

(19) 世界知的所有権機関
国際事務局(43) 国際公開日
2007年1月25日 (25.01.2007)

PCT

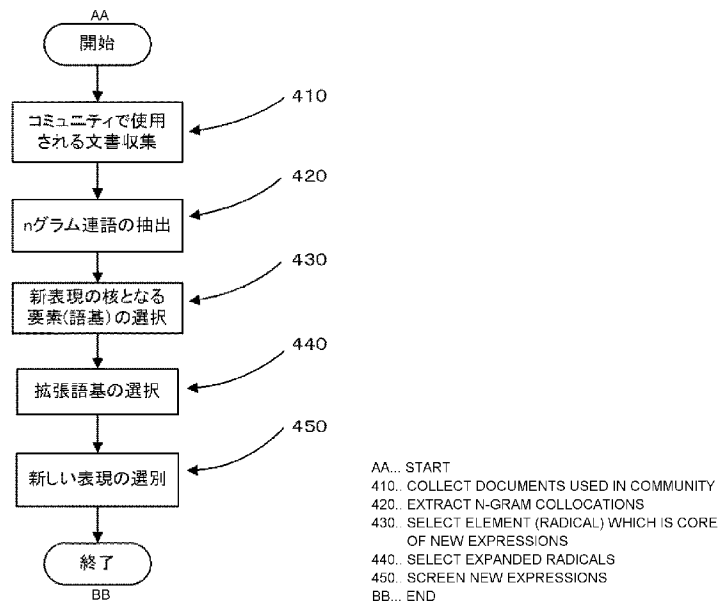
(10) 国際公開番号
WO 2007/010836 A1

- (51) 国際特許分類:
G06F 17/21 (2006.01) G06F 17/30 (2006.01)
G06F 17/28 (2006.01)
- (21) 国際出願番号: PCT/JP2006/314000
- (22) 国際出願日: 2006年7月13日 (13.07.2006)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願2005-207810 2005年7月15日 (15.07.2005) JP
- (71) 出願人 (米国を除く全ての指定国について): ヒューレット・パカード デベロップメント カンパニー エル. ピー. (HEWELETT-PACKARD DEVELOPMENT COMPANY, L.P.) [US/US]; 77070 テキサス州 ヒューストン 20555 ステイトハイウェイ 249 Texas (US).
- (72) 発明者; および
- (75) 発明者/出願人 (米国についてのみ): 小田 弘美 (ODA, Hiromi) [JP/JP]; 〒1688585 東京都杉並区高井戸東3丁目29-21 日本ヒューレット・パカード株式会社内 Tokyo (JP).
- (74) 代理人: 大西 昭広 (ONISHI, Akihiro); 〒1680072 東京都杉並区高井戸東3丁目8-13 日本ヒューレット・パカード株式会社内 Tokyo (JP).
- (81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SY,

[続葉有]

(54) Title: COMMUNITY SPECIFIC EXPRESSION DETECTING DEVICE AND METHOD

(54) 発明の名称: コミュニティ特有表現検出装置及び方法



(57) Abstract: The prior art concerning collections of community specific expressions includes collections of technical terms including nouns and compound nouns in technical fields. However, application to new expressions other than nouns is difficult. Even in the field of collection of unknown words and new words, the objective is limited substantially to nouns, and no techniques of collecting new expressions systematically have been proposed. The invention solves the above problem by (a) means for extracting n-gram collocations specific in a predetermined community from a set of documents used in the community, (b) means for selecting a radical which might be a core of specific expressions, (c) means for expanding the selected radical toward the front and back, and (d) means for screening the expanded radicals according to the grammar.

(57) 要約: コミュニティ固有表現の収集に関係する従来技術では、専門的分野における名詞・複合名詞からなる専門用語の収集に関するものがあるが、名詞以外の新しい表現には応用が難しい。また、未知語・

[続葉有]



WO 2007/010836 A1



TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

- (84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR),

添付公開書類:

— 国際調査報告書

2文字コード及び他の略語については、定期発行される各PCTガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

新語の収集の分野においても、対象はほぼ名詞に限定されて、新しい表現を規則的に収集するという手法は提案されていない。 所定のコミュニティで使用される文書集合の中から、(a) コミュニティに固有のnグラム連語を抽出する手段、(b) 固有の表現の核となる可能性のある語基を選択する手段、(c) 前記選択された語基をその前後に拡張する手段、(d) 前記拡張された語基を文法に従って選別する手段によって、上記問題を解決している。

明 細 書

コミュニティ特有表現検出装置及び方法

技術分野

- [0001] 語形成論に基づき、コミュニティで使用される表現の中から当該コミュニティ特有の表現を検出する装置及び方法に関する。

背景技術

- [0002] 特定の興味やテーマをめぐって活発な議論が交されている人々のコミュニティにおいては、往々にしてそのコミュニティ独自の表現が発生する。例えば、日本酒の味を議論するコミュニティにおいては、「老ね(ヒネ)、ヒキのある、キレル、・・・」といった表現が用いられる。ワインを好む人々の間では「フルボディ、ミディアムドライ、樽香、後口、・・・」のような表現が見られる。これらは専門知識を有する人々の用いる難解な専門用語ではなく、ワインや日本酒の味に親しむ人であれば、その味を言い表す表現として自然にその意味が理解される種類の語彙である。また、高校・大学生等の「若者語」として集められている表現もコミュニティ固有の表現と考える事ができる。最近では、インターネットの掲示板などに集まる人々の成すコミュニティにおいて多くの新しい表現が見いだされるようになっている。

特許文献1:特開2002-297589「未知語収集方法」

特許文献2:特開平5-113997「辞書データ収集装置」

特許文献3:特開2004-265440「未知語登録装置および方法並びに記憶媒体」

特許文献4:特開2005-309853「専門的記述と非専門的記述間の語彙変換方法・プログラム・システム」

非特許文献1:中川祐志, 湯本紘彰, & 辰則 (2003). 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, 10(1), 27-45.

非特許文献2:辻慶太, & 芳鐘冬樹 (2004). 専門分野において重要となる新語の特定に向けた基礎研究. 言語処理学会第10回年次大会発表論文集, (pp. 189-191).

非特許文献3:藤井敦, 伊藤克亘, 秋葉友良 (2003), IPA未踏ソフトウェア創造事業「CYCLONE: 最強事典サイトの構築」, www.ipa.go.jp/about/news/event/pdf/29A7_f

ujii.pdf

非特許文献4:米川明彦(1998)「若者語を科学する」東京:明治書院

発明の開示

発明が解決しようとする課題

[0003] コミュニティ固有表現の収集に関係する既存技術には、主に専門用語の収集と未知語の収集に関するものがある。専門用語の収集については、非特許文献1、非特許文献2を始めとした研究があるが、ほとんどは専門的分野における名詞、複合名詞からなる専門用語の収集に関するものである。このように限定する事によって、単名詞の重なりや接続関係等に着目したスコアに基づいたアルゴリズムを用いる事ができるが、名詞以外の表現には応用が難しい。

また、未知語・新語の収集については、辞書の構築等においても重要なテーマであり、特開2002-297589「未知語収集方法」(特許文献1)、特開2004-265440「未知語登録装置および方法並びに記憶媒体」(特許文献3)等、既存特許にもこのテーマを扱った技術が存在する。

[0004] しかしながら、非特許文献3等の報告にもあるように日本語における未知語の検出は困難な問題であり、特開2002-297589「未知語収集方法」(特許文献1)の方法もそうであるが、基本的には辞書に登録されていないものを人手やヒューリスティクスによって収集しているものが多い。また、これら未知語の検出においても対象はほぼ名詞に限定されており、真に新しい表現の収集という問題に焦点を絞ったものはまれである。

