11, 5/11	1/8, 2/9, 5/14 (1.0793)	0.26007	2
3	1/9, 2/9, 6/9	1/8, 2/11, 6/14 (0.7353)	0.21689	3
4	3/11, 5/14, 6/14	3/8, 5/11, 6/9 (1.4961)	0.32131	1

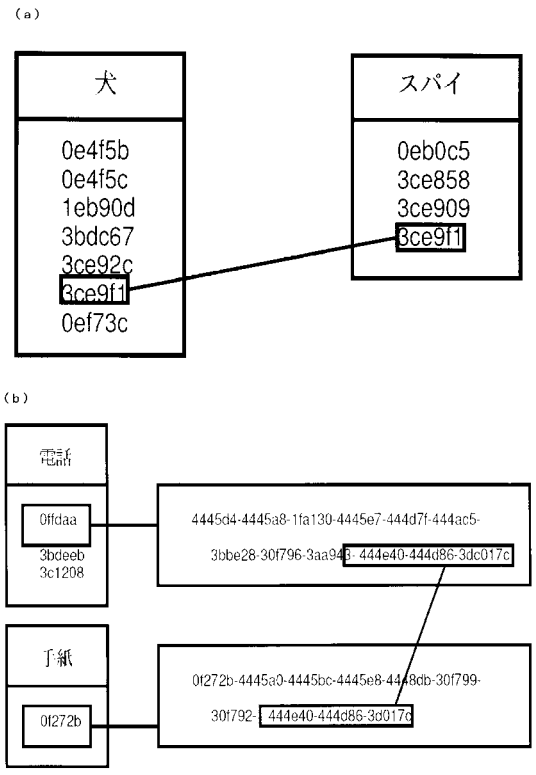
【図５】


動詞を含む文節
1
2
3
4
5
6
7
8
9
動詞を含む文節
A
B
C
D
E
F
G
H
I

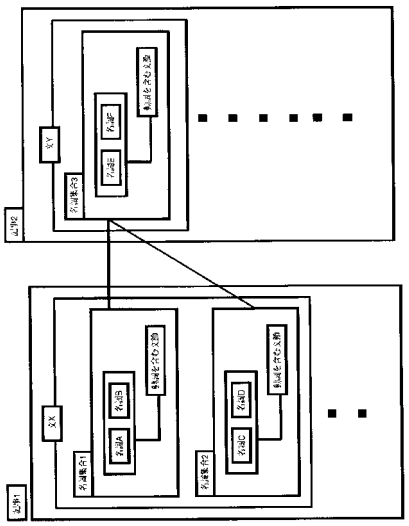
【図６】


EDR辞書
単語辞書
日本語単語辞書(26万語)
英語単語辞書(19万語)
対訳辞書
日英対訳辞書(24万語)
英日対訳辞書(16万語)
概念辞書
概念体系辞書・概念記述辞書(41万概念)
共起辞書
日本語共起辞書(93万句)
英語共起辞書(46万句)
専門用語辞書
日本語専門用語単語辞書(11万語)
英語専門用語単語辞書(7万語)
その他
EDRコーパス
日本語コーパス(20万文)
英語コーパス(12万文)

【図７】


(a)
犬
0e4f5b
0e4f5c
1eb90d
3bdc67
3ce92c
3ce9f1
0ef73c
スパイ
0eb0c5
3ce858
3ce909
3ce9f1
(b)
電話
0ffdaa
3bdeeb
3c1208
4445d4-4445a8-1fa130-4445e7-444d7f-444ac5-
3bbe28-30f796-3aa943-444e40-444d86-3dc017c
手紙
0f272b
0f272b-4445a0-4445bc-4445e8-4448db-30f799-
30f792-444e40-444d86-3d017c

