(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5143057号
(P5143057)
(45)発行日　平成25年2月13日(2013.2.13)　(24)登録日　平成24年11月30日(2012.11.30)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 210A
G06F 17/30 170A

請求項の数 9　(全 28 頁)

(21)出願番号　特願2009-48550(P2009-48550)
(22)出願日　平成21年3月2日(2009.3.2)
(65)公開番号　特開2010-204866(P2010-204866A)
(43)公開日　平成22年9月16日(2010.9.16)
審査請求日　平成22年5月28日(2010.5.28)

(73)特許権者　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(74)代理人　100070150
弁理士　伊東　忠彦
(74)代理人　100124844
弁理士　石原　隆治
(72)発明者　近藤　光正
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　中辻　真
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

最終頁に続く

(54)【発明の名称】重要キーワード抽出装置及び方法及びプログラム

(57)【特許請求の範囲】
【請求項１】
Ｗｅｂ文書中に含まれるテキストから重要なキーワードを抽出する重要キーワード抽出装置であって、
前記Ｗｅｂ文書を取得して該Ｗｅｂ文書の特徴量を抽出し、主要コンテンツを抽出する主要コンテンツ抽出手段と、
オンライン百科辞典に代表される、文書集合内において一意の見出し語を持つ文書で、かつ、文書集合内において引用関係もしくは参照関係を持つ文書集合（以下、辞典文書集合と記す）の見出し語を形態素解析用の辞書として登録しているキーワード辞書を参照して、前記主要コンテンツから重要キーワード候補を抽出する重要キーワード候補抽出手段と、
前記Ｗｅｂ文書内の前記重要キーワード候補に重みを付ける出現頻度算出手段と、
知名度や話題性が高く、キーワードについて語る際に様々な話題が挙がる情報量が豊富なキーワードの重要度が高くなるように前記重要キーワード候補の固有重要度を算出し、第１の記憶手段に格納するキーワード重要度算出手段と、
前記重要キーワード候補の前記Ｗｅｂ文書中の位置に基づいて出現位置キーワード重要度を求め、第２の記憶手段に格納する位置情報算出手段と、
前記出現頻度算出手段による重要キーワード候補の重み、前記第１の記憶手段に格納されている重要キーワード候補の固有重要度及び前記第２の記憶手段に格納されている前記出現位置キーワード重要度を乗算した値に基づいて、最終重要度付きのキーワード集合を

出力するキーワード出力手段と、を有し、
前記主要コンテンツ抽出手段は、
前記Ｗｅｂ文書を所定の分割規則に基づいてセグメントに分割するＷｅｂ文書分割手段と、
前記セグメント毎に主要コンテンツ判定のための特徴量を抽出する特徴量抽出手段と、
前記セグメント毎の特徴量に基づいて機械学習アルゴリズムを用いて主要コンテンツか否かの判定を行う主要コンテンツ判定手段と、
主要コンテンツと判断された部位を結合して前記主要コンテンツとして出力する主要コンテンツ出力手段を含み、
前記特徴量抽出手段は、
前記Ｗｅｂ文書で表示される文字列の特徴量、タグ情報の特徴量、アンカーリンク情報の特徴量を抽出し、各特徴量間の比率を用いて最終的な特徴量とする手段を含み、
前記出現頻度算出手段は、
検索エンジンが収集したＷｅｂ文書集合における出現確率を考慮した重みであるＷｅｂIDFを用いて前記重要キーワード候補に重みを付ける手段、
または、
前記重要キーワード候補の前記Ｗｅｂ文書内の出現頻度を求め、該出現頻度と文書集合内での出現分布に基づいて各重要キーワード候補に重み付けを行うBM25を用いて該重要キーワード候補に重みを付ける手段、
のいずれかを有することを特徴とする重要キーワード抽出装置。
【請求項２】
前記出現頻度算出手段は、
形態素解析手段の形態素解析結果から得られた形態素の固有名詞を含む前記重要キーワード候補の出現頻度を用いて、該重要キーワード候補に重みを付ける手段を含む
請求項１記載の重要キーワード抽出装置。
【請求項３】
前記キーワード重要度算出手段は、
前記辞典文書集合内のリンク構造を用いて、前記重要キーワード候補の固有重要度を求める手段を含む請求項１または２記載の重要キーワード抽出装置。
【請求項４】
前記キーワード重要度算出手段は、
検索エンジンに投入された回数が多いキーワードほど、重要なキーワードとなるような値を用いて、前記重要キーワード候補の固有重要度を求める手段を含む請求項１乃至３のいずれか１項記載の重要キーワード抽出装置。
【請求項５】
Ｗｅｂ文書中に含まれるテキストから重要なキーワードを抽出する重要キーワード抽出方法であって、
主要コンテンツ抽出手段が、前記Ｗｅｂ文書を取得して該Ｗｅｂ文書の特徴量を抽出し、主要コンテンツを抽出する主要コンテンツ抽出ステップと、
重要キーワード候補抽出手段が、オンライン百科辞典に代表される、文書集合内において一意の見出し語を持つ文書で、かつ、文書集合内において引用関係もしくは参照関係を持つ文書集合（以下、辞典文書集合と記す）の見出し語を形態素解析用の辞書として登録しているキーワード辞書を参照して、前記主要コンテンツから重要キーワード候補を抽出する重要キーワード候補抽出ステップと、
出現頻度算出手段が、前記Ｗｅｂ文書内の前記重要キーワード候補に重みを付ける出現頻度算出ステップと、
キーワード重要度算出手段が、知名度や話題性が高く、キーワードについて語る際に様々な話題が挙がる情報量が豊富なキーワードの重要度が高くなるように前記重要キーワード候補の固有重要度を算出し、第１の記憶手段に格納するキーワード重要度算出ステップと、

位置情報算出手段が、前記重要キーワード候補の前記Ｗｅｂ文書中の位置に基づいて出現位置キーワード重要度を求め、第２の記憶手段に格納する位置情報算出ステップと、
キーワード出力手段が、前記出現頻度算出ステップによる重要キーワード候補の重み、前記第１の記憶手段に格納されている重要キーワード候補の固有重要度及び前記第２の記憶手段に格納されている前記出現位置キーワード重要度を乗算した値に基づいて、最終重要度付きのキーワード集合を出力するキーワード出力ステップと、を行い、
前記主要コンテンツ抽出ステップでは、
前記Ｗｅｂ文書を所定の分割規則に基づいてセグメントに分割するＷｅｂ文書分割ステップと、
前記セグメント毎に主要コンテンツ判定のための特徴量を抽出する特徴量抽出ステップと、
前記セグメント毎の特徴量に基づいて機械学習アルゴリズムを用いて主要コンテンツか否かを判定する主要コンテンツ判定ステップと、
主要コンテンツと判断された部位を結合して前記主要コンテンツとして出力する主要コンテンツ出力ステップと、を含み、
前記特徴量抽出ステップでは、
前記Ｗｅｂ文書で表示される文字列の特徴量、タグ情報の特徴量、アンカーリンク情報の特徴量を抽出し、各特徴量間の比率を用いて最終的な特徴量とするステップを含み、
前記出現頻度算出ステップでは、
検索エンジンが収集したＷｅｂ文書集合における出現確率を考慮した重みであるＷｅｂIDFを用いて前記重要キーワード候補に重みを付けるステップ、
または、
前記重要キーワード候補の前記Ｗｅｂ文書内の出現頻度を求め、該出現頻度と文書集合内での出現分布に基づいて各重要キーワード候補に重み付けを行うBM25を用いて該重要キーワード候補に重みを付けるステップ、
のいずれかを行うことを特徴とする重要キーワード抽出方法。
【請求項６】
前記出現頻度算出ステップは、
形態素解析手段の形態素解析結果から得られた形態素の固有名詞を含む前記重要キーワード候補の出現頻度を用いて、該重要キーワード候補に重みを付ける
請求項５記載の重要キーワード抽出方法。
【請求項７】
前記キーワード重要度算出ステップは、
前記辞典文書集合内のリンク構造を用いて、前記重要キーワード候補の固有重要度を求める
請求項５または６記載の重要キーワード抽出方法。
【請求項８】
前記キーワード重要度算出ステップは、
検索エンジンに投入された回数が多いキーワードほど、重要なキーワードとなるような値を用いて、前記重要キーワード候補の固有重要度を求める
請求項５乃至７のいずれか１項記載の重要キーワード抽出方法。
【請求項９】
請求項１乃至４のいずれか１項記載の重要キーワード抽出装置を構成する各手段としてコンピュータを機能させるための重要キーワード抽出プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、重要キーワード抽出装置及び方法及びプログラムに係り、特に、文書中に含まれるテキストから重要なキーワードを抽出する重要キーワード抽出装置及び方法及びプログラムに関する。

【背景技術】
【０００２】
従来の重要キーワード抽出法は、キーワードの出現頻度や複合語を構成する形態素の連接頻度等の"頻度"を用いて抽出する手法が主流であった（例えば、非特許文献１参照）。
【先行技術文献】
【非特許文献】
【０００３】
【非特許文献１】中川裕志、森辰則、湯本紘彰、「出現頻度と連接頻度に基づく専門用語抽出。自然言語処理」Vol.10 No.1, pp.27-45, 2003年1月
【発明の概要】
【発明が解決しようとする課題】
【０００４】
従来の重要キーワード抽出法は、キーワードの出現頻度や連接頻度等の"頻度"を用いる手法が主流であったため、事前に比較的量の多い文書集合を要する必要があった。また、頻度が高いからといって、重要であるという判定を行う手法は場合によっては上手く作用しない場合が多く、精度的にもあまりよい結果を得られていない。
【０００５】
本発明は、上記の点に鑑みなされたもので、キーワードそのものの固有の重要性を算出でき、文書内の出現頻度等に左右されずに重要キーワードを頑健に抽出することが可能な重要キーワード抽出装置及び方法及びプログラムを提供することを目的とする。
【課題を解決するための手段】
【０００６】
図１は、本発明の原理構成図である。
【０００７】
本発明（請求項１）は、Ｗｅｂ文書中に含まれるテキストから重要なキーワードを抽出する重要キーワード抽出装置であって、
Ｗｅｂ文書を取得して該Ｗｅｂ文書の特徴量を抽出し、主要コンテンツを抽出する主要コンテンツ抽出手段１０と、
オンライン百科辞典（例えば、Wikipedia（登録商標））に代表される文書集合内において一意の見出し語を持つ文書で、かつ、文書集合内において引用関係もしくは参照関係を持つ文書集合（以下、辞典文書集合と記す）の見出し語を形態素解析用の辞書として登録しているキーワード辞書１０１を参照して、主要コンテンツから重要キーワード候補を抽出する重要キーワード候補抽出手段２と、
Ｗｅｂ文書内の重要キーワード候補に重みを付ける出現頻度算出手段３と、
知名度や話題性が高く、内容的に興味深く、キーワードについて語る際に様々な話題が挙がる情報量が豊富なキーワードの重要度が高くなるように重要キーワード候補の固有重要度を算出し、第１の記憶手段１０５に格納するキーワード重要度算出手段４と、
重要キーワード候補のＷｅｂ文書中の位置に基づいて出現位置キーワード重要度を求め、第２の記憶手段１０６に格納する位置情報算出手段５と、
出現頻度算出手段３による重要キーワード候補の重み、第１の記憶手段１０５に格納されている重要キーワード候補の固有重要度及び第２の記憶手段１０６に格納されている出現位置キーワード重要度を乗算した値に基づいて、最終重要度付きのキーワード集合を出力するキーワード出力手段６と、を有し、
主要コンテンツ抽出手段１０は、
Ｗｅｂ文書を所定の分割規則に基づいてセグメントに分割するＷｅｂ文書分割手段１２と、
セグメント毎に主要コンテンツ判定のための特徴量を抽出する特徴量抽出手段１３と、
セグメント毎の特徴量に基づいて機械学習アルゴリズムを用いて主要コンテンツか否かの判定を行う主要コンテンツ判定手段１４と、
主要コンテンツと判断された部位を結合して主要コンテンツとして出力する主要コンテ