また、社会言語学において、高校生・大学生の用いる「若者語」の収集と分析を行う分野が存在する(非特許文献4)。コミュニティ固有の表現についての既存研究としては、本願発明に近いと思われるが、社会言語学分野で、若者語や流行語を規則的に収集するという手法は提案されていない。

課題を解決するための手段

[0005] 以下の装置を開示することにより課題を解決している。

(1)

以下の(a)から(d)の手段を有する所定のコミュニティで使用される文書集合から前

記所定のコミュニティに特有な表現を検索する装置、

- (a) 前記コミュニティに特有に使用されるnグラム連語を抽出する手段、
- (b) 前記特有な表現の核となる可能性のある第一の語基を選択する手段、
- (c) 前記第一の語基の有意度、及び、前記第一の語基の前又は後の要素を取込んだ第二の語基の有意度を用いて算出された値に基づいて拡張語基を選択する手段、
- (d) 前記拡張語基の中から当該言語の語形成規則に従って前記所定のコミュニティに特有な表現を選別する手段。

[0006] (2)

さらに、前記文書集合を、所定の用語リストに含まれる用語をキーワードとしてデータ検索することによって収集する手段を含むことを特徴とする(1)に記載の装置。

(3)

前記nグラム連語を抽出する手段は、複数のコミュニティで使用される文書を用い、前記所定のコミュニティで使用されるnグラム連語の有意度と、他のコミュニティで使用されるnグラム連語との有意度との比較に基づいて前記nグラム連語を抽出する手段を含むことを特徴とする(1)及び(2)に記載の装置。

[0007] さらに、以下の方法を開示することにより課題を解決している。

(4)

以下の(a)から(d)のステップを有する、所定のコミュニティで使用される文書集合から前記所定のコミュニティに特有な表現を検索する方法、

- (a) 前記コミュニティに特有に使用されるnグラム連語を抽出するステップ、
- (b) 前記特有な表現の核となる可能性のある第一の語基を選択するステップ、
- (c) 前記第一の語基の有意度、及び、前記第一の語基の前又は後の要素を取込んだ第二の語基の有意度を用いて算出された値に基づいて拡張語基を選択するステップ、
- (d) 前記拡張語基の中から当該言語の語形成規則に従って、前記所定のコミュニティに特有な表現を選別するステップ。

(5)

さらに、前記文書集合を、所定の用語リストに含まれる用語をキーワードとしてデータ検索することによって収集するステップを含むことを特徴とする(4)に記載の方法。

[0008] さらに、以下のプログラムを開示することにより課題を解決している。

(6)

コンピュータを制御して、以下の(a)から(d)の手段を動作させ、所定のコミュニティで使用される文書集合から前記コミュニティに特有な表現を検索するプログラム、

(a) 前記コミュニティに特有に使用されるnグラム連語を抽出する手段、

(b) 前記特有な表現の核となる可能性のある第一の語基を選択する手段、

(c) 前記第一の語基の有意度、及び、前記第一の語基の前又は後の要素を取込んだ第二の語基の有意度を用いて算出された値に基づいて拡張語基を選択する手段、

(d) 前記拡張語基の中から当該言語の語形成規則に従って前記所定のコミュニティに特有な表現を選別する手段。

(7)

さらに、前記文書集合を、所定の用語リストに含まれる用語をキーワードとしてデータ検索することによって収集する手段を含むことを特徴とする(6)に記載のプログラム。

発明の効果

[0009] 本願発明に従って、所望のコミュニティで使用される表現を収集しその意味を理解することは、コミュニティのメンバーにとってコミュニケーションが容易になり、さらに、そのアイデンティティを確認するのに役に立てることが出来る。また、そのコミュニティの特徴や性格を分析する目的に役立てる事ができる。

さらに、商品の開発等においてユーザのコミュニティで交される議論の内容を分析することが有用であると思われるが、この場合当該コミュニティ固有の表現を収集しその意味を理解する事は、この目的に大きく貢献すると考えられる。

また、本願発明は、主要品詞間の語法の拡張であり、他の言語にも応用可能である。英語の例を挙げると、「He 747'ed to Chicago.」という表現が可能である。これは航空機の型番を動詞化したものである。また、「The web-logging is becoming a social phenomenon.」という表現も用いられるが、これは「Web-log(ウェブに書き込む)」とい

う動詞が名詞化された例である。

発明を実施するための最良の形態

[0010] 以下に最良の形態を説明する。

実施例 1

[0011] 図1は、本願発明を実施する場合のシステム例を示している。ネットワーク140には、ユーザPC110、サイトサーバ(1)120、サイトサーバ(2)130等が接続されている。使用者がユーザPC110を操作することにより、ネットワーク140に接続されているサイトサーバ(1)120、サイトサーバ(2)130等をアクセスし、検索ツール等を使用して必要な情報を取得する。本願発明はインターネットでの検索を実施例として示すが、これに限らず、情報が検索できるシステムならば他の方法でも応用できる。取得した情報をユーザPC上のコンピュータプログラムで処理し、所望の結果を得ることが出来る。

[0012] 図2は、本願発明の一部を実施するユーザPCを示している。筐体200の中には、記憶装置210、メインメモリー220、出力装置230、中央制御装置(CPU)240、操作装置250、ネットワークI/O260が含まれている。使用者が操作装置250を操作し、ネットワークI/Oを通して、必要な情報をインターネットの各サイトから入手する。中央制御装置240は記憶装置210に記憶されている文書処理プログラムをメモリにダウンロードし、インターネットから検索された情報を用いて所定のデータ処理を行い出力装置230に結果を表示する。

[0013] 図3は、本願発明によるコミュニティ固有表現検出装置のブロック図を示している。310はコミュニティ文書検索部、314はウェブサイト、316は用語リスト格納部、320は文書処理部、330はnグラム連語抽出部、335は有意度判定部、340は語基選択部、350は語基の左右拡張部、354は左側拡張規則格納部、356は右側拡張規則格納部、360は新表現の選別部、365は言語規則格納部、370は出力部を表す。

以下、これらの詳細について説明する。

[0014] [基本アルゴリズム]

図4に示すフローチャートに従って、本願発明の基本アルゴリズムを説明する。

ステップ410:コミュニティで使用する文書の収集

ステップ420:nグラム連語の抽出

ステップ430:新表現の核となる要素(語基)の選択

ステップ440:拡張語基の選択

ステップ450:新しい表現の選別

[0015] [アルゴリズムの詳細]

以下にアルゴリズムの詳細について説明する。

(1)所定のコミュニティで使用される文書の収集(図4 ステップ410)

先ず、所定のコミュニティで使用される文書集合を次のステップで収集する。図5に示されるアルゴリズムを参照。

ステップ510:用語の指定による候補文書の取得

ステップ520:候補文書の前処理

ステップ530:ノイズ文書の除去

ステップ540:他のコミュニティ文書の検索の要否

以下、各ステップについて詳細に説明する。

[0016] (1-1)ステップ510:候補文書の取得

本願発明を実施する為には、所定の用語を含む用語リストを用いて、所定のコミュニティの関係者が使用する文書を収集する。ここで用語リストは用語リスト格納部(図3:316)に格納されている。

ここで用語リストとは、一つのコミュニティにおけるキーワードとなる用語の集合である。例えば、一つのコミュニティとして「ワインの愛好家」を選択すると、用語リストの構成要素は「ワインの銘柄」である。ワインの用語リスト中に記載されている銘柄に従い、インターネットの検索ツールを使用して、ワインに関する情報を収集する(図3:314)。ここで、銘柄としては、「アウスレーゼ」、「シャトー キュレ ボン」、「シャトー マルゴー」、「ヴィン サント トスカーノ」等の銘柄を指定することが出来る。この用語をキーワードとして、データベースから候補の文書を検索する。データベースとしてはこのような情報が格納されているデータベースならば何でも構わないが、本実施例ではインターネットの検索エンジンを使用して、候補の文書を検索する方法について説明する。