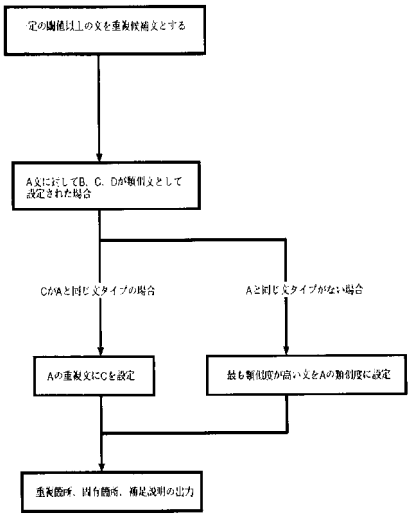
【図８】



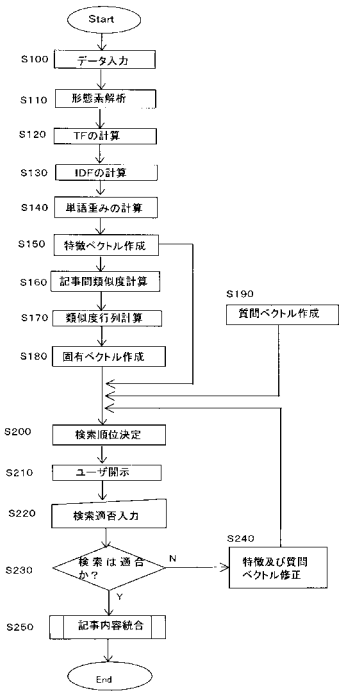
【図９】

文タイプ名	数	割合(%)
要旨	199/1396	14.26
予定	109/1396	7.81
理由	11/1396	0.79
分析	69/1396	4.94
補足説明	105/1396	7.52
様態・伝聞	116/1396	8.31
比況・推量	3/1396	0.21
その他	784/1396	56.16

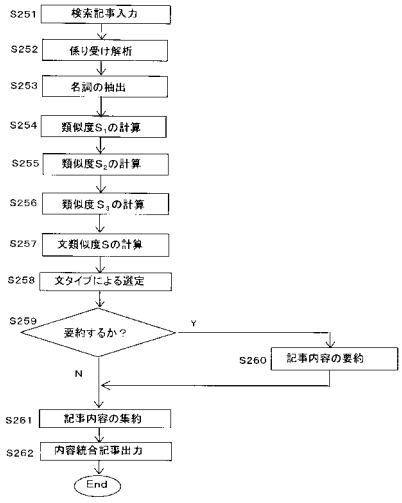
【図１０】



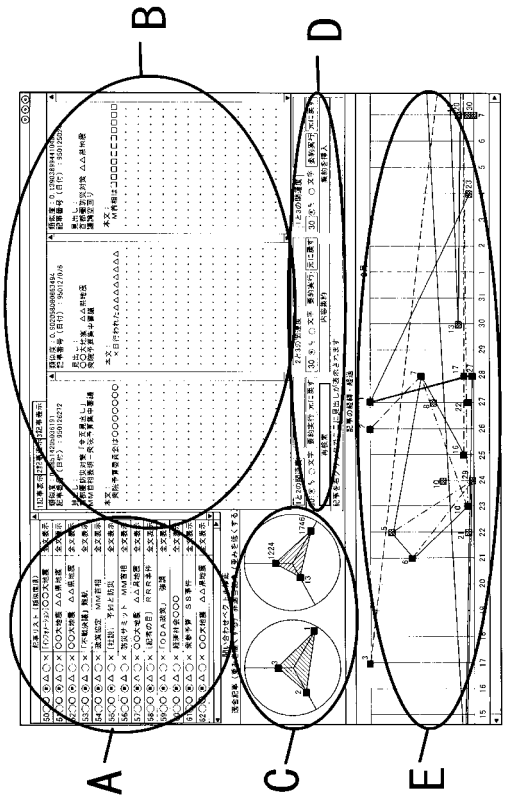
【図１１】



【図１２】



【図１３】

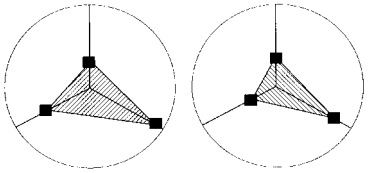


【図１４】

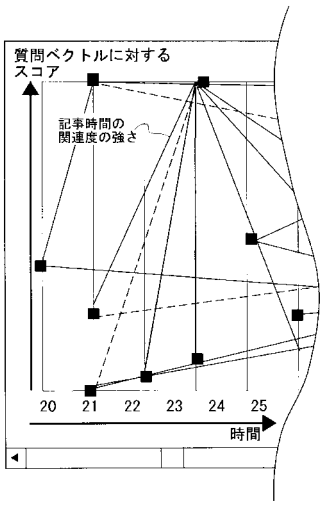
(a)