ンツ出力手段１５を含み、
主要コンテンツ抽出手段１０の特徴量抽出手段１３は、
Ｗｅｂ文書で表示される文字列の特徴量、タグ情報の特徴量、アンカーリンク情報の特徴量を抽出し、各特徴量間の比率を用いて最終的な特徴量とする手段を含み、
出現頻度算出手段３は、
検索エンジンが収集したＷｅｂ文書集合における出現確率を考慮した重みであるＷｅｂIDFを用いて重要キーワード候補に重みを付ける手段、
または、
重要キーワード候補のＷｅｂ文書内の出現頻度を求め、該出現頻度と文書集合内での出現分布に基づいて各重要キーワード候補に重み付けを行うBM25を用いて該重要キーワード候補に重みを付ける手段、のいずれかを有する。
【０００８】
また、本発明（請求項２）は、上記請求項１の出現頻度算出手段３が、
形態素解析手段の形態素解析結果から得られた形態素の固有名詞を含む重要キーワード候補の出現頻度を用いて、該重要キーワード候補に重みを付ける手段を含む。
【０００９】
また、本発明（請求項３）は、上記請求項１または２のキーワード重要度算出手段４が、
辞典文書集合内のリンク構造を用いて、重要キーワード候補の固有重要度を求める手段を含む。
【００１０】
また、本発明（請求項４）は、上記請求項1乃至３のキーワード重要度算出手段４が、
検索エンジンに投入された回数が多いキーワードほど、重要なキーワードとなるような値を用いて、重要キーワード候補の固有重要度を求める手段を含む。
【００１７】
図２は、本発明の原理を説明するための図である。
【００１８】
本発明（請求項５）は、Ｗｅｂ文書中に含まれるテキストから重要なキーワードを抽出する重要キーワード抽出方法であって、
主要コンテンツ抽出手段が、Ｗｅｂ文書を取得して該Ｗｅｂ文書の特徴量を抽出し、主要コンテンツを抽出する主要コンテンツ抽出ステップ（ステップ１）と、
重要キーワード候補抽出手段が、オンライン百科辞典（例えば、Wikipedia（登録商標））に代表される文書集合内において一意の見出し語を持つ文書で、かつ、文書集合内において引用関係もしくは参照関係を持つ文書集合（以下、辞典文書集合と記す）の見出し語を形態素解析用の辞書として登録しているキーワード辞書を参照して、主要コンテンツから重要キーワード候補を抽出する重要キーワード候補抽出ステップ（ステップ２）と、
出現頻度算出手段が、Ｗｅｂ文書内の重要キーワード候補に重みを付ける出現頻度算出ステップ（ステップ３）と、
キーワード重要度算出手段が、知名度や話題性が高く、キーワードについて語る際に様々な話題が挙がる情報量が豊富なキーワードの重要度が高くなるように重要キーワード候補の固有重要度を算出し、第１の記憶手段に格納するキーワード重要度算出ステップ（ステップ４）と、
位置情報算出手段が、重要キーワード候補のＷｅｂ文書中の位置に基づいて出現位置キーワード重要度を求め、第２の記憶手段に格納する位置情報算出ステップ（ステップ５）と、
キーワード出力手段が、出現頻度算出ステップによる重要キーワードの重み、第１の記憶手段に格納されている重要キーワード候補の固有重要度及び第２の記憶手段に格納されている出現位置キーワード重要度を乗算した値に基づいて、最終重要度付きのキーワード集合を出力するキーワード出力ステップ（ステップ６）と、を行い、
主要コンテンツ抽出ステップ（ステップ１）では、

Ｗｅｂ文書を所定の分割規則に基づいてセグメントに分割するＷｅｂ文書分割ステップと、
セグメント毎に主要コンテンツ判定のための特徴量を抽出する特徴量抽出ステップと、
セグメント毎の特徴量に基づいて機械学習アルゴリズムを用いて主要コンテンツか否かを判定する主要コンテンツ判定ステップと、
主要コンテンツと判断された部位を結合して主要コンテンツとして出力する主要コンテンツ出力ステップと、を含み、
特徴量抽出ステップでは、
Ｗｅｂ文書で表示される文字列の特徴量、タグ情報の特徴量、アンカーリンク情報の特徴量を抽出し、各特徴量間の比率を用いて最終的な特徴量とするステップを含み、
出現頻度算出ステップ（ステップ３）では、
検索エンジンが収集したＷｅｂ文書集合における出現確率を考慮した重みであるＷｅｂIDFを用いて重要キーワード候補に重みを付けるステップ、
または、
重要キーワード候補のＷｅｂ文書内の出現頻度を求め、該出現頻度と文書集合内での出現分布に基づいて各重要キーワード候補に重み付けを行うBM25を用いて該重要キーワード候補に重みを付けるステップ、
のいずれかを行う。

【００１９】

また、本発明（請求項６）は、請求項１１の出現頻度算出ステップ（ステップ３）において、
形態素解析手段の形態素解析結果から得られた形態素の固有名詞を含む重要キーワード候補の出現頻度を用いて、該重要キーワード候補に重みを付ける。

【００２０】

また、本発明（請求項７）は、請求項１１または１２のキーワード重要度算出ステップ（ステップ４）において、
辞典文書集合内のリンク構造を用いて、重要キーワード候補の固有重要度を求める。

【００２１】

また、本発明（請求項８）は、請求項１１乃至１３のキーワード重要度算出ステップ（ステップ４）において、
検索エンジンに投入された回数が多いキーワードほど、重要なキーワードとなるような値を用いて、重要キーワード候補の固有重要度を求める。

【００２３】

本発明（請求項９）は、請求項１乃至４のいずれか１項記載の重要キーワード抽出装置を構成する各手段としてコンピュータを機能させるための重要キーワード抽出プログラムである。

【発明の効果】

【００２４】

本発明によれば、文書中の重要キーワードの抽出が可能となり、それらを用いた文書管理システムや、重要キーワードからの広告配信等での応用が可能である。さらに、従来の技術では、文書中におけるキーワードの出現頻度のみから重要キーワードの特定を行っていたが、オンライン百科事典（例えば、Wikipedia（登録商標））のリンク構造や実際の検索クエリの投入回数を用いることで、従来の技術では実現できなかったキーワードそのものの固有の重要性を算出でき、文書内の出現頻度等に左右されない重要キーワードの頑健な抽出が可能になる。

【図面の簡単な説明】

【００２５】

【図１】本発明の原理構成図である。
【図２】本発明の原理を説明するための図である。
【図３】本発明の一実施の形態における重要キーワード抽出装置の構成図である。

【図４】本発明の処理の概要を示す図である。
【図５】本発明の一実施の形態における主要コンテンツ抽出部の構成図である。
【図６】本発明の一実施の形態におけるＷｅｂ文書取得・入力部の構成図である。
【図７】本発明の一実施の形態におけるＷｅｂ文書分割部の構成図である。
【図８】本発明の一実施の形態における特徴量抽出部の構成図である。
【図９】本発明の一実施の形態における主要コンテンツ判定部の構成図である。
【図１０】本発明の一実施の形態における主要コンテンツ出力部の構成図である。
【図１１】本発明の一実施の形態における特徴量のパラメータ推定方法のフローチャートである。
【図１２】本発明の一実施の形態における主要コンテンツ例（その１）である。
【図１３】本発明の一実施の形態における主要コンテンツ例（その２）である。
【発明を実施するための形態】
【００２６】
以下、図面と共に本発明の実施の形態を説明する。
【００２７】
本発明は、文書中における重要キーワードを抽出するシステムである。従来の技術においては比較的量の多い文書集合を事前に用意する必要があったが、本発明では文書が少ない場合においても精度が落ちにくい手法を提案する。
【００２８】
本発明では、「文書集合内において一意の見出し語を持つ文書で、かつ文書集合内において引用関係もしくは参照関係を持つ文書集合」を用いて、重要キーワードを抽出する。例をあげると、オンライン百科事典である「Wikipedia(登録商標)」や、「はてなブックマーク（登録商標）」、「マイペディア（登録商標）」等の文書集合がそれに該当する。キーワード候補は、これらの文書集合における文書の見出し語を用いる。百科事典のような一意の見出し語を有する文書集合の見出し語を用いることで事象を一意に示す複数形態素から構成される複合語の切り出しが可能になる。さらに、このような見出し語は一般的に重要だと思われるキーワードを網羅しているため、掲載されている見出し語をキーワード候補として用いることで、重要なキーワードの絞込みが可能になる。
【００２９】
本実施の形態において、対象とする文書集合は「文書集合内において一意の見出し語を持つ文書で、かつ文書集合内において引用関係もしくは参照関係を持つ文書集合」とするが、以下では、Wikipedia（登録商標）を例に挙げて説明する。
【００３０】
図３は、本発明の一実施の形態における重要キーワード抽出装置の構成図である。
【００３１】
同図に示す重要キーワード抽出装置は、主要コンテンツ抽出部１０を有する文書入力部１、重要キーワード候補抽出部２、出現頻度算出部３、キーワード重要度算出部４、位置情報算出部５、重要キーワード出力部６、キーワード辞書１０１、ＷｅｂＩＤＦ記憶部１０２、事前に求められているWikipedia（登録商標）のリンク構造からのキーワード重要度を格納するキーワード重要度１記憶部１０３、事前に求められている検索エンジンの検索クエリ投入回数を用いたキーワード重要度を格納するキーワード重要度２記憶部１０４、固有重要度記憶部１０５、位置重要度記憶部１０６から構成される。
【００３２】
図４は、本発明の処理の概要を示す図である。
【００３３】
ステップ１００）　文書入力部１の主要コンテンツ抽出部１０は、入力された対象文書がＷｅｂ文書であるかを判定し、Ｗｅｂ文書である場合はステップ２００に移行し、Ｗｅｂ文書でない場合はステップ３００に移行する。
【００３４】
ステップ２００）　Ｗｅｂ文書である場合は、主要コンテンツの抽出を行う（主要コン