[0017] (1-2)ステップ520:候補文書の前処理

前処理では、先ずウェブページの情報から文書に相当するものを取り出し文書解析を行なう。次に、分かち書きを行ない内容語、助詞、助動詞等を抽出し、これらの文書の特徴を表す特徴値を求める。これらの特徴値を用いて、以下の様にノイズ文書を除去する。また、収集しようとする文書の典型とみなされるような少量のモデル文書を前もって選定しておく。

[0018] (1-3)ステップ530:ノイズ文書の除去

インターネットのウェブページから自動的にこれらの情報を収集した文書には様々な情報が含まれており、そのままでは利用できない場合が多い。本実施例ではこれらの文書の中から、ガービッジ文書、リスト文書、及び日記型文書に該当する文書をノイズ文書として除去している。

以下に、ガービッジ文書、リスト文書、及び日記型文書について説明する。

(a)ガービッジ文書

内容語数が少ない文書、あるいは、固有名詞比率の低い文書等の条件の全てを満足する文書を言う。内容語数とは、一つのウェブページに記載されている文書に含まれているに内容語の数である。内容語とは助詞・助動詞を除いた、名詞、動詞、形容詞、副詞に該当する単語である。また、ここで言う固有名詞とは、世間一般に固有名詞であると認識されている名詞である。固有名詞比率とは一つのウェブページに出現する固有名詞の数と内容語数との比率である。

(b)リスト文書

固有名詞比率が高い文書、内容語と助詞・助動詞との相関係数が低い文書等の条件の全てを満足する文書をリスト情報文書と定義する。これはインターネットのサイトにおいて、ある領域における対象物に関する情報が単なるリストとして格納されている文書である。

[0019] (c)日記型文書

あるコミュニティに関する固有名詞比率が低い文書、内容語nグラムに基づくモデル文書との相関度が低い文書、助詞・助動詞nグラムに基づくモデル文書との相関度が高い文書等の条件の全てを満足する文書を日記型文書と定義する。これらは言わば

個人の日記書き込みサイトとして利用されている文書、及び、デパートの売場に関するサイトなど、主として他の情報が記載されて要る文書である。以上の定義に基づいて、ガービッジ文書、リスト文書、及び、日記型文書をノイズ文書として除去する。

[0020] (1-4)ステップ540:他のコミュニティ文書の検索の要否

ステップ510からステップ530により、所定のコミュニティで使用する文書集合が収集される。ステップ540では、他のコミュニティで使用する文書集合を同様に収集する。

[0021] 次にこれらの収集された複数のコミュニティで使用する文書集合を用いて、これらのコミュニティで固有に使用される新しい表現を選別する。

以上により、複数のコミュニティで使用する文書集合が作成される(図3:320)。

[0022] (2)nグラム連語の抽出(図4 ステップ420)

(2-1)コミュニティ固有の連語抽出

単語レベルのn-gram連語(nグラム連語)を統計的手法により、特定のコミュニティで 사용되는場合に有意に出現するものを抽出する。これらをコミュニティ固有の連語と呼ぶ。これらの詳細について説明する。

nグラム連語とは、連続した1以上の語であって、一語の場合はユニグラム(Uni-gram)、二語の場合はバイグラム(Bi-gram)、三語の場合はトライグラム(Tri-gram)と呼ばれる。本実施例では、バイグラム、トライグラムを用いている(図3:330)。

[0023] (2-2)有意度による判定

単純にnグラム連語を求めると数多くのnグラム連語が得られるが、全てのnグラム連語が有効であるとは限らない。そこで、二つのコミュニティで使用する文書集合を比較し、一方のコミュニティで使われているnグラム連語が、一方に有意に偏って出現するnグラム連語を選択する(Z検定)。本願明細書では、2つの文書集合においてそれぞれのnグラム連語の出現する比率を比較し、その比率差を検定する方法を用いる(図3:330)。ここで、あるnグラム連語Wが2つの文書集合d1, d2に共に表れたと考え、その頻度がw1, w2であったとする。文書集合d1に表れた用語の総数をn1, 文書d2のそれをn2とする。するとWがそれぞれの文書集合に表れた割合は次のようになる。

[0024] (式1) $p1 = w1/n1$,

(式2) $p2 = w2/n2$

ここで、標本比率を実際のデータから得られた比率とすると、 $p1$ 及び $p2$ は標本比率である。

ここで、 $p1 > p2$ である場合に、これが有意であるかどうかを検定する、すなわち、 n グラム連語 W は $d1$ の文書の方に有意に偏って出現するかどうかを検定するということを意味する(片側検定)。

ここで、帰無仮説と対立仮説は次のようになる。

$H0: p1 = p2$ 帰無仮説

$H1: p1 > p2$ 片側検定における対立仮説

検定を行うために、まず実際には知られていない母比率 π (式3)を標本比率から推定する。

(式3) $\pi = (n1 \cdot p1 + n2 \cdot p2) / (n1 + n2)$

ここから z を(式4)で計算する。

(式4) $z = (p1 - p2) / \sqrt{(\pi(1-\pi) \cdot (1/n1 + 1/n2))}$

帰無仮説を棄却し、対立仮説を採用するには、5%の危険率において、 $z > 1.65$ でなくてはならないことになる。

[0025] このようにして、全ての連語について検定を行い、文書集合の中に現れる n グラム連語であって、一方のコミュニティで使用される文書に有意に出現する n グラム連語、及び、他方のコミュニティで使用される文書に有意に出現する n グラム連語をそれぞれ選択することが出来る。従って、双方のコミュニティで共通に使用されるものは選択されないこととなる。

本願実施例では、ワインの愛好家が使用する文書集合と、日本酒の愛好家が使用する文書集合とに特徴的に現れる2グラム、3グラムのリストを取り出し、 Z 検定を行っている。ここで、 Z 検定の結果、ワインの愛好家が使用する文書集合から、 Z 値が1.65以上の n グラムを選択する。

[0026] (3) 新表現の核となる要素(語基)の選択(図4 ステップ430)

ここで、上記の方法によって抽出された n グラムについて、その中から新しい表現の

核となる要素を取り出す(図3:340)。そのためには、nグラム連鎖をひとまず切り離し、そこに生じる全ての要素(形態素)のリストを作る。そこから、核となる可能性のないものを除外する。ここで、核となる可能性の無いものとしては、助詞、助動詞、接続詞、活用語尾等の機能語、「、」、「。」、「?」等の切れ目要素がある。また、「平仮名一文字」、「片仮名一文字」のものも除外する。これによって、新表現の核となる可能性のある要素のリスト(核リスト)が作成される。

[0027] (4) 拡張語基の選択(図4 ステップ440)

(4-1) 語基の拡張

それぞれの語基候補について、連語パターンの分布に基づき、前後の要素を取り込んで拡張する必要があるかどうかを判断する(図3:350)。

ここで、(式5)の様に Z_{ratio} を定義する。

$$(式5) \quad Z_{ratio} = Z[X] / \text{Avg}Z([X][X+1])$$

ここで、 $Z[X]$ とは現在着目しているnグラム語基のZ値である。核要素をXとし、それに1語拡張した要素を $[X+1]$ とし、2語拡張した要素を $[X+2]$ とする。 $\text{Avg}Z([X][X+1])$ とはnグラム語基から「右」に一語拡張したときの、 $[X][X+1]$ に相当する全ての $(n+1)$ グラムの語基のZ値の平均値である($0 < Z_{ratio}$)。