記事リスト（類似度順）		
50	[インフォメーション]○○大地震	全文表示
51	○○大地震 △△県地震	全文表示
52	○○大地震 △△県地震	全文表示
53	「不戦決議」難航	全文表示
54	政策協定 MM首相	全文表示
55	[社説] 予知と防災	全文表示
56	防災サミット MM首相	全文表示
57	○○大地震 △△県地震	全文表示
58	[記者の目] RRR事件	全文表示
59	「ODA政策」 強調	全文表示
60	経済社会○○○	全文表示
61	衆参予算 SS事件	全文表示
62	○○大地震 △△県地震	全文表示

(b)

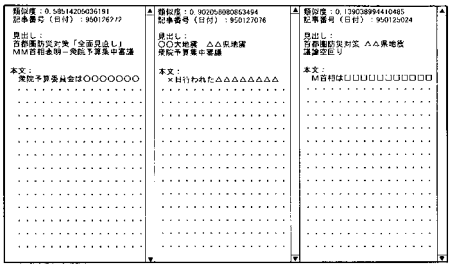


【図１５】

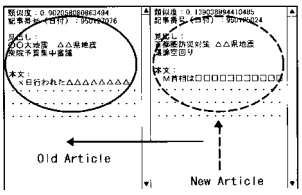


【図１６】

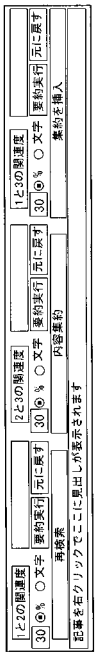
(a)



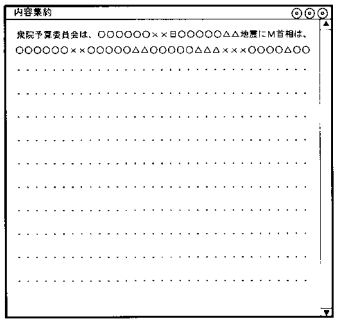
(b)