テンツの抽出処理）。
【００３５】
ステップ３００）　重要キーワード候補抽出部２は、文書内のテキストからキーワード候補を抽出する（キーワード候補の抽出処理）。
【００３６】
ステップ４００）　出現頻度算出部３において文書中に出現するキーワード候補の出現頻度を求め、キーワード重要度算出部４ではキーワード候補の重要度（固有重要度）を求め、位置情報算出部５は、キーワード候補が文書中に出現する位置により重要度（位置重要度）を求め、重要キーワード出力部６において、固有重要度と位置重要度に基づいて、キーワードの順序付けを行う（キーワードの重要度算出処理）。
【００３７】
以上が、文書中における重要キーワードを抽出する処理の主な流れである。
【００３８】
［１］主要コンテンツの抽出処理（ステップ２００）：
当該処理は、文書入力部１の主要コンテンツ抽出部１０が行う処理である。
【００３９】
入力文書がＷｅｂ文書である場合、ナビゲーションリンクや広告テキスト等のＷｅｂ文書の内容とは関係のないテキストが存在する。そのため、それら不要テキストを除去する主要コンテンツ部分の抽出を行う必要がある。Ｗｅｂ文書か否かの判定は、ファイルの拡張子を用いて行うものとする。
【００４０】
［１－１］処理の流れ：
文書入力部１の主要コンテンツ抽出部１０は、図５に示すように、Ｗｅｂ文書取得・入力部１１、Ｗｅｂ文書分割部１２、特徴量抽出部１３、主要コンテンツ判定部１４、主要コンテンツ出力部１５から構成される。ここで、特徴量抽出部１３と主要コンテンツ判定部１４は、Ｗｅｂ文書分割部１２で分割されたＷｅｂ文書毎に処理を行う。
【００４１】
Ｗｅｂ文書取得・入力部１１は、処理するＷｅｂ文書（データ）の入力を行うもので、図６に示すように、データ入力部１１１、Ｗｅｂ文書ファイル入力部１１２、ＵＲＬ入力部１１３、Ｗｅｂ文書取得部１１４、文字コード変換部１１５から構成される。
【００４２】
Ｗｅｂ文書分割部１２は、取得したＷｅｂ文書を分割するものであり、図７に示すように、広告対象領域部１２１、ノイズとなるタグや領域除去部１２２、Ｗｅｂ文書の分割部１２３を有する。広告対象領域部１２１では、Ｗｅｂ文書に広告対象領域が存在する場合は、当該広告対象領域を抽出する。ノイズとなるタグや領域除去部１２２は、広告対象領域または入力されたＷｅｂ文書のノイズとなるタグや領域を除去し、Ｗｅｂ文書分割部１２３において文書を分割する。
【００４３】
特徴量抽出部１３は、Ｗｅｂ文書分割部１２で分割されたＷｅｂ文書毎に、主要コンテンツ判定のための特徴量を抽出するものであり、図８に示すように、アンカーリンク情報特徴量抽出部１３１、タグ情報特徴量抽出部１３２、Ｗｅｂ文書で表示される文字列特徴量抽出部１３３、特徴量の正規化部１３４、特徴量の比率特徴量抽出部１３５から構成される。
【００４４】
主要コンテンツ判定部１４は、特徴量抽出部１３で抽出された特徴量から主要コンテンツか否かの判定を行うもので、図９に示すように、特徴量入力部１４１、テキスト判定部１４２、主要コンテンツ判定部１４３から構成される。
【００４５】
主要コンテンツ出力部１５は、図１０に示すように、タグ付きテキストを要求された場合に用いられるタグ付テキスト出力部１５１、タグなしテキストを要求された場合に用い

られるタグなしテキスト出力部１５２、タグ付テキスト出力部１５１またはタグなしテキスト出力部１５２から出力された主要コンテンツと判定されたセグメントを結合して出力するデータ出力部１５３から構成される。

【００４６】

主要コンテンツ抽出部１０の処理は、概略以下のようになる。

【００４７】

（１）Ｗｅｂ文書取得・入力部１１が、処理するＷｅｂ文書（データ）の入力を行う。

【００４８】

（２）Ｗｅｂ文書分割部１２が、（１）で取得したＷｅｂ文書を分割する。

【００４９】

以下の処理は、分割したＷｅｂ文書毎に行う。

【００５０】

（３）特徴量抽出部１３が主要コンテンツ判定のための特徴量を抽出する。

【００５１】

（４）主要コンテンツ判定部１４が、（３）で抽出した特徴量から主要コンテンツか否かの判定を行う。

【００５２】

（５）主要コンテンツ出力部１５が、（４）で主要コンテンツと判定された部位を結合して最終出力とする。

【００５３】

以下に主要コンテンツ抽出部１０の動作を詳細に説明する。

【００５４】

［１－２］Ｗｅｂ文書の入力：

Ｗｅｂ文書の入力は、Ｗｅｂ文書取得・入力部１１で行われる処理である。

【００５５】

Ｗｅｂ文書取得・入力部１１のデータ入力部１１１は、ユーザから入力された主要コンテンツを抽出したＷｅｂ文書のＵＲＬ、もしくは、ファイルそのものを取得する。入力がＵＲＬの場合はＵＲＬ入力部１１３及びＷｅｂ文書取得部１１４において、そのＵＲＬ先のＷｅｂ文書を取得し、ファイルが直接入力された場合はＷｅｂ文書ファイル入力部１１２がそのファイルを取得する。文字コード変換部１１５は、Ｗｅｂ文書ファイル入力部１１２及びＷｅｂ文書取得部１１４から取得したＷｅｂ文書の文字コードをＵＴＦ－８に変換し統一する。

【００５６】

［１－３］Ｗｅｂ文書の分割：

Ｗｅｂ文書の分割は、Ｗｅｂ文書分割部１２で行われる処理である。

【００５７】

最初に、Ｗｅｂ文書分割部１２の広告対象領域抽出部１２１は、インターネット広告等のコンテンツタグを含む領域がある場合、その領域を抽出する。ここで、インターネット広告とは、google（登録商標）や、Overture（登録商標）等の広告会社が広告配信のための主要コンテンツ絞込みに用いるタグである。google（登録商標）の広告の場合、＜!--google_ad_section_start--＞から、＜--google_ad_section_end--＞までがその領域に該当する。これらのタグはＷｅｂ文書によって文字列が少々異なったり大文字で表記されるので、大文字と小文字を区別しない正規表現を用いたり、ワイルドカードの正規表現を用いる等により、多少の文字列表記の違いを吸収する処理を行う。以下、正規表現を用いる処理の説明の際には、多少の違いを吸収する処理を行っているものとする。

【００５８】

以降の処理は、インターネット広告の領域が存在する場合、上記で述べた領域を抽出する処理を行い、インターネット広告の領域がない場合は、最初に入力されたＷｅｂ文書に対して処理を行う。

【００５９】

次に、ノイズとなるタグや領域除去部１２２は、余計なタグや領域、特定の文字列を除去する処理を行う。除去されるタグや領域は、Ｗｅｂ文書のＨＴＭＬを説明するコメントタグであったり、JavaScriptであったり、formタグであったりする。除去するタグと領域を以下に記載する。

【００６０】

・"＜!"で始まり、"--＞"で終わるコメントタグ；
・"＜script＞"タグから、"＜/script＞"タグで囲まれる領域；
・"＜style＞"タグから、"＜/style＞"タグで囲まれる領域；
・"＜select＞"タグから、"＜/select＞"タグで囲まれる領域；
・"＜noscript＞"タグから、"＜/ noscript ＞"タグで囲まれる領域；
・"＜form＞"タグから、"＜/form＞"タグで囲まれる領域；
・連続した空白文字列（単一の空白は除く）；
・連続したタブ文字列（単一のタブは除く）；

ノイズとなるタグや領域除去部１２２は、以上のタグ、領域、文字列を正規表現を用いて除去する。タグ内にalt属性やclass属性が存在する場合も考えられるため、その場合はそれらを含めたタグを考慮した正規表現を用いて分割を行う（例：＜style class="hoge"＞）。

【００６１】

次に、Ｗｅｂ文書の分割部１２３は、Ｗｅｂ文書の分割を行う。分割の規則は、以下のタグを用いて分割を行う。

【００６２】

・　＜div＞
・　＜/div＞
・　＜td＞
・　＜/td＞

タグ内にalt属性やclass属性が存在する場合も考えられるため、その場合は、それらを含めたタグを考慮した正規表現を用いて分割を行う（例：＜div class="hoge".）。

【００６３】

Ｗｅｂ文書分割部１２で分割されたＷｅｂ文書の一つ一つを『セグメント』と呼び、これらはメモリ（図示せず）に格納される。以降、特徴量抽出部１３と主要コンテンツ判定部１４の処理は、当該セグメント毎に行う。

【００６４】

［１－４］特徴量の抽出：

ここでは、特徴量抽出部１３が抽出する特徴量について述べる。特徴量抽出部１３では、メモリ（図示せず）に格納されたセグメント毎に、以降で述べる特徴量を抽出し、Ｗｅｂ文書の主要コンテンツ部分の判定を行う。

【００６５】

［１－４－１］Ｗｅｂ文書で表示される文字列

当該特徴量は、Ｗｅｂ文書で表示される文字列特徴量抽出部１３３で抽出される特徴量である。ここで述べる文字列とは、ＨＴＭＬタグ等のＷｅｂブラウザで表示されない文字列を含まないものとする。

【００６６】

Ｗｅｂ文書で表示される文字列として、文字列の量、句読点の数がある。

【００６７】

［１－４－１－１］文字列の量：

一般的にＷｅｂ文書の主要コンテンツ部分は、主要コンテンツでない部分と比較すると多くの文字列が含まれている。また、全体的に文字列の少ないＷｅｂ文書においても同様のことがいえる。そのためＷｅｂ文書で表示される文字列特徴量抽出部１３３では、分割されたＷｅｂ文書に含まれる文字列の数を特徴量とする。そして、特徴量の正規化部１３４では、文字列の量を正規化して特徴とする手法と、文字列の絶対値を用いて特徴量とす

る手法の２つを実行し、最終的な文字列の量の特徴量とする。

【００６８】

・文字列の量の正規化を行い特徴量とする手法：

特徴量の正規化部１３４では、全てのセグメントにおいて最大の文字列の量をもつセグメントの特徴量を１とする正規化を行う。例えば、全てのセグメントにおいて最大の文字列の量が２００で、あるセグメント内の文字列の量が１００だった場合には、そのセグメントの文字列の量の特徴量は０．５となる。このような正規化を行うことで、全体的に文字列の少ないＷｅｂ文書においても主要コンテンツの抽出が可能になる。

【００６９】

・文字列の量の絶対値を用いて特徴量とする手法：

上記で述べた正規化を行い特徴量とする手法は、全体的に文字列の少ないＷｅｂ文書において有効であったが、正規化を行うことで、全体的に文字列の量が多く、主要コンテンツ部分のセグメント間の文字列の量の差が大きい場合に不都合が生じる。例えば、全てのセグメントにおいて最大の文字列の量が１０００で、あるセグメント内の文字列の量が１００だった場合、そのセグメントの文字列の量の特徴量は０．１になる。文字列の量としては多いはずだが、正規化を行うことで、このような弊害が生じる。そこで、文字列の絶対値を用いて特徴量とする手法を行う必要がある。具体的には、特徴量の正規化部１３４では、ある特定の値を超えた場合にその文字列の特徴量を１とする手法を用いる。例えば、あるセグメント内の文字列の量が１００の場合、文字列の量が５以上の場合の特徴量が１となり、文字列の量が１０以上の場合の特徴量が１となり，…，文字列の量が１０５以上の特徴量は０となり，…，文字列の量が２００以上の特徴量は０となるように特徴量を作成する。このように、ある特定の文字列量を超えた場合に特徴量を１とする手法を用いることで、特徴量の最大値は１のままで文字列量の絶対値を特徴量とすることができる。また、例における文字列の量の絶対値の特徴量の間隔は５としたが、場合において適切な間隔を用いるのが好ましい。８，１６，３２，６４といった２の乗数を用いて特徴量の間隔とする手法も考えられる。文字列の量がｘ以上の…の最大のｘも同様に、場合において適切な値に変更する。主要コンテンツ判定における計算量を減らしたい場合にはｘの値を小さくすればよい。