正確に言えば、nグラム語基から「左」に一語拡張したときの $\text{Avg}Z([X-1][X])$ も考えられる。従って、以下本願明細書では、 Z_{ratio} と言ったときには、特段の説明がない限り、nグラム語基から「左」または「右」に一語拡張したときの双方を含むものとする。さらに、データ処理の便宜上、 Z_{ratio} の対数をとって、(式6)を定義する。

$$(式6) \quad LZ = 10 * \log(Z_{ratio})$$

[0028] (4-2) 右側拡張規則

図6のアルゴリズムに示すように、nグラム語基から右に一語拡張したときには、以下の規則を適用する(図3:356)。但し、 $[X+1]$ 、及び、 $[X+2]$ の最後の語が切れ目要素の場合には除外する。

[0029] 第一の条件

(i) $Z([X],[X+1]) > \text{Avg} Z([X],[X+1],[X+2])$ 、かつ、

(ii) $LZ > \text{第1閾値}$

を満たす場合には $[X+1]$ へ拡張する候補として選択される (610、620、650)。ここで、第1の閾値は本実施例では5.0とし、 $Z([X],[X+1])$ は $([X],[X+1])$ で表現される $(n+1)$ グラム語基のZ値、 $AvgZ([X],[X+1],[X+2])$ は、 $[X],[X+1],[X+2]$ に相当する全ての $(n+2)$ グラムのZ値の平均値である。なお、第一の条件で使用されるLZに対する第一の閾値は高く設定されている。この値が高い場合には、Zの値による判定のみでも十分に新表現として認定され得ると判断されるので、Jratio (後述) の値に関わらず、新表現の可能性のあるものとして選択する。

第一の条件、すなわち、(i) 及び(ii) の双方の条件を満足する場合には、拡張された語基の候補として選択される (650)。(i) の条件を満たさない場合には拡張する候補としては選択されない (660)。(i) の条件を満たすが、(ii) の条件を満たさない場合には、次に示す第二の条件で判別する (630、640)。

[0030] 第二の条件

(iii) $LZ > \text{第2閾値}$ 、かつ、

(iv) $Jratio = N_{jun}/N_{all} > \text{第3閾値}$

を満たす場合には $[X+1]$ へ拡張する候補として選択される (630、640、650)。

第二の条件で使用されるLZに対する第二の閾値は、実施例では3.0と設定されており、LZがこの値より大きく、かつ、Jratioが 0.1以上の値を取る時にのみ、新表現の可能性があると判定される。

ここで、Jratioとは $[X+2]$ 要素が切れ目要素である割合のことである ($0 \leq Jratio \leq 1$)。また、第3閾値は本実施例では0.1とし、 N_{jun} は切れ目要素と認定された先端要素 $[X+2]$ の数、 N_{all} は対象となった $[X+2]$ に相当する $(n+2)$ グラムの数である。

第二の条件、すなわち、(iii) 及び(iv) の双方の条件を満足する場合には、拡張された語基の候補として選択される (650)。(iii) 及び(iv) のいずれかの条件を満たさない場合には拡張された語集は選択されない (660)。

[0031] (4-3) 左側拡張規則

基本的に右側拡張規則と同様である (図3:354)。前述の(i)、(ii)、(iii)の条件は全く同じである。但し、(iv)において、切れ目要素のカウント方法が異なる。右側拡張規則では、[老][ねる]のような例に現れる[ねる]のように、着目している動詞の活用語

尾は切れ目要素と見なさない。しかし、左側拡張規則では、着目している語基の左側に存在する動詞の活用語尾が、着目している語基の新しい表現の接頭辞として用いられることは考えにくい。従ってこの場合には切れ目要素としてカウントされる。すなわち、左側では切れ目要素としてカウントされる要素が追加される。

[0032] (4-4) 右側拡張規則適用例

右側拡張規則について実例を使って説明する。語基として選択されたフルーティー(Z値は147.14)を右側に拡張することについて説明する。

語基	拡張	Z値
[X]	[X+1] [X+2]	
[フルーティー]	[さ]	5.66
[フルーティー]	[さ] [が]	2.00
[フルーティー]	[さ] [は]	2.00

ここで、注目している語基は、「フルーティー」である。まず、右に一個延ばして検討する。[フルーティー]、[さ]は前述の[X][X+1]に対応する。

[0033] この時のZ値は以下ようになる。

$$Z([X][X+1]) = Z([フルーティー][さ]) = 5.66$$

さらに右に一個延ばして([X][X+1][X+2])を検討する。ここでは2つの連語が見つかる。すなわち[フルーティー][さ][が]、及び、[フルーティー][さ][は]である。

$$[フルーティー][さ][が]のZ値 = Z([フルーティー][さ][が]) = 2.00$$

$$[フルーティー][さ][は]のZ値 = Z([フルーティー][さ][は]) = 2.00$$

ここで、[X+2]の要素、すなわち、「が」「は」をkOne要素と呼ぶ。この例のように複数のkOne要素がある場合には、これらのZ値の平均値を求める。この場合、どちらも2.00であるので、平均値は2.00となる。

すなわち、 $AvgZ([X][X+1][X+2]) = 2.00$ 、次にLZを求める。

$$Zratio = Z([X][X+1]) / AvgZ([X][X+1][X+2]) = 5.66 / 2.00 = 2.83$$

$$LZ = 10 * \log(Zratio) = 4.52 \text{ となる。}$$

[0034] 次に、このkOne要素について、切れ目を示す「切れ目要素」であるかどうかを調べる。すなわち、「フルーティーさ」という新しい表現の候補の後に、文法的切れ目を示

す要素があるかどうかをチェックする。もしあれば、その候補(「フルーティーさ」)が文法的にひとまとまりの要素として扱われていることを示唆し、新表現の候補となる。ここでは、「が」「は」共に格助詞であり、文法的切れ目を示す要素である。つまり要素(「フルーティーさ」)とつながってさらに大きな一まとまりの表現や語を作ることは考えにくい。kOne要素のうち切れ目要素である割合をJratioと呼ぶ。ここでは、2つとも切れ目要素であるから、 $Jratio = 2/2 = 1$ となる。

[0035] これらの準備をした上で、新表現としての可能性のあるものを検出していく。まず、第一の条件について検討する。

第一の条件

(i) $Z([X],[X+1]) > AvgZ([X],[X+1],[X+2])$ 、かつ、

(ii) $LZ > \text{第1閾値}$

(i) の条件は、 $Z(\text{「フルーティー」}[\text{さ}]) = 5.66$ 、及び、 $AvgZ([X][X+1][X+2]) = 2.00$ であるので満足する。

(ii) の条件は、 $LZ = 10 * \log(Zratio) = 4.52$ 、第1閾値 = 5.0となり、この条件を満足しない。従って第一の条件は満足しないので、次に第二の条件について検討する。

[0036] 第二の条件

(iii) $LZ > \text{第2閾値}$ 、かつ、

(iv) $Jratio = Njun / Nall > \text{第3の閾値}$

(iii) の条件は、 $LZ = 4.52$ 、第2の閾値は3.0であるので満足する。(iv) の条件は、 $Jratio = 2/2 = 1$ 、であり、第3の閾値は0.1であるので、満足する。

以上より、第二の条件を満足するので、[フルーティー]から[フルーティーさ]へ拡張される。ちなみに[フルーティーさ]のZ値 = $Z(\text{「フルーティー」}[\text{さ}]) = 5.66$ である。