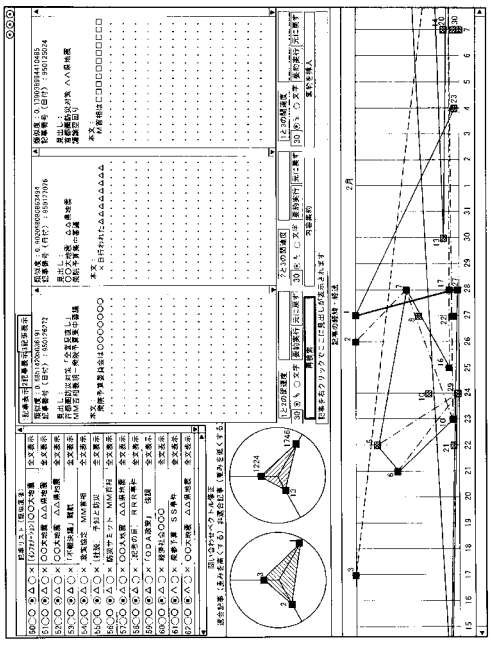
【図１７】



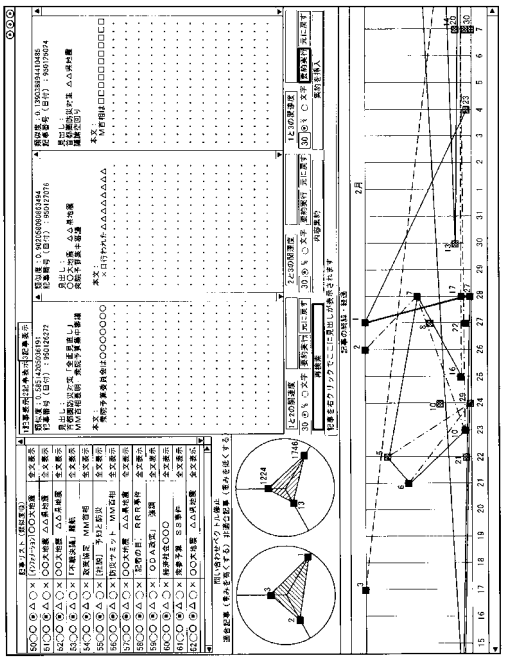
【図１８】



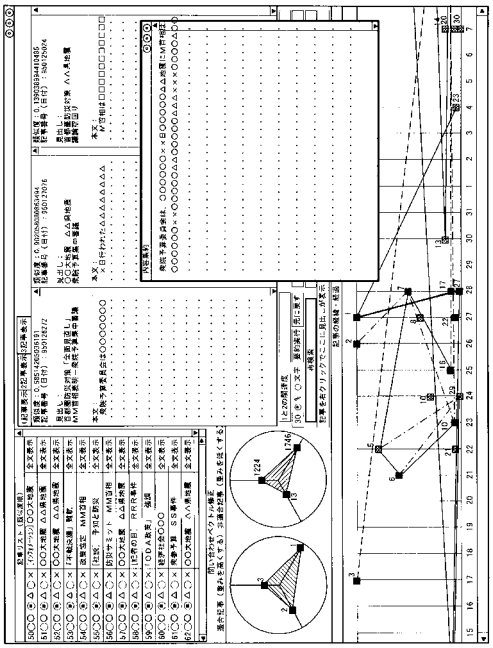
【図１９】



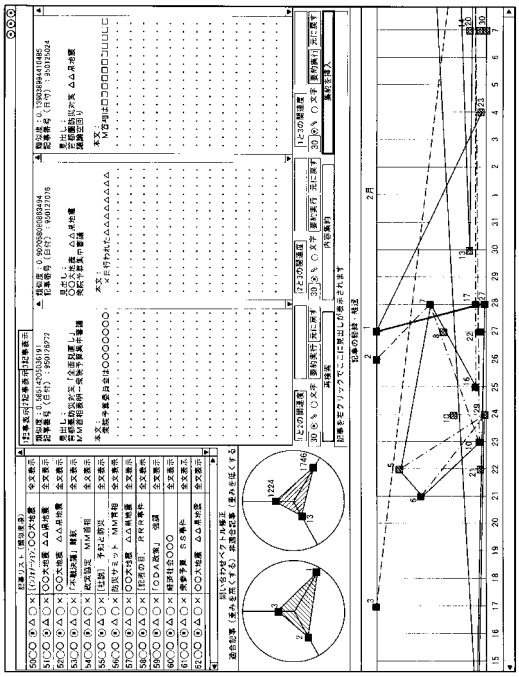
【図２０】



【図２１】



【図２２】



フロントページの続き

(56)参考文献 特開２００５－３２７２２５（ＪＰ，Ａ）
特開２００４－１８５５１５（ＪＰ，Ａ）
特開平１０－１３４０６６（ＪＰ，Ａ）
特開２００３－２２８５８１（ＪＰ，Ａ）
荒谷寛和、外２名，ウェブページ間類似度に基づく推薦リンクを用いたウェブ検索システムの設計，電子情報通信学会技術研究報告（ＡＩ２００４－１２～１８），日本，社団法人電子情報通信学会，２００４年 7月２２日，第１０４巻，第２３３号，ｐ．７－１２
中村貞吾、外２名，文タイプと文間関係に基づく要約処理，言語処理学会第４回年次大会ワークショップ論文集，日本，言語処理学会，１９９８年 3月２７日，ｐ．５０－５５
中山聡、外４名，ＥＤＲコーパスを利用した動詞の語義分類，電子情報通信学会技術研究報告（ＮＬＣ９５－４０～４５），日本，社団法人電子情報通信学会，１９９５年１０月２０日，第９５巻，第３２１号，ｐ．２３－３０
平田陽一、外３名，Ｗｅｂ検索における意味的適合フィードバック機構，情報処理学会研究報告（２０００－ＤＢＳ－１２２），日本，社団法人情報処理学会，２０００年 7月２８日，第２０００巻，第６９号，ｐ．１３７－１４４
篠原直嗣、外２名，類似文の比較による省略可能な格要素の認定，情報処理学会研究報告（２０００－ＮＬ－１３９），日本，社団法人情報処理学会，２０００年 9月２２日，第２０００巻，第８６号，ｐ．１０１－１０８
金田晃征、外１名，情報検索と情報集約による情報取得システム，情報処理学会研究報告（２００７－ＮＬ－１７９），日本，社団法人情報処理学会，２００７年 5月２４日，第２００７巻，第４７号，ｐ．３１－３６

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １７／３０