【００７０】

また、セグメント内に文字列が全くない場合も考えられる。その場合は、以降で説明する特徴量を抽出するまでもなく、主要コンテンツ判定部１４内のテキスト判定部１４２で主要コンテンツでないと判断できる。そのため、実際の実行時には以下の特徴量抽出は行わず、該当セグメントを非主要コンテンツとして判別する。この処理は、特徴量の学習を行う際には行わない。

【００７１】

［１－４－１－２］句読点の数：

Ｗｅｂ広告等のノイズとなりやすいセグメントは、文字列の量は多いが句読点の数が少ない傾向にある。そのため、句読点の数を特徴とする。具体的には、セグメント内の文字列に含まれる、『、』、『，』、『。』、『．』、『！』、『・』、『？』、『…』の数を特徴量としてカウントする。この特徴量も、文字列の量で述べた正規化による特徴量と、絶対値による特徴量の二通りを算出する。算出方法においては、[１－４－１－１]の文字列の量で述べた手法と同じものを用いる。

【００７２】

［１－４－２］タグ情報

ここでは、タグ情報特徴量抽出部１３２で扱うＨＴＭＬタグ等のタグ情報に関する特徴量について述べる。タグ情報には、

・テキスト系のHTMLタグの数；

・テキスト系のHTMLタグの連続出現数；

・リンクリストタグの数、Ｗｅｂ文書で表示される文字列を含まない文字列の量；

がある。

【００７３】
［１－４－２－１］テキスト系のＨＴＭＬタグの数：
あるセグメント内において、Ｗｅｂ文書で表示される文字列が多い場合、テキストに関するＨＴＭＬタグが多く含まれる。また、ブログ等のＣＧＭ(Consumer Generated Media)においては、Ｗｅｂ文書で表示される文字列は少ないが、ユーザが改行タグを多用する事例が多くみられる。そこで、タグ情報特徴量抽出部１３２は、テキストに関するＨＴＭＬタグの数を特徴量として用いる。この特徴量も、Ｗｅｂ文書で表示される文字列で用いた特徴量の正規化を同様に行い、最終的な特徴量とし、メモリ（図示せず）に格納する。例えば、全てのセグメントにおいて最大のＨＴＭＬタグの量が１０で、あるセグメント内のＨＴＭＬタグの量が５だった場合には、そのセグメントのＨＴＭＬタグの量の特徴量は０．５となる。そして、実際に使用するテキスト系のＨＴＭＬタグは、以下のタグを対象とする。
【００７４】
・ ＜p＞
・ ＜/p＞
・ ＜br＞
・ ＜/br＞
・ ＜font＞
・ ＜/font＞
タグ内にsize属性やclass属性が存在する場合も考えられるため、その場合は、それらを含めたタグを考慮した正規表現を用いてカウントを行う（例：＜font size="+1"＞）。
【００７５】
［１－４－２－２］テキスト系のＨＴＭＬタグの連続出現数：
Ｗｅｂ文書で表示される文字列が集中して記述してあるセグメントは、テキスト系のタグが多く存在すると同時に、テキスト系のHTMLタグが連続して出現する。ここでいう、連続して出現するというのは、他のアンカーリンク等のＨＴＭＬタグが間に出現しないということである。そこで、［１－４－２－１］で述べたテキスト系のＨＴＭＬタグの連続出現数を特徴量とする。この特徴量も、Ｗｅｂ文書で表示される文字列で用いた特徴量の正規化を同様に行い、最終的な特徴量とし、メモリ（図示せず）に格納する。例えば、全てのセグメントにおいて最大のＨＴＭＬタグの連続量が１０で、あるセグメント内のＨＴＭＬタグの連続量が５だった場合には、そのセグメントの文字列の量の特徴量は０．５となる。
【００７６】
［１－４－２－３］リンクリストタグの数：
あるセグメント内においてリンクリストタグが多い場合、そのセグメントにはナビゲーションリンク等の多くのアンカーリンクが存在し、そのセグメントは主要コンテンツとならない可能性が高い。そこで、リンクリストタグの数を特徴量とする。この特徴量も、Ｗｅｂ文書で表示される文字列で用いた特徴量の正規化を同様に行い、最終的な特徴量とし、メモリ（図示せず）に格納する。例えば、全てのセグメントにおいて最大のリンクリストタグの量が１０で、あるセグメント内のリンクリストタグの量が５だった場合には、そのセグメントのリンクリストタグの量の特徴量は、０．５となる。そして、具体的に使用するリンクリストタグは、以下のタグを対象とする。
【００７７】
・ ＜li＞
・ ＜ul＞
・ ＜dl＞
・ ＜dd＞
・ ＜ol＞
タグ内にalt属性やclass属性が存在する場合も考えられるため、その場合は、それらを含めたタグを考慮した正規表現を用いてカウントを行う（例：＜font class="hoge"）。

【００７８】
［１－４－２－４］Ｗｅｂ文書で表示される文字列を含まない文字列（ＨＴＭＬタグを含む）の量：
あるセグメント内において、Ｗｅｂで表示されない文字列（ＨＴＭＬタグを含む）が多い場合、そのセグメントは広告等の主要コンテンツでない可能性が高い。そこで、Ｗｅｂ文書で表示される文字列以外の文字列（ＨＴＭＬタグを含む）量を特徴量とする。この特徴量も、Ｗｅｂ文書で表示される文字列で用いた特徴量の正規化を同様に行い、最終的な特徴量とし、メモリ（図示せず）に格納する。例えば、全てのセグメントにおいて最大のＷｅｂで表示されない文字列の量が１００で、あるセグメント内のＷｅｂで表示されない文字列の量が５０だった場合には、そのセグメントのＷｅｂで表示されない文字列の量の特徴量は０．５となる。
【００７９】
［１－４－３］アンカーリンク情報：
以下では、特徴量抽出部１３のアンカーリンク情報特徴量抽出部１３１で扱うアンカーリンクに関する特徴量の抽出方法について述べる。アンカーリンク情報には、
・アンカーリンクの数；
・各アンカーリンクの文字列の平均量；
・すべてのアンカーリンク文字列の合計値；
・最大文字列のアンカーリンクＵＲＬの量；
・広告に関するアンカーリンクを含むか；
がある。
【００８０】
［１－４－３－１］アンカーリンクの数：
あるセグメントにおいて、アンカーリンクの数が多数含まれているセグメントは主要コンテンツでない可能性が高い。そこで、アンカーリンクの数を特徴量として用いる。具体的には、＜a href=…＞…＜/a＞タグで表されるアンカーリンクの数を特徴量とする。この特徴量も、特徴量の正規化部１３４において、Ｗｅｂ文書で表示される文字列で用いた特徴量の正規化を同様に行い、最終的な特徴量とし、メモリ（図示せず）に格納する。例えば、全てのセグメントにおいて最大のアンカーリンクの数が１０で、あるセグメント内のアンカーリンクの数が５だった場合には、そのセグメントのアンカーリンクの数の特徴量は０．５となる。アンカーリンクタグには、class属性やalt属性が含まれる場合もあるので、アンカーリンクタグの数は正規表現を用いてカウントする。
【００８１】
［１－４－３－２］各アンカーリンクの文字列の平均量：
各アンカーリンクの文字列が平均して多い場合、そのセグメントは、関連記事等のナビゲーションリンクである可能性が高い。また、アンカーリンクの文字列が平均して少ない場合、主要コンテンツ内に含まれるキーワード検索リンクである可能性が高い。そこで、セグメントに含まれるアンカーリンクの文字列の平均量を特徴量として用いる。アンカーリンクの文字列とは、＜a href='…'＞○○○＜/a＞の○○○の部分に該当する。この特徴量も特徴量の正規化部１３４においてＷｅｂ文書で表示される文字列で用いた特徴量の正規化を同様に行い、最終的な特徴量とし、メモリ（図示せず）に格納する。
【００８２】
［１－４－３－３］すべてのアンカーリンク文字列の合計値：
セグメント内に含まれるアンカーリンクの文字列の合計量が多い場合、そのセグメントはナビゲーションリンクである可能性が高い。そこで、セグメント内に含まれるアンカーリンクの文字列の合計量を特徴量として用いる。アンカーリンクの文字列とは、＜a href='…'○○○＜/a＞の○○○の部分に該当する。この特徴量も特徴量の正規化部１３４でＷｅｂ文書で表示される文字列で用いた特徴量の正規化を同様に行い、最終的な特徴量とし、メモリ（図示せず）に格納する。
【００８３】

［１－４－３－４］最大文字列のアンカーリンクＵＲＬの量：
アンカーリンク先のＵＲＬ文字列が非常に長い場合、そのセグメントは広告である可能性が高い。そこで、セグメント内で最大長のアンカーリンク先のＵＲＬの文字列を特徴量として用いる。ここで述べるアンカーリンク先のＵＲＬ文字列とは、＜a href='△△△'…＞…＜/a＞の△△△の部分に該当する。この特徴量も特徴量の正規化部１３４でＷｅｂ文書で表示される文字列で用いた特徴量の正規化を同様に行い、最終的な特徴量とし、メモリ（図示せず）に格納する。

【００８４】
［１－４－３－５］広告に関するアンカーリンクを含むか：
広告に関するＵＲＬを含むアンカーリンクは特徴的な文字列を含む可能性が高い。例えば、「adclick」、「adnet」、「banner」等がそれにあたる。そこで、アンカーリンク情報特徴量抽出部１３１では、このような広告となりやすい文字列を含んだＵＲＬを含むアンカーリンクが存在する場合、特徴量を１とし、存在しない場合を０とする特徴量を抽出し、メモリ（図示せず）に格納する。広告になりやすい文字列は、広告除去用のFirefox用アドインであるadblock plugin等のサイトに記載されているため、それを用いる。

【００８５】
［１－４－４］特徴量の比率：
ここでは、特徴量の比率特徴量抽出部１３５で扱う、［１－４－１］～［１－４－３］で述べた特徴間の比率を用いた特徴量について述べる。

【００８６】
［１－４－４－１］テキスト系のタグ数とテキスト系のタグの連続出現数の比率：
テキスト系のタグが多数あり、またテキスト系のタグの連続出現数が多いセグメントは、テキストが密に書かれているため、主要コンテンツである可能性が高い。しかしながら、テキスト系のタグは多数あるが、テキスト系のタグの連続出現数が少ないセグメントは、セグメントのサイズが大きいだけで、主要コンテンツでない可能性が高いといえる。そこで、特徴量の比率特徴量抽出部１３５では、テキスト系のタグ数とテキスト系のタグの連続出現数の比率を特徴量として用いる。具体的には、テキスト系のタグ数を分母とし、テキスト系のタグの連続出現数を分子とした値を特徴量として用いる。ここで、テキスト系のタグ数が０の場合は、分母が０となってしまうため、この場合のテキスト系のタグ数とテキスト系のタグの連続出現数の比率の特徴量は０とする。本特徴量も特徴量の正規化部１３４においてＷｅｂ文書で表示される文字列で用いた特徴量の正規化を同様に行い、最終的な特徴量とし、メモリ（図示せず）に格納する。この特徴量が大きければ大きいほど主要コンテンツである可能性が高い。