[0037] (4-5) 左側拡張規則適用例

左側拡張規則について実例を使って説明する。語基として選択された[受け](Z値は73.01)を左側に拡張することについて説明する。

語基		拡張	Z値
[X-2]	[X-1]	[X]	
	[も]	[受け]	6.83

[に]	[も]	[受け]	2.83
	[女性]	[受け]	6.83
[、]	[女性]	[受け]	2.00
[あまり]	[女性]	[受け]	2.00

右側拡張規則の例と同様であるので左側にも拡張する。

[0038] 先ず、第一の条件について検討する。

(i) $Z([X-1],[X]) > \text{Avg } Z([X],[X-1],[X-2])$ 、かつ、

(ii) $LZ > \text{第1閾値}$

$Z([X-1][X]) = 6.83$ 、及び、 $\text{Avg } Z([X][X-1][X-2]) = 2.00$ であるので、(i)の条件は満足する。 $LZ=5.33$ 、第1閾値は5.0であるので、(ii)の条件も満足する。

以上より、[受け]から[女性受け]へ拡張される。ちなみに[女性受け]のZ値= $Z([女性受け]) = 5.33$ である。

[0039] (5) 新しい表現の選別 (図4 ステップ450)

拡張の条件に合致するものの中から、語形成規則に合致するものを新表現として選り出す (図3:360)。新しい表現を生み出す可能性の高い語は、日本語形成規則に従っていないのではなく、その形成規則は限られている (図3:365)。新しい表現として選別するためには、語法の拡張の起きている部分が名詞、動詞、形容詞、形容動詞等を形成するための規則を遵守しているか否かを確認する必要がある。図7に示すフローチャートに従って説明する。

710:名詞化規則

720:動詞化規則

730:形容詞化規則

740:形容動詞化規則

750:全ての条件を満たさない場合は候補として選別しない。

760:何れかの条件を満たす場合は候補として選別する。

以下詳細に説明する。

[0040] (5-1) 名詞化規則 (ステップ710)

名詞化形成規則に合致するものは、語基の拡張の候補として選別される。名詞化と

しては、「語基＋接尾辞」、「動詞連用形名詞化」、「複合名詞」などが挙げられる。それぞれについて、日本語としての規則を満足しているか確認する必要がある。

(a) 語基＋接尾辞

名詞以外の形容詞などを名詞化する場合は、それらの語尾に「さ」、「み」などを追加する場合がある。例として以下のものが挙げられる。

「さ」（薄さ、悲しさ、ほめられたさ）

「け」（寒け、ねむけ、吐きけ、かざりけ）

「み」（強み、いやみ、すごみ）

[0041] (b) 動詞連用形名詞化

語基の右側に格助詞・名詞をつけることによって、動詞連用形を名詞用法する場合も起こり得る。例えば、以下の様な例が挙げられる。

「走る」から「走り」、「歩き」

「遊ぶ」から「遊び」

(c) 複合名詞

複合名詞とみなされるものは、語基の拡張の候補として選別される。例えば、以下の様な例が挙げられる。

語尾に[米]を付けた場合 [掛け][米]、[麴][米]、[純][米]、[赤][米]

語尾に[香]を付けた場合 [バナナ][香]、[吟醸][香]、[熟成][香]

(d) 英語の名詞化について

本願発明は日本語だけでなく外国語にも応用できる。英語を例にとって説明する。英語で元来名詞以外の品詞として使用されているものが名詞として使用されている場合がある。例えば、以下の様な接尾辞を付加することによって名詞化されている。

「ness」: pleasantness, ugliness

「ing」: gathering

「ful」: earful

「dom」: femidom

「hood」: brotherhood, womanhood

[0042] (5-2) 動詞化規則(ステップ720)

動詞化形成規則に合致するものも、語基の拡張の候補として選別される。動詞化の例として「名詞＋する」、「動詞の一般活用形」等が考えられる。拡張の候補として選択されたものが、日本語としての規則を満足しているか確認する必要がある。

(a)「名詞＋動詞化接尾辞」の形態であるか

名詞に「する」、「ぶる」のような動詞化接尾辞、またはその活用形が結合されている場合には、語基の動詞化拡張の候補として選別される。例えば、「お茶」に「する」をつけて「お茶する」とする場合、「美人」に「ぶる」をつけて「美人ぶる」が挙げられる。

(b)動詞の一般活用形

拡張された語基が、「名詞＋動詞化接尾辞」の形態を除いた、動詞の一般活用形である場合にも語基の拡張の候補として選別される。例えば、名詞に動詞の活用語尾をつけて動詞化してしまう生産的な例として、以下のような例が挙げられる、「デモる、デモらない、デモれば」。同様に、「ゲバる、ハモる、ツモる、グーグる」といった新しい動詞をこの方法で作ることができる。

[0043] (c)英語の動詞化について

本願発明は日本語だけでなく外国語にも応用できる。英語を例にとって説明する。英語で元来名詞として使用されているものが動詞として使用されている場合がある。

Are you googling?

元来名詞である「google」が「googleを使って検索する」という動詞として使用されている例である。

I 747'ed to Chicago.

元来航空機の型番である「747」が「747航空機に乗った」という動詞として使用されている例である。

その他、以下の様な接尾辞によって動詞化されている。

「ify」: Frenchify

「en」: enliven, soften

「ize」: pluralize

[0044] (5-3)形容詞化規則(ステップ730)

形容詞化形成規則に合致するものも語基の拡張の候補として選別される。拡張の

候補として選択されたものが日本語としての規則を満足しているか確認する必要がある。

「い」(しんどい、四角い)

「こい」(ネチっこい)

「ぼい」(おんなっぼい、それっぼい)

[0045] (5-4) 形容動詞化規則(ステップ740)

形容動詞化形成規則に合致するものも語基の拡張の候補として選別される。拡張の候補として選択されたものが日本語としての規則を満足しているか確認する必要がある。

「風」(王朝風、レゲー風)

「な」(マックな[人])

「げ」(うれしげ、よさげ、なにげ)

以上のステップ710からステップ740までの何れかの条件を満足する場合には、語基の拡張の候補として選別される(760)。いずれの条件も満たさない場合には、語基の拡張の候補として選別されない(750)。

[0046] [実験結果]

以上のアルゴリズムに従って、実際のデータを用いた実験結果を示す。なお、本実験では、対象とするコミュニティとして「日本酒の味覚を議論するコミュニティ」と「ワインの味覚を議論するコミュニティ」を例として取り上げている。日本酒およびワインの銘柄名を「キーワード」として、インターネットの検索ツールを使用してそれぞれの文書集合を収集した。

[0047] (1) 名詞化

(1-1) 語基+接尾辞

形容詞を名詞化する例について説明する。ここでは形容詞「フルーティ」を名詞化し「フルーティさ」とする例について説明する。

語基	拡張	Z値
[X]	[X+1] [X+2]	
[フルーティー]	[さ]	5.66

[フルーティー] [さ] [が] 2.00

[フルーティー] [さ] [は] 2.00

[フルーティー]から[フルーティーさ]へ拡張されることは前述のとおりである。

次に、拡張された語基が名詞化形成規則(語基+接尾辞)を満足するか否か検討する。名詞以外の形容詞などを名詞化する場合は、これらの語に「さ」、「み」などを追加する。この実施例ではこの条件を満足している。

以上より、新しい語基として「フルーティー」の名詞である「フルーティーさ」が選択される。ちなみに、「フルーティー」+「さ」の判定のためのLZ値は4.52である。

[0048] (1-2) 動詞連用形名詞化

語基として選択された[受け](Z値は73.01)を左側に拡張することについて説明する。

拡張	語基	Z値
[X-2] [X-1]	[X]	
[も]	[受け]	6.83
[に] [も]	[受け]	2.83
[女性]	[受け]	6.83
[、] [女性]	[受け]	2.00
[あまり] [女性]	[受け]	2.00

[受け]から[女性受け]へ拡張されることは前述の通りである。そこで、拡張された語基が規則(動詞連用形名詞化)を満たすか否か検討する。[女性]は名詞であることは明らかである。また[受け]は後ろに格助詞が続く連語が見られ、動詞連用形による名詞化がなされていると考えられることから、[女性][受け]は動詞連用形による名詞化であると考えられるのでこの条件も満足する。