【００８７】
［１－４－４－２］Ｗｅｂで表示される文字列とタグの比率：
あるセグメント内において、Ｗｅｂで表示される文字列が多い場合は、主要コンテンツとなる可能性が高いが、同じセグメント内において、ＨＴＭＬタグ等のタグが多い場合もある。この場合、前述の［１－４－４－１］で述べたようにセグメントのサイズが大きいだけで主要コンテンツでない可能性がある。そこで、特徴量の比率特徴量抽出部１３５では、Ｗｅｂで表示される文字列とタグの比率を特徴量として用いることでこの場合に対処する。具体的には、Ｗｅｂで表示される文字列の数を分子とし、タグの数を分母とした値を特徴量とする。この特徴量が大きければ大きいほど、主要コンンテンツである可能性が高い。本特徴量も、特徴量の正規化部１３４において、Ｗｅｂ文書で表示される文字列で用いた特徴量の正規化を同様に行い、最終的な特徴量とし、メモリ（図示せず）に格納する。タグの数が０の場合は分母が０となってしまうため、特徴量は１とする。

【００８８】
［１－４－４－３］アンカーリンクの数とリンクリストタグの数の比率：
あるセグメント内において、アンカーリンクの数や、リンクリストタグの数が多ければ多いほど、そのセグメントは主要コンテンツでない可能性が高いが、セグメントが広いためにこれらの特徴量が偶然高くなってしまう場合も考えられる。そこで、特徴量の比率特

徴量抽出部１３５では、アンカーリンクの数とリンクリストタグの数の比率を特徴量として用いる。具体的にはアンカーリンクの数を分母とし、リンクリストタグの数を分子とし、特徴量とする。この特徴量が大きければ大きいほど、セグメントの面積に対し密度の高いリンク数が存在することになり、主要コンテンツでない可能性が高い。本特徴量も、特徴量の正規化部１３４で、Ｗｅｂ文書で表示される文字列で用いた特徴量の正規化を同様に行い、最終的な特徴量とし、メモリ（図示せず）に格納する。アンカーリンクの数が０の場合は分母が０となってしまうため、特徴量は０とする。

【００８９】

［１－５］主要コンテンツの判定：

ここで、主要コンテンツの判定部１４が、特徴量抽出部１３で求められ、メモリ（図示せず）に格納されている［１－４］で述べた特徴量を用いて、主要コンテンツか否かの判定を用いて行う手法について述べる。判定には、Support Vector Machine(SVM)や最大エントロピー法、ナイーブベイズ法等の機械学習アルゴリズムを用いて判定を行う。

【００９０】

図１１は、本発明の一実施の形態における特徴量のパラメータ推定方法のフローチャートである。最初に、人手で主要コンテンツか否かを特徴量を抽出したセグメント毎に判定し、訓練データを作成する（ステップ３０１～３０３）。ここで、［１－４－１－１］で、Ｗｅｂで表示される文字列が存在しない場合は主要コンテンツと見做さないと記述したが、機械学習を用いた手法において負例として学習に有効であるため、訓練データにはそのようなデータも採用する。そして、そのセグメントの特徴量を用いて学習を行い（ステップ３０４）、特徴量毎の重みを算出する（ステップ３０５）。速度を重視する場合は、最大エントロピー法で学習し、精度を重視する場合には、二次の多項式カーネルを用いたSupport Vector Machineを用いて学習を行う。そしてこれらの学習したパラメータを用いて、特徴量入力部１４１がセグメントの特徴量を主要コンテンツ判定部１４３に入力し、主要コンテンツ判定部１４３がセグメント毎に主要コンテンツか否かを判定する。

【００９１】

［１－６］主要コンテンツ部分の出力：

上記の［１－５］で説明した主要コンテンツ判定部１４が主要コンテンツか否かの判定を行った後、主要コンテンツ出力部１５のデータ出力部１５３は、学習器によって主要コンテンツと判断されたセグメントのみを、結合して最終出力とする。出力の例を図１２、図１３に示す。

【００９２】

ここで、情報検索の事前処理として本装置を用いたい場合には、タグ付きテキスト出力部１５１を用いてＨＴＭＬタグ等のタグを残して出力する。また、情報推薦等でＷｅｂ文書の内容を解析したい場合には、タグなしテキスト出力部１５２を用いてＨＴＭＬタグ等のタグを削除して出力する。

【００９３】

［１－７］精度向上のための処理：

［１－４］で述べた特徴量を抽出する事前処理として、不要文字列等を除去する手法が有効である。以下に記述する不要文字列を事前に除去しておくことで、主要コンテンツの判定精度を高める。

【００９４】

・ ；
・&lt；
・&gt；
・&amp；
・&laquo；
・&raquo；

これらの文字列は、ＨＴＭＬタグ等で用いる記号をＷｅｂブラウザ上で表示する際に用いる特殊文字である。また、上記で挙げた特殊文字以外のＨＴＭＬ特殊文字も削除の対象

とする。特殊文字は実際表示される文字列に対して、文字列の量が少ないため、学習の際のノイズとなりやすい。

【００９５】

［１－８］実装理由による特徴量の選択：

主要コンテンツ抽出手法をユーザＰＣ等に組み込む場合、［１－４］で述べた全ての特徴量を用いて処理することは処理量的に難しい。そのため、抽出する特徴量を絞り込むことで処理量を削減する。ここで、機械学習による学習モデルは、絞り込んだ特徴量モデル毎に学習モデルを作成する。

【００９６】

［２］キーワード候補の抽出処理（ステップ３００）：

以下では、重要キーワード候補抽出部２について説明する。

【００９７】

従来の一般的なキーワード抽出手法は形態素解析や固有表現抽出を用いた手法であったが、形態素解析手法では、複合名詞の抽出に関する問題や重要キーワードの絞込みに関しての問題があり、また、固有表現抽出においては、人名、組織名、地名といった狭い範囲のキーワード抽出しかできないため、ユーザの興味を網羅するキーワード候補の抽出ができなかった。そこで、本発明では、ユーザ参加型オンライン百科事典であるWikipedia（登録商標）の見出し語をキーワード候補として用いることで、これらの解決を試みた。Wikipedia（登録商標）の見出し語のため、体系的にまとめられており、かつ実世界の事象を一意に表す特徴を持つため、ユーザの興味対象を幅広く網羅した言語資源であるといえる。また、Wikipedia（登録商標）に記載されているキーワードは、重要で、キーワードについて語る際に様々な話題が挙がる情報量が豊富なキーワードのみが登録されているため、重要なキーワードの絞込みが可能となる。具体的なキーワード抽出手法は、Wikipedia（登録商標）の見出し語を形態素解析用の辞書としてキーワード辞書１０１に登録し、そのキーワード辞書１０１を用いた形態素解析結果からキーワードを抽出する。なお、キーワード辞書１０１に登録する際、最も長いキーワードを抽出するよう辞書の重み付けを行う。

【００９８】

［３］キーワードの重要度算出処理（ステップ４００）：

前述の［２］では、Wikipedia（登録商標）の見出し語を用いて重要だと思われるキーワード候補の絞込みを実現したが、［３］では、出現頻度算出部３、キーワード重要度算出部４により、その中からさらに重要なキーワードを上位に位置づけるために、キーワードの重要度を算出する手法について述べる。

【００９９】

［３－１］ＷｅｂＩＤＦ：

検索エンジンが収集した大規模なＷｅｂ文書集合における出現確率を考慮した重みがWebIDFであり、WebIDF記憶部１０２に格納されているものとする。具体的には、出現頻度算出部３は、検索エンジンにキーワードを入力し、その結果得られたＷｅｂ文書のヒット数からＩＤＦ値を算出し、これをWebIDFとする。Ｗｅｂ文書集合中で多数出現するキーワードは、一般的なキーワードで特徴的なキーワードではないと判断する重みである。以下に、キーワードｋのWebIDF算出式を示す。

【０１００】

【数１】

$$WebIDF(k) = \log_2\left(\frac{N}{n_k + 1} + 1\right) \qquad (1)$$

ここで、Ｎはキーワードの閾値で、ｎ$_k$はキーワードｋのヒット数である。ここでキーワードｋの検索エンジンのヒット数の閾値Ｎは、ｎ$_k$を降順にソートし、不必要なキーワードが少なくなってきた辺りのｎ$_k$の値を用いる。その理由として、ｎ$_k$が高ければ高いほど一般的なキーワードである可能性が高く、また、閾値Ｎを設定すると閾値Ｎ以上のｎ$_k$を持つキーワードｋのWebIDF値は常に負となり、常にスコアとして低い値となるためである。なお、Ｎの推奨値は20000000である。

【０１０１】

［３－２］ＢＭ２５：

ここでは、キーワード重要度算出部４による文書内の出現頻度と文書集合内での出現分布を用いたキーワード重み付けの手法について述べる。これらの重み付けには既存の技術であるＢＭ２５を用いる。

【０１０２】

【数２】

$$BM25(d,k)=\frac{(k_1+1)\cdot tf(d,k)}{k_1\left((1-b)+b\cdot\frac{dl}{avdl}\right)+tf(d,k)}\cdot\log\left(\frac{N-df(k)+0.5}{df(k)+0.5}\right) \qquad (2)$$

ここで、ｄは文書であり、ｋは文書に含まれるキーワードである。そして、tf(d,k)はｄ内のキーワードｋの出現頻度、dlは文書長、avdlは文書集合内における平均文書長、Nは総文書数である。ｋ$_1$とｂは自由パラメータで、ｋ$_1$はtf(d,k)に関するパラメータ、ｂはdlに関するパラメータである。

【０１０３】

また、文書集合を用意できなく、文書１枚からキーワードを抽出したい場合には、以下の算出式を用いる。

【０１０４】

【数３】

$$BM25(d,k)=\frac{(k_1+1)\cdot tf(d,k)}{k_1+tf(d,k)} \qquad (3)$$

なお、式に関するノーテーションは（２）式と同様である。

【０１０５】

［３－３］タイトルと本文のキーワード重要度の算出：

上記の［３－２］では、出現頻度算出部３によるBM25を用いた出現頻度による重み付け手法を述べた。本節では、キーワード重要度算出部４での文書のタイトル中での出現頻度と、本文中での出現頻度を用いたキーワード重要度の算出手法について述べる。

【０１０６】

具体的には、キーワード重要度算出部４は、文書のタイトル（もしくはファイル名）から算出したBM25の値と、本文から算出したBM25の値の線形和を用いてスコアを算出する。以下に算出式を示す。

【０１０７】

【数４】

$$BM25_{all}(d,k) = (1-\alpha)\cdot BM25_{title}(d,k) + \alpha \cdot BM25_{body}(d,k) \qquad (4)$$

ここで、$BM25_{all}$はタイトルと本文の出現頻度を用いたBM25スコアで、$BM25_{title}$はタイトルでの出現頻度を用いたBM25スコア、$BM25_{body}$は本文でのBM25スコアとなる。αはパラメータである。αの推奨値は0.7とする。

【０１０８】

［３－４］形態素の固有名詞を用いたキーワード重要度算出：

一般的なニュース記事内において、人名や組織名等の固有名詞は最初に正式名称で記述された後に、省略形で記述される場合が多い。例えば、「麻生次郎」と最初に書かれた後に「麻生」と書かれる場合がそれに該当する。省略形で記述された場合、記事の主題に関するキーワードでも他の重要でないキーワードより出現回数が少ない場合が生じる問題があった。そこで、形態素解析結果を用いて簡易的に省略形に対応する方法を提案する。一般的な形態素解析器は、品詞として、人名（姓、名）、地名、組織名の出力が可能である。そのため、形態素解析結果から得られた形態素の固有名詞を含むキーワード候補の出現頻度を、その形態素の出現頻度と置き換えることで、この問題に対処する。例えば、「麻生次郎」と言うキーワードが１回出現し、「麻生」という形態素の固有名詞が３回出てきた場合、「麻生次郎」の出現頻度を３回とする。

【０１０９】

逆に、地名等は省略されずに記述されるため、キーワードの出現回数が大きくなりすぎてしまう問題がある。そのため、形態素解析結果から得られた地名の名詞の出現回数を提案する。以下に、キーワード重要度算出部４で用いる、上記で述べた手法の算出式を示す。

【０１１０】

$$\begin{aligned} tf(d,k) &= \alpha_{person} \cdot match(d,k,m_k) \\ tf(d,k) &= \alpha_{location} \cdot match(d,k,m_k) \\ tf(d,k) &= \alpha_{organization} \cdot match(d,k,m_k) \end{aligned} \qquad (5)$$

ここで、tf(d,k)は文書ｄにおける出現頻度によるキーワードｋのスコア、α_{person}、$\alpha_{location}$、$\alpha_{organization}$はそれぞれ、人名、地名、組織名に関する係数である。そして、match(d,k,m_k)は文書ｄにおけるキーワードｋに含まれる形態素m_kの出現頻度である。上記の式（５）で算出したtf(d,k)は、［３－２］，［３－３］で述べた出現頻度を用いたキーワード重要度の算出式で用いる。

【０１１１】

［３－５］キーワードの重要性の算出（キーワード固有重要度）：

本節では、キーワード重要度算出部４でのキーワードの重要性の算出について述べる。キーワードの重要性とは、知名度や話題性の高い、キーワードについて語る際に様々な話題が挙がる情報量が豊富な（内容の深い）キーワード程重要であると定義し、このキーワードの重要度をキーワード固有重要度と呼ぶ。当該キーワード重要度算出部４で算出されたキーワード固有重要度は固有重要度記憶部１０５に格納する。

【０１１２】

［３－５－１］Wikipedia（登録商標）内のリンク構造を用いた手法：

以下では、キーワード重要度算出部４において、Wikipedia（登録商標）内のリンク構造を用いてキーワード固有重要度を算出する手法について説明する。

【０１１３】

HITSやPageRankといった一般的なＷｅｂ文書のランキング手法は、Ｗｅｂページのリンク構造を用いて、Ｗｅｂ文書のランキングを行っている。しかしながら、Wikipedia（登

録商標）の文書には、１つの文書に付き１つの見出し語（キーワード）がついているため、Wikipedia（登録商標）のリンク構造から得られた文書のランキングをキーワードのランキングと見做すことができる。そのため、本装置では、キーワード重要算出部４において、Wikipedia（登録商標）に特化したランキング手法を適用することで、キーワード重要度１記憶部１０３に格納されているキーワード重要度を用いてキーワード固有重要度を算出し、固有重要度記憶部１０５に格納する。ベースとなるアルゴリズムには、HITSアルゴリズムを用いる。

【０１１４】

HITSアルゴリズムには、全てのＷｅｂ文書をauthority（コンテンツ）とhub（リンク集）の２つから構成されると定義する。そして、良いhubから多数リンクされるauthority程良いauthorityであるという仮説と、良いauthorityに多数リンクしているhub程良いhubであると言う二つの仮説を繰り返し実行することでＷｅｂ文書のランキングを行う。しかしながら、HITSアルゴリズムはＷｅｂ世界におけるＷｅｂ文書のリンク構造をモデルにしたアルゴリズムのため、リンク構造が非常に密なWikipedia（登録商標）にそのまま適用した場合、やや難がある。そこで、本装置では、Wikipedia（登録商標）の特徴的な構造と密なリンク構造に対応させた手法を提案する。そして、本アルゴリズムから算出したauthorityの値による順位を、本手法が提案する減衰関数に近似させ、最終的なキーワード固有重要度とする。

【０１１５】

●テキスト量の考慮：

Wikipedia（登録商標）の見出し語は、知名度が高く話題性の高い見出し語ほど、テキストの記述量が多い傾向がある。そこで、authority値の算出の際に、自文書のテキスト量が多ければ多いほどその文書は重要であるといった重み、text(k)を考慮する。

【０１１６】

●自リンクと被リンクの比率：

一般的にWikipedia（登録商標）の見出し語は、有名なキーワードほど、自リンクと被リンクの数が多くなっている。しかしながら、地名やジャンル名のような広い概念を持つキーワードは、引用しやすいキーワードのため、自リンク数に比べて圧倒的に被リンクの数が多い傾向がある。通常のHITSアルゴリズムは良いhubから多数リンクされているauthorityは良いauthorityであるといった仮説を用いるが、圧倒的に被リンクが多い場合においては、これらの仮説は成り立たないと予想される。また、その一方で、最近知名度が高くなってきている新人俳優や話題語等の見出し語は、誕生してから日が浅いため引用数は少ないが自リンクは多い傾向にある。そのため少ない被リンク数においても、authorityを高める必要がある。これらの問題を解決するために、authority値の算出の際に、flink(k)/blink(k)を考慮する。ここで、flink(k)はキーワードｋの文書内に含まれる自リンクの数を表す、blink(k)はキーワードｋの文書にリンクしている被リンクの数を表す。

【０１１７】

●明らかにauthority算出とならない見出し語の扱い：

Wikipedia（登録商標）の見出し語には「～年」や「～一覧」といった明らかにauthorityとならない見出し語が存在する。これらの見出し語は自リンクが非常に多く、被リンクも非常に多い場合があるためノイズとなりやすい。そこで、明らかにauthorityとならない見出し語のauthority値は常に変更しないことで、この問題に対処する。

【０１１８】

●hubの平均的なリンクの質：

Wikipedia（登録商標）の文書には、自リンクが多数あるが、hubとして質の悪い文書がある。そこで、リンク先キーワードのauthorityが平均的に高いhubは重要であるといった仮説に変更することで、自リンクは多いがhubとして質の低い文書のhub値を下げる重み

【０１１９】

【数５】

$$\sum_{k'} \log(a(k')+1) / K'$$

を考慮する。

【０１２０】

●リダイレクトの扱い：

Wikipedia（登録商標）の文書には、見出し語の異表記を解消するために、redirect（リダイレクト）が存在する。例えば、「イチロー」には「鈴木イチロー」、「ＩＣＨＩＲＯ」のredirectがある。Redirectは異表記のキーワードを一意に纏める効果だけでなく、キーワードの被リンク構造に大きな影響を持つため、redirectキーワードを親ノードに纏めることで、異表記のキーワード固有重要度を算出し、被リンクの問題も解決する。

【０１２１】

そして、最終的なWikipedia（登録商標）ランキングアルゴリズムは以下の式で定義される。

【０１２２】

【数６】

$$a(k) = \frac{\log(flink(k)+1}{\log(blink(k)+1} \cdot text(k) \cdot \sum_{k'} h(k')$$

$$h(k) = \frac{\sum_{k'} \log(a(k')+1}{K'} \cdot \sum_{k'} a(k') \qquad (6)$$

ここで、a(k)は、キーワードｋのauthority値で、h(k)はhub値である。そして、flink(k)はキーワードｋから自リンク数、blink(k)はキーワードｋからの被リンク数である。text(k)は、キーワードｋが見出し語になっているWikipedia（登録商標）文書の文字数（アンカーリンク対象の文字列は除く）であり、K'はキーワードｋが見出し語になっているWikipedia（登録商標）文書内に含まれるリンク数の総数となる。なお、式（６）は交互に繰り返し計算を行うものとし、１０回ほど繰り返し行うとよい。

【０１２３】

そして、上記の式（６）で算出したauthority値を用いて、降順にキーワードを順位付けする。そして、以下のキーワード固有重要度算出式を用いてスコアの近似を行い、Wikipedia（登録商標）内のリンク構造を用いたキーワード固有重要度WKIS(k)とし、固有重要度記憶部１０５に格納する。

【０１２４】

【数７】

$$WKIS(k) = \exp\left(\frac{\log(y_1 - y_0 + 1)(K - k_r + 1)^a}{K^a}\right) \div y_0 - 1 \qquad (7)$$

ここで、y_1はキーワード固有重要度の上界であり、y_0はキーワード固有重要度の下界である。そして、k_rはキーワードｋのauthorityの値による順位、Ｋはキーワードの総数、ａはスコアの勾配係数で、ａの値が大きくなればなるほどスコアの勾配が急になる。本関数の特徴は、キーワードの候補数（ｘ軸の要素数）に左右されることなく、また、上界と下界を設定でき、またキーワードの候補数の上位１０％以内において、最大値と最小値の差が３０％～７０％以内に収まる減衰曲線を描き、そしてパラメータ係数にて減衰度合いを整数値を用いることで容易に設定できる特徴をもつスコア関数である（勿論、減衰度合いの設定は実数値でも設定可能である）。y_1、y_0の推奨値は、それぞれ1，0.1で、勾配係数の推奨値は3から7である。

【０１２５】

Wikipedia（登録商標）内のリンク構造から算出したキーワード固有重要度は、ユーザ参加型のオンライン百科事典のWikipedia（登録商標）内においての重要なキーワードが上位に位置付けられるため、一般に知名度は低いが実際には内容が深く重要なキーワードが上位に位置づけられる。

【０１２６】

［３－５－２］検索エンジンの検索クエリ投入回数を用いた手法：

この手法は、実際の検索エンジンに投入された回数が多い検索クエリほど、重要なキーワードであるとみなす手法である。検索クエリの投入回数は、ポータルサイト上で投入された検索クエリのような大規模データであることが好ましい。ここで、対象となる検索クエリは、Wikipedia（登録商標）に存在するキーワードであるとする。この手法もWikipedia（登録商標）内のリンク構造を用いた手法と同じく式（７）のキーワード固有重要度算出式を用いてスコアの近似を行い、検索クエリを投入回数を用いたキーワード固有重要度QKIS(k)とし、固有重要度記憶部１０５に格納する。

【０１２７】

【数８】

$$QKIS(k) = \exp\left(\frac{\log(y_1 - y_0 + 1)(K - k_r + 1)^a}{K^a}\right) + y_0 - 1 \qquad (8)$$

ここで、y_1はキーワード固有重要度の上界であり、y_0はキーワード固有重要度の下界である。そして、k_rはキーワードｋの検索回数による順位、Kはキーワードの総数、aはスコアの勾配係数で、aの値が大きくなればなるほどスコアの勾配が急になる。本関数の特徴は、キーワードの候補数（ｘ軸の要素数）に左右されることなく、また、上界と下界を設定でき、またキーワードの候補数の上位１０％以内において、最大値と最小値の差が３０％～７０％以内に収まる減衰曲線を描き、そしてパラメータ係数にて減衰度合いを整数値を用いることで容易に設定できる特徴をもつスコア関数である（勿論、減衰度合いの設定は実数値でも設定可能である）。