以上より、新しい語基として[女性][受け]が選択される。ちなみに、[女性][受け]の判定のためのLZ値は5.33である。

[0049] (1-3) 複合名詞

語基として選択された[雪](Z値は66.96)を左側に拡張することについて説明する。

語基	拡張	Z値
----	----	----

[X]	[X+1] [X+2]	
[雪]	[の]	4.00
[雪]	[の] [中]	2.00
[雪]	[温]	4.00
[雪]	[温] [で]	2.00
[雪]	[室]	4.00

前述の条件にあてはめて検討すると[雪]から[雪温]へ拡張されることが分かる。ここでの詳細な説明は割愛する。次に拡張された語基が名詞化形成規則(複合名詞)を満足するか否か検討する。[雪]及び[温]は名詞であることは明らかであるのでこの条件も満足する。

以上より、新しい語基として[雪温]が選別される。ちなみに、[雪温]の判定のためのLZ値は3.01である。

その他の複合名詞として拡張された例としては以下のものがある。

[米]を語基として、[掛け][米]、[麴][米]、[純][米]、[赤][米]

[香]を語基として、[バナナ][香]、[吟醸][香]、[熟成][香]

[様]を語基として、[マスカット][様]、[リンゴ][様]、[果実][様]

[度]を語基として、[アミノ酸][度]、[アルコール][度]、[日本酒][度]

[0050] (2)動詞化

(2-1)「名詞+動詞化接尾辞」

「名詞+する」の様な動詞化パターンの検出について説明する。ここでは、語基として「悪酔い」(Z値は24.01である)を選択し右側へ拡張する。

左側拡張	語基	Z値
[X-2] [X-1]	[X]	
[悪酔い]	[する]	4.00
[から] [悪酔い]	[する]	2.00
[使用]	[する]	2.00

前述の条件にあてはめて検討すると「悪酔い」を「悪酔いする」へ拡張し新しい語基とすることが出来る。ここでの詳細な説明は割愛する。

[0051] 次に、拡張された語基が動詞化規則(「名詞＋する」)を満足するか否かについて検討する。この例では、名詞に「する」または「する」の活用形が結合されているので、この条件を満たす。

以上より、新しい語基として「悪酔いする」が選別される。ちなみに、[雪温]の判定のためのLZ値は3.01である。

ここで、「悪酔いする」は普通に使用される言葉であると考えられるが、「ワインの味覚を議論するコミュニティ」と比較して、「日本酒の味覚を議論するコミュニティ」では有意差を持って出現していることが分かる。

その他の動詞化として拡張された例としては以下のものがある。

[醸造]を語基として[醸造][する]、[調和]を語基として[調和][する]、[登場]を語基として[登場][する]、[倍增]を語基として[倍增][する]

[0052] (2-2) 動詞の一般活用形

動詞が文法に従って活用する場合に、「語基＋拡張部」が一個の新しい動詞を形成する例について説明する。

例えば、日本酒コミュニティで用いられるパターンから、[老][ね](読み:ひね)、[老][ねた](読み:ひねた)、[老][ね][が、を(格助詞)](読み:ひねが、ひねを)等のデータが得られる。

語基	右側拡張	Z値
[老]	[ねる](読み:ひねる)	2.05
[老]	[ねた](読み:ひねた)	2.05

前述のアルゴリズムに従って、老ねる(読み:ひねる)(動詞一段活用形)が候補として選択される。ここで、[老](読み:おい)は、一般名詞として辞書に登録されており、動詞としては[老いる](読み:おいる)という上二段動詞が登録されている。データと動詞活用規則から、[老ねる](読み:ひねる)という下二段動詞としての拡張が起きていると判断される。また、[老][ね]+[格助詞]等のデータから、動詞連用形[老ね](読み:ひね)が名詞として用いられる名詞化が起きていることが分かる。ここから、[老ねる](読み:ひねる)がこのコミュニティにおいて新しい表現として共通の言葉として使用されている事が推測される。

図面の簡単な説明

- [0053] [図1]本願発明を実施するシステム例を示す図である。
- [図2]本願発明の一部を実施するPCのブロック図である。
- [図3]本願発明によるコミュニティ固有表現検出装置のブロック図である。
- [図4]本願発明のフローチャートである。
- [図5]本願発明の文書収集のフローチャートである。
- [図6]拡張した語基の適否を判断するフローチャートである。
- [図7]拡張した語基が語形成規則に合致しているかを判定するフローチャートである。
- 。

符号の説明

- [0054] 110:ユーザPC
- 120:サイトサーバ(1)
- 130:サイトサーバ(2)
- 140:ネットワーク
- 200:筐体
- 210:記憶装置
- 220:メインメモリー
- 230:出力装置
- 240:中央制御装置(CPU)
- 250:操作装置
- 260:ネットワークI/O

請求の範囲

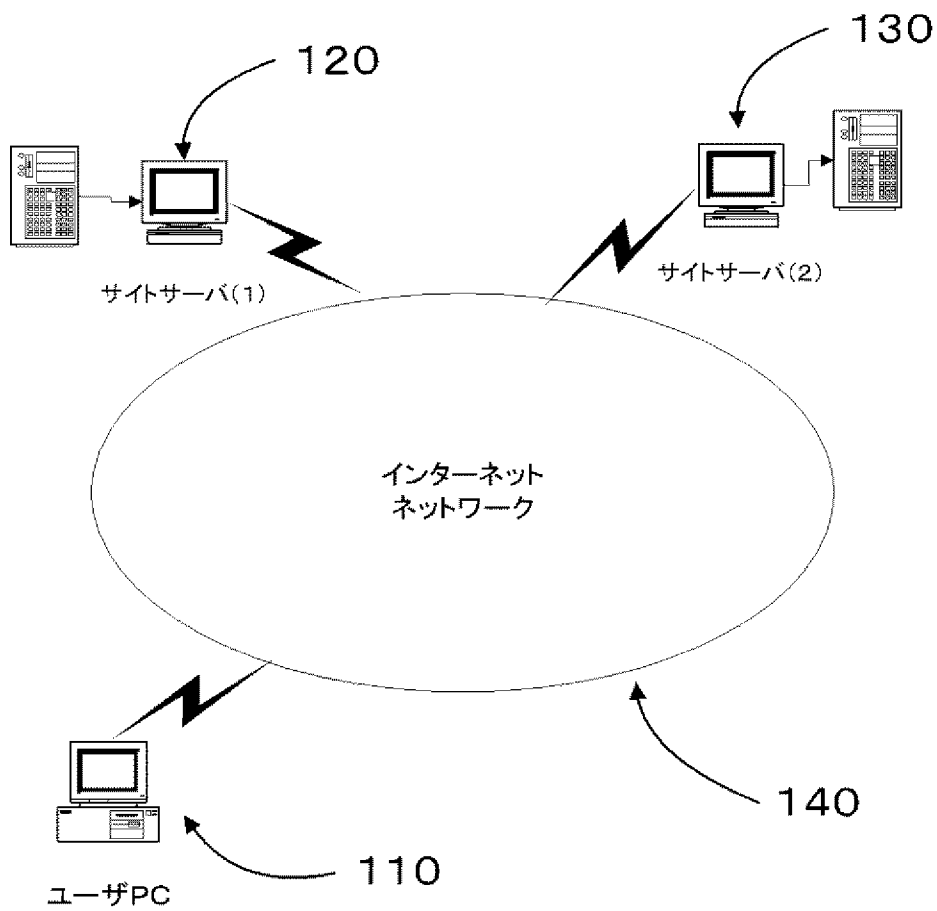
- [1] 以下の(a)から(d)の手段を有する、所定のコミュニティで使用される文書集合から前記所定のコミュニティに特有な表現を検索する装置、
- (a)前記コミュニティに特有に使用されるnグラム連語を抽出する手段、
- (b)前記特有な表現の核となる可能性のある第一の語基を選択する手段、
- (c)前記第一の語基の有意度、及び、前記第一の語基の前又は後の要素を取込んだ第二の語基の有意度を用いて算出された値に基づいて拡張語基を選択する手段、
- (d)前記拡張語基の中から当該言語の語形成規則に従って前記所定のコミュニティに特有な表現を選別する手段。
- [2] さらに、前記文書集合を、所定の用語リストに含まれる用語をキーワードとしてデータ検索することによって収集する手段を含むことを特徴とする請求項1に記載の装置。
- [3] 前記nグラム連語を抽出する手段は、複数のコミュニティで使用される文書を用い、前記所定のコミュニティで使用されるnグラム連語の有意度と、他のコミュニティで使用されるnグラム連語との有意度との比較に基づいて前記nグラム連語を抽出する手段を含むことを特徴とする請求項1及び2に記載の装置。
- [4] 前記拡張語基を選択する手段は、さらに、前記第二の語基の数、及び、前記第二の語基に取込まれた要素が切れ目要素である数を用いて算出された値に基づいて前記拡張語基を選択する手段を含むことを特徴とする請求項1及び2に記載の装置。
- [5] 前記語形成規則に従って選別する手段は、名詞化規則、動詞化規則、形容詞化規則、及び、形容動詞化規則のうち少なくとも1つの語形成規則を含むことを特徴とする請求項1及び2に記載の装置。
- [6] 以下の(a)から(d)のステップを有する、所定のコミュニティで使用される文書集合から前記所定のコミュニティに特有な表現を検索する方法、
- (a)前記コミュニティに特有に使用されるnグラム連語を抽出するステップ、
- (b)前記特有な表現の核となる可能性のある第一の語基を選択するステップ、