【０１２８】
また、Wikipedia（登録商標）には見出し語として存在するが、検索エンジンの検索クエリには存在しない場合がある。その場合はy_0を該当キーワードのスコアとする。y_1，y_0の推奨値は、それぞれ、1，0.1で、勾配係数の推奨値は3から7である。
【０１２９】
キーワード重要度２記憶部１０４に格納されている検索エンジンの検索クエリ投入回数から算出したキーワード固有重要度は、実際の検索クエリ投入回数によって重要度が決まるため、検索クエリとして投入されやすく、重要なキーワードが上位に来る傾向にある。
【０１３０】
［３－６］最終的なキーワード固有重要度算出：
前述の［３－５］では、Wikipedia（登録商標）内のリンク構造から算出したキーワード固有重要度WKIS(k)と、検索エンジンの検索クエリ投入回数から算出したキーワード固有重要度QKIS(k)について述べた。しかしながら、QKIS(k)は検索クエリデータの収集期間が短い場合、その間にインターネット上で起きた話題に強く影響されてしまう問題があり、さらにインターネットサイト名等の生活的クエリが多く含まれる傾向にある。そのため、キーワード重要度算出部４では、固有重要度記憶部１０５に格納されているWKIS(k)とQKIS(k)の線形和を、最終的なキーワード固有重要度とすることで、話題性が高く内容が深く、かつ、検索クエリとして投入されやすいキーワードを上位に位置づける重要度を算出する。
【０１３１】
Keyword_score(k)=α・WKIS(k)+(1-α)・QKIS(k)　　（９）
QKIS(k)よりも、WKIS(k)のキーワード重要の方が一般的に精度が良いことが実験により確認できたため、αの値は0.5～0.8辺りの範囲で調整する。
【０１３２】
［３－７］文書中での出現位置におけるキーワード重要度算出：
以下では、位置情報算出部５における処理を説明する。
【０１３３】
ニュース記事やコラム等の一般的な記述がなされる記事においては、文の先頭に来れば来るほど、重要なキーワードが含まれている可能性が高い。そのため、文の先頭にあればあるほど、そのキーワードは重要であるという重要度を算出する。この手法も、Wikipedia（登録商標）内のリンク構造を用いた手法と同じく式（７）のキーワード固有重要度算出式を用いて、出現位置を用いたキーワード重要度Pos(k)の算出を行い、位置重要度記憶部１０６に格納する。
【０１３４】
【数９】

$$Pos(k)=\exp\left(\frac{\log(y_1-y_0+1)(P-k_p+1)^a}{P^a}\right)+y_0-1 \qquad (10)$$

ここで、y_1は出現位置を用いたキーワード重要度の上界であり、y_0は出現位置を用いたキーワード重要度の下界である。そしてk_pはキーワードｋの文位置（文の位置は、文の先頭から１文ずつ（句点までを１文とする）数え上げたものを用い、さらにそのキーワードにおいて、最も先頭の位置を用いる）、Pは最後尾の文の位置、aはスコアの勾配係数で、aの値が大きくなればなるほどスコアの勾配が急になる。本関数の特徴は、キーワードの候補数（ｘ軸の要素数）に左右されることなく、また、上界と下界を設定でき、またキーワードの候補数の上位１０％以内において、最大値と最小値の差が３０％～７０％以内に

収まる減衰曲線を描き、そしてパラメータ係数にて減衰度合いを整数値を用いることで容易に設定できる特徴をもつスコア関数である（勿論、減衰度合いの設定は実数値でも設定可能である）。y_1、y_0の推奨値は、それぞれ1、0.5で勾配係数の推奨値は1から5である。ブログ記事のような一般的な記述がなされていない文書を多く処理する場合には、y_0の値を大きくし、勾配係数の値も小さくするのがよい。

【０１３５】

［３－８］文書における最終的なキーワード重要度：

以下に、重要キーワード出力部６の動作を説明する。

【０１３６】

文書d内における最終的なキーワード重要度FS(d,k)は、［３－７］節までに述べた手法により求められ、固有重要度記憶部１０５と位置重要度記憶部１０６に格納されている各値を乗算することによって算出する。FS(d,k)が高いほど重要なキーワードと見做す。

【０１３７】

$$FS(d,k)=BM25_{all}(d,k)\cdot WebIDF(k)\cdot Keyword\_score(k)\cdot Pos(k) \qquad (11)$$

FD(d,k)は全てのキーワード候補に対して、キーワードの重要度を算出するため、応用アプリケーション等で少数のキーワードしか表示できない場合、上位３～５位のキーワードをその文書内における重要キーワードとして表示する。

【０１３８】

なお、上記の重要キーワード抽出装置の各構成要素の動作をプログラムとして構築し、重要キーワード抽出装置として利用されるコンピュータにインストールして実行させる、または、ネットワークを介して流通させることが可能である。

【０１３９】

また、構築されたプログラムをハードディスクや、フレキシブルディスク・ＣＤ－ＲＯＭ等の可搬記憶媒体に格納し、コンピュータにインストールする、または、配布することが可能である。

【０１４０】

なお、本発明は、上記の実施の形態に限定されることなく、特許請求の範囲内において種々変更・応用が可能である。

【産業上の利用可能性】

【０１４１】

本発明は、Ｗｅｂ文書のキーワードを抽出する技術、例えば、情報検索の事前処理や情報推薦等に適用可能である。

【符号の説明】

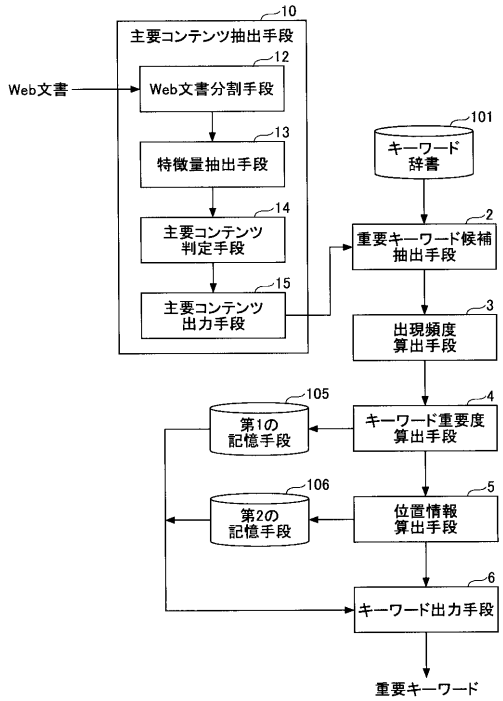
【０１４２】

１　文書入力部
２　重要キーワード候補抽出手段、重要キーワード候補抽出部
３　出現頻度算出手段、出現頻度算出部
４　キーワード重要度算出手段、キーワード重要度算出部
５　位置情報算出手段、位置情報算出部
６　キーワード出力手段、重要キーワード出力部
１０　主要コンテンツ抽出手段、主要コンテンツ抽出部
１１　Ｗｅｂ文書取得・入力部
１２　Ｗｅｂ文書分割手段、Ｗｅｂ文書分割部
１３　特徴量抽出手段、特徴量抽出部
１４　主要コンテンツ判定手段、主要コンテンツ判定部
１５　主要コンテンツ出力手段、主要コンテンツ出力部
１０１　キーワード辞書
１０２　ＷｅｂＩＤＦ記憶部
１０３　キーワード重要度１記憶部
１０４　キーワード重要度２記憶部

１０５　第１の記憶手段、固有重要度記憶部
１０６　第２の記憶手段、位置重要度記憶部
１１１　データ入力部
１１２　Ｗｅｂ文書ファイル入力部
１１３　ＵＲＬ入力部
１１４　Ｗｅｂ文書取得部
１１５　文字コード変換部
１２１　広告対象領域抽出部
１２２　ノイズとなるタグや領域除去部
１２３　Ｗｅｂ文書の分割部
１３１　アンカーリンク情報特徴量抽出部
１３２　タグ情報特徴量抽出部
１３３　Ｗｅｂ文書で表示される文字列特徴量抽出部
１３４　特徴量の正規化部
１３５　特徴量の比率特徴量抽出部
１４１　特徴量入力部
１４２　テキスト判定部
１４３　主要コンテンツ判定部
１５１　タグ付テキスト出力部
１５２　タグなしテキスト出力部
１５３　データ出力部

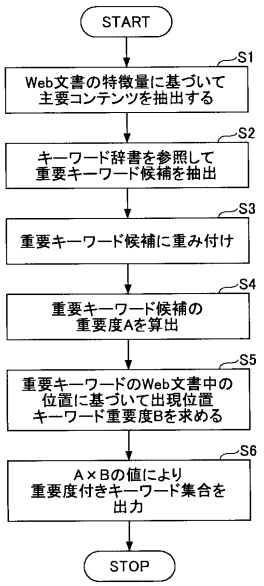
【図１】

本発明の原理構成図



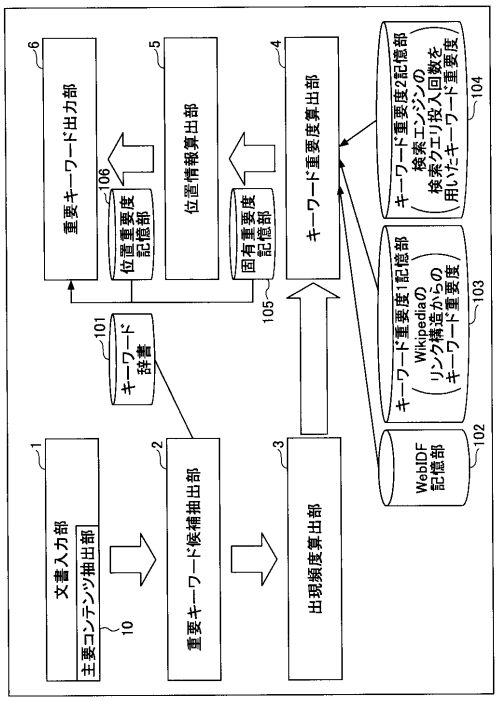
【図２】

本発明の原理を説明するための図



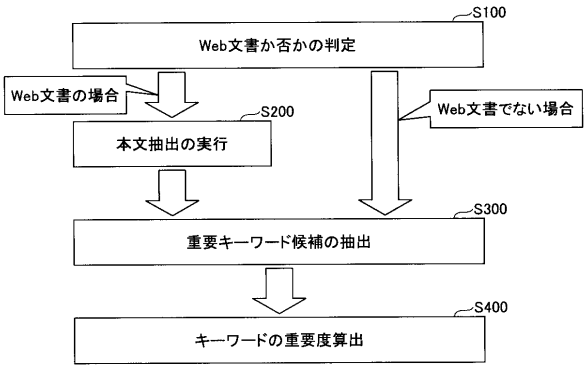
【図３】

本発明の一実施の形態における重要キーワード抽出装置の構成図



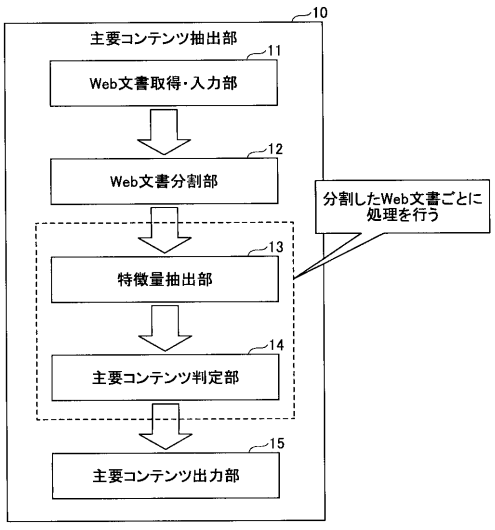
【図４】

本発明の処理の概要を示す図



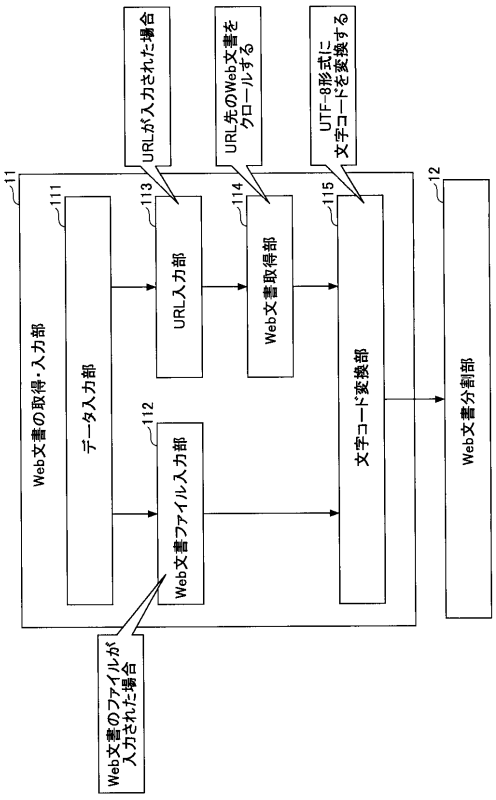
【図５】

本発明の一実施の形態における主要コンテンツ抽出部の構成図



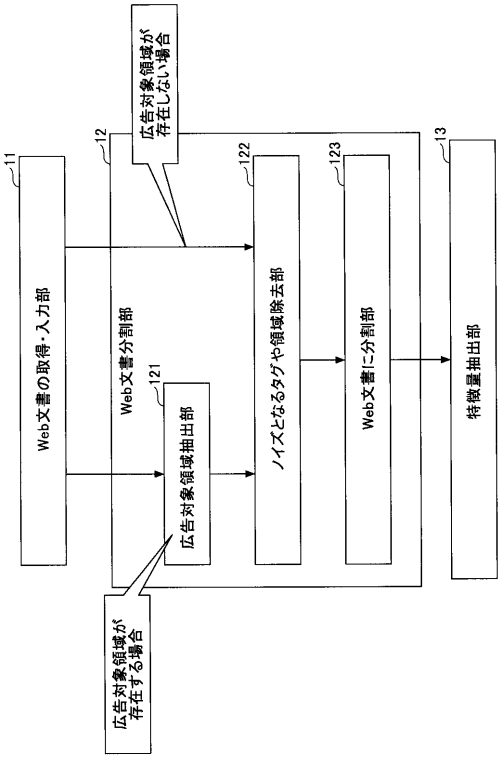
【図６】

本発明の一実施の形態におけるWeb文書取得・入力部の構成図



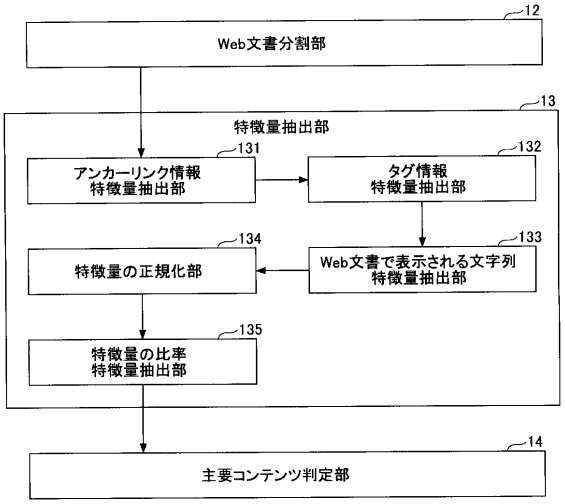
【図７】

本発明の一実施の形態におけるWeb文書分割部の構成図



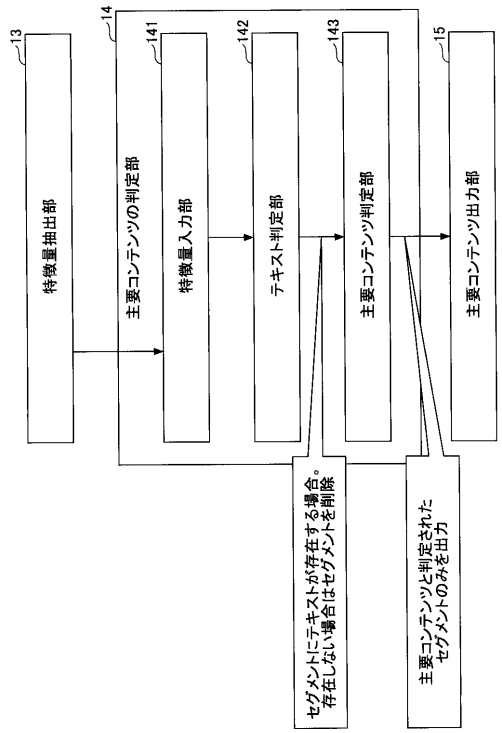
【図８】

本発明の一実施の形態における特徴量抽出部の構成図



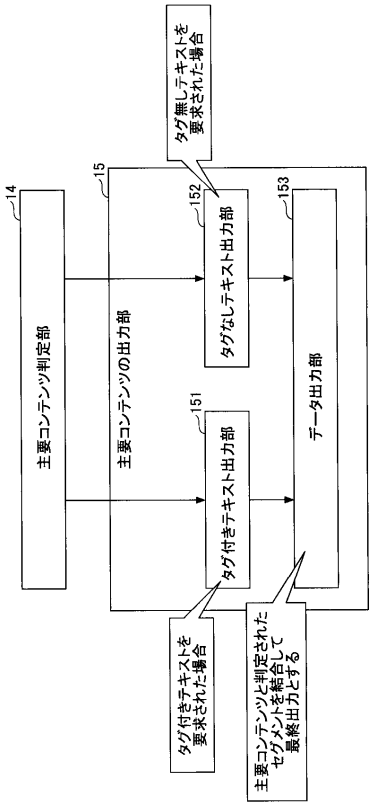
【図９】

本発明の一実施の形態における主要コンテンツ判定部の構成図



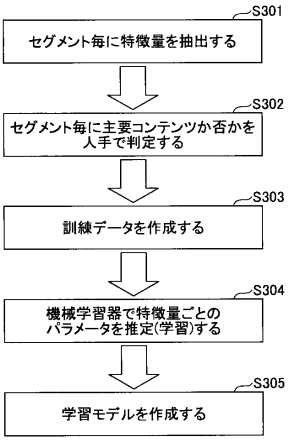
【図１０】

本発明の一実施の形態における主要コンテンツ出力部の構成図



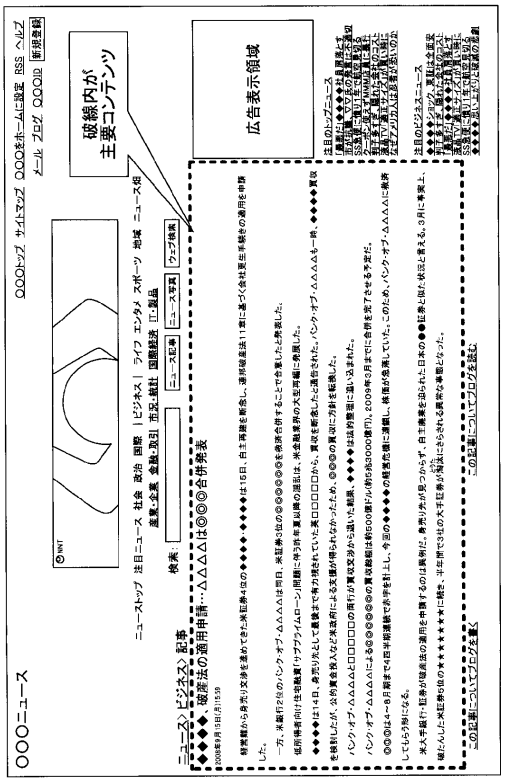
【図１１】

本発明の一実施の形態における特徴量のパラメータ推定方法のフローチャート



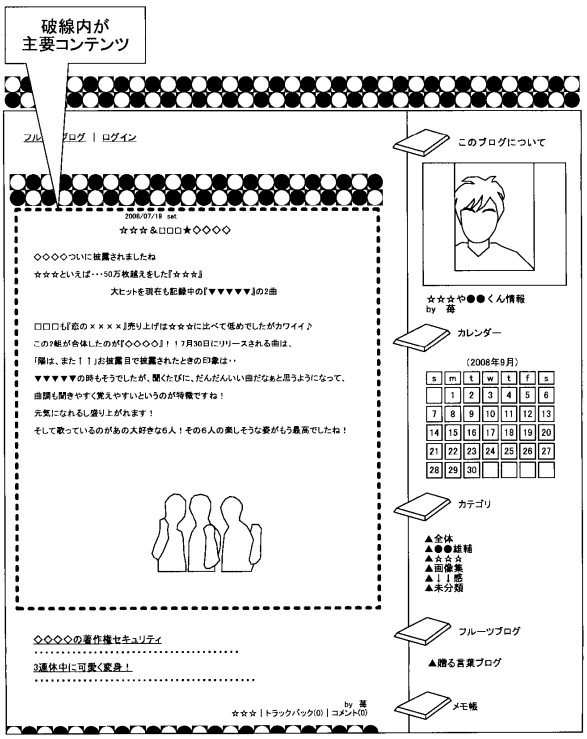
【図１２】

本発明の一実施の形態における主要コンテンツ例(その1)



【図１３】

本発明の一実施の形態における主要コンテンツ例(その2)



フロントページの続き

(72)発明者　内山　匡
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

審査官　鈴木　和樹

(56)参考文献　近藤光正、外３名，PC上のWeb閲覧履歴からのクエリ抽出技術を用いたモバイル情報検索システム，２００８年度人工知能学会全国大会（第２２回）論文集　［ＣＤ－ＲＯＭ］，２００８年６月１１日，ｐ．１－４（２Ｐ２－１０）
近藤光正、外３名，ＨＩＴＳに基づくＷｉｋｉｐｅｄｉａランキングアルゴリズムとユーザ履歴を用いた個人適応型クエリ推薦，電子情報通信学会　第１９回データ工学ワークショップ論文集　［ｏｎｌｉｎｅ］，日本，電子情報通信学会データ工学研究専門委員会，２００８年　４月７日，ｐ．１－８（Ｂ２－４），Internet<URL:http://www.ieice.org/iss/de/DEWS/DEWS2008/proceedings/files/b2/b2-4.pdf>

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０