(c)前記第一の語基の有意度、及び、前記第一の語基の前又は後の要素を取込んだ第二の語基の有意度を用いて算出された値に基づいて拡張語基を選択するステップ、

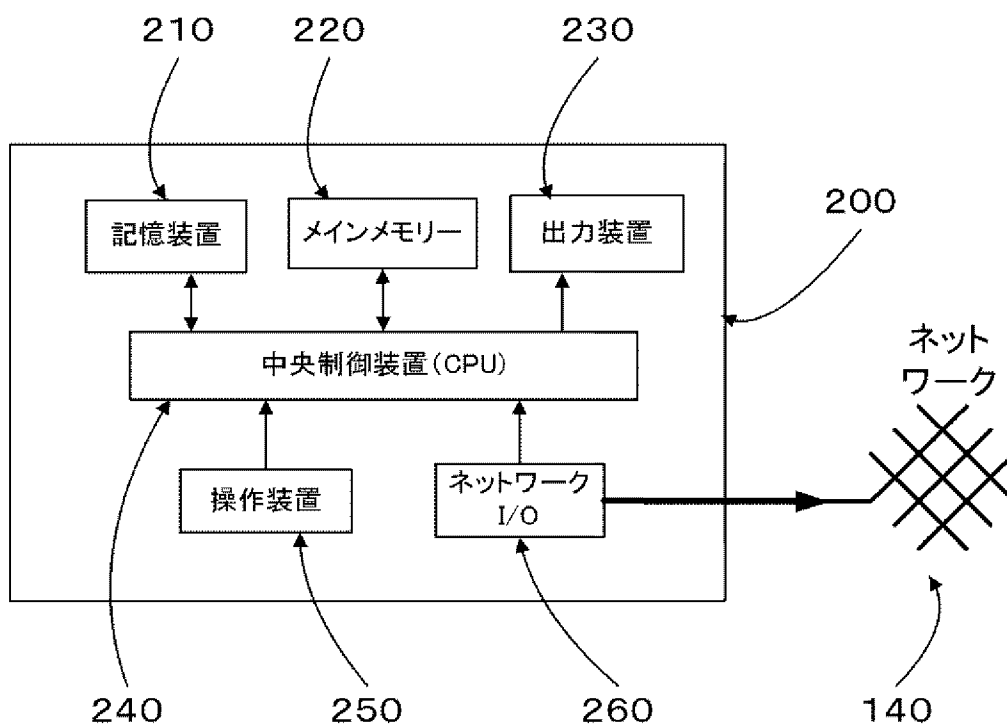
(d)前記拡張語基の中から当該言語の語形成規則に従って、前記所定のコミュニティに特有な表現を選別するステップ。

- [7] さらに、前記文書集合を、所定の用語リストに含まれる用語をキーワードとしてデータ検索することによって収集するステップを含むことを特徴とする請求項6に記載の方法。
- [8] 前記nグラム連語を抽出するステップは、複数のコミュニティで使用される文書を用い、前記所定のコミュニティで使用されるnグラム連語の有意度と、他のコミュニティで使用されるnグラム連語との有意度との比較に基づいて前記nグラム連語を抽出するステップを含むことを特徴とする請求項6及び7に記載の方法。
- [9] コンピュータを制御して、以下の(a)から(d)の手段を動作させ、所定のコミュニティで使用される文書集合から前記コミュニティに特有な表現を検索するプログラム、
- (a)前記コミュニティに特有に使用されるnグラム連語を抽出する手段、
- (b)前記特有な表現の核となる可能性のある第一の語基を選択する手段、
- (c)前記第一の語基の有意度、及び、前記第一の語基の前又は後の要素を取込んだ第二の語基の有意度を用いて算出された値に基づいて拡張語基を選択する手段、
- (d)前記拡張語基の中から当該言語の語形成規則に従って前記所定のコミュニティに特有な表現を選別する手段。
- [10] さらに、前記文書集合を、所定の用語リストに含まれる用語をキーワードとしてデータ検索することによって収集する手段を含むことを特徴とする請求項9に記載のプログラム。
- [11] 前記nグラム連語を抽出する手段は、複数のコミュニティで使用される文書を用い、前記所定のコミュニティで使用されるnグラム連語の有意度と、他のコミュニティで使用されるnグラム連語との有意度との比較に基づいて前記nグラム連語を抽出する手段を含むことを特徴とする請求項9及び10に記載のプログラム。

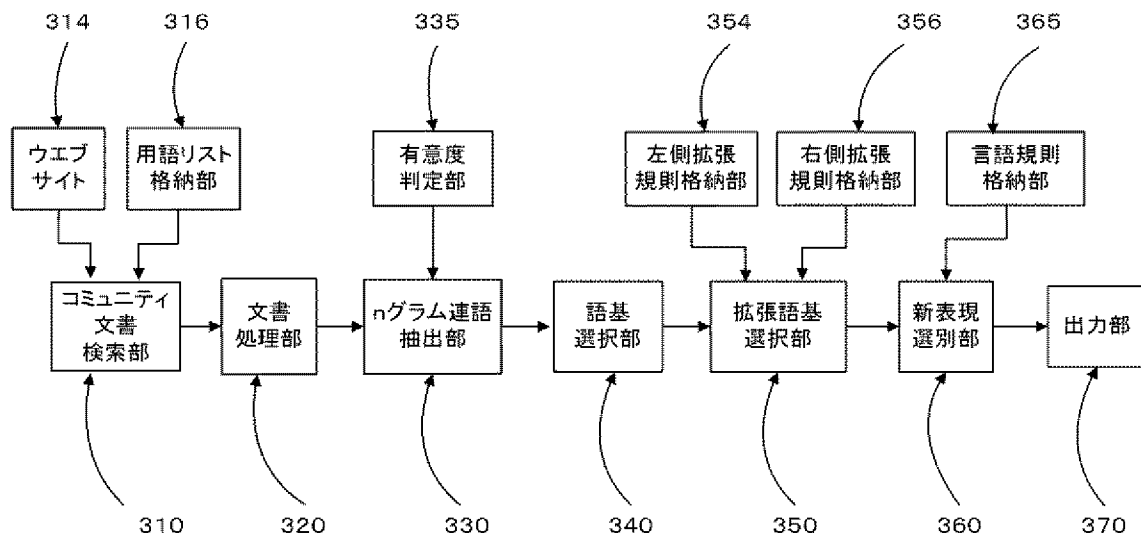
[図1]



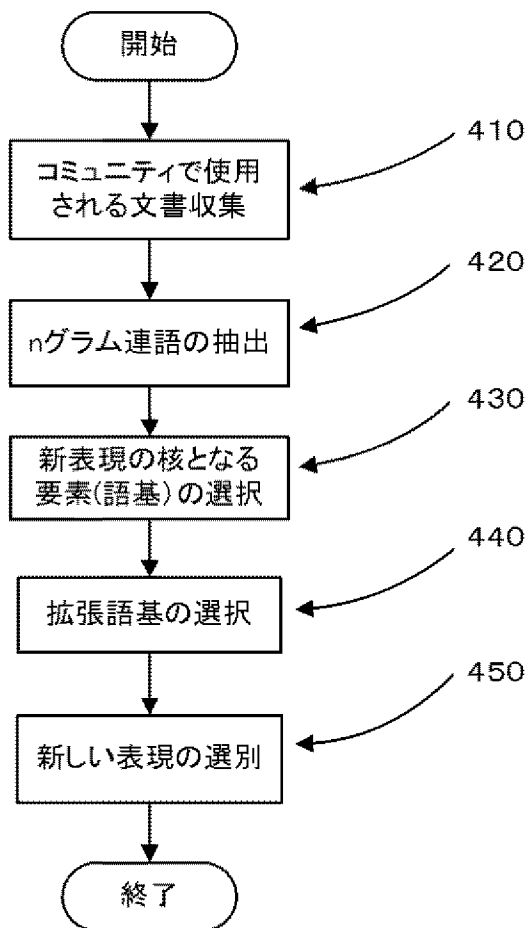
[図2]



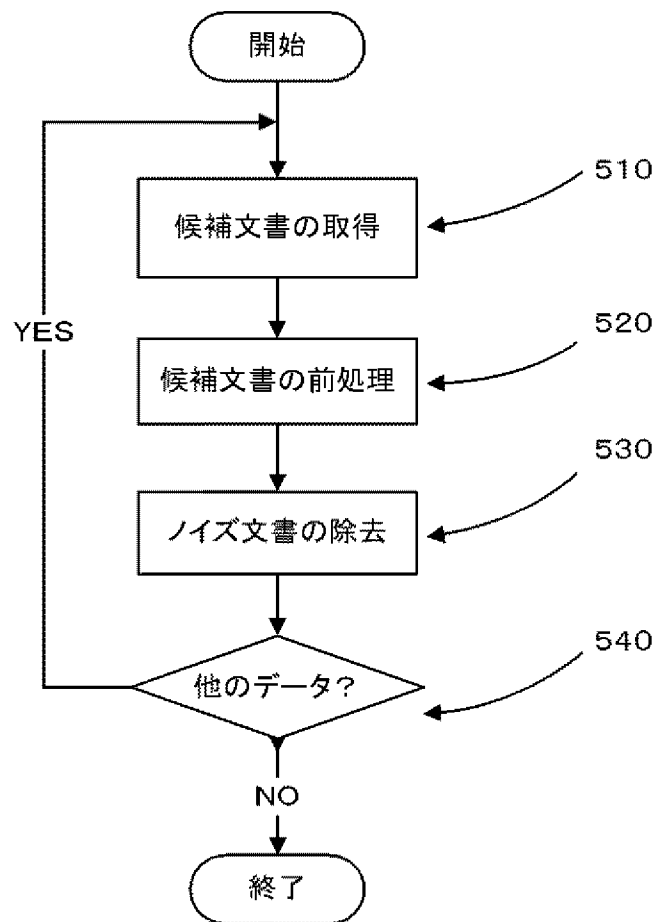
[図3]



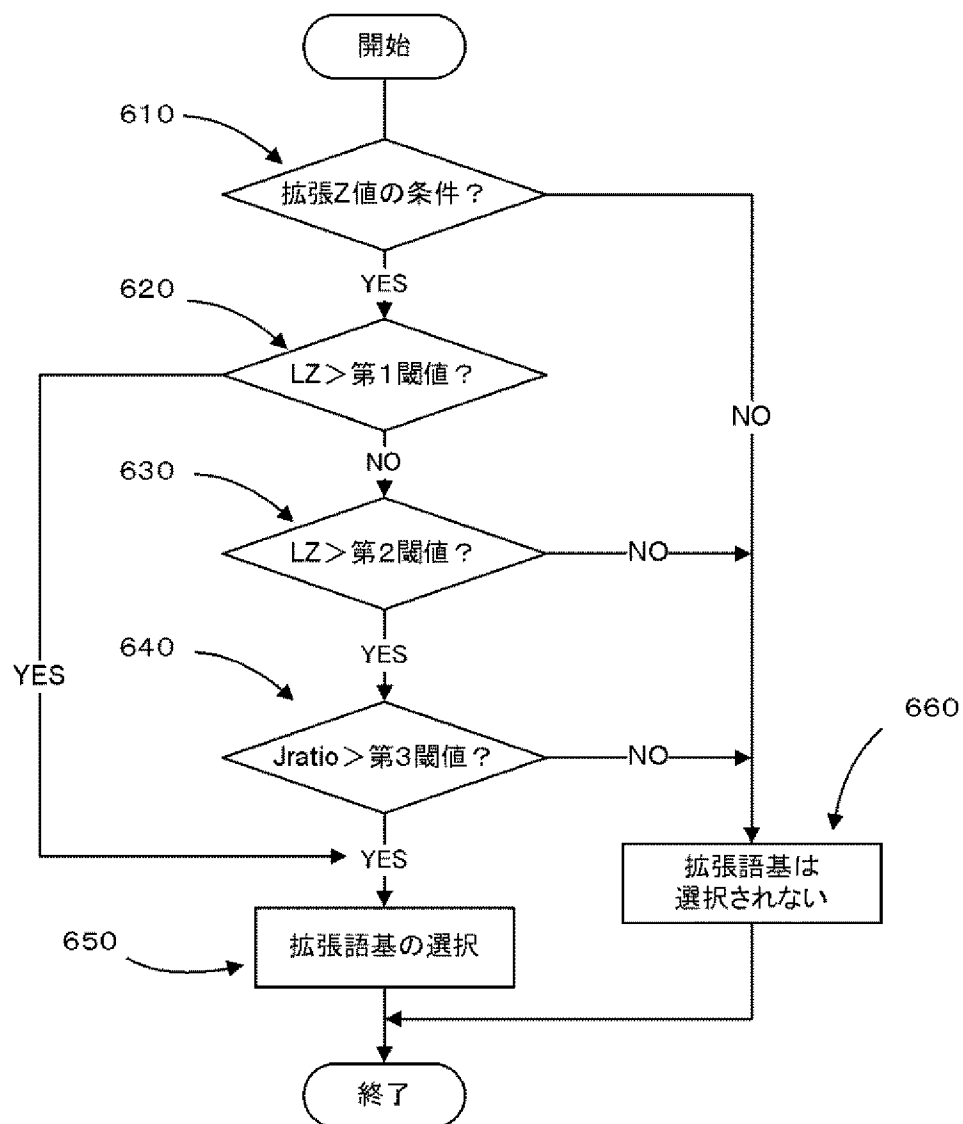
[図4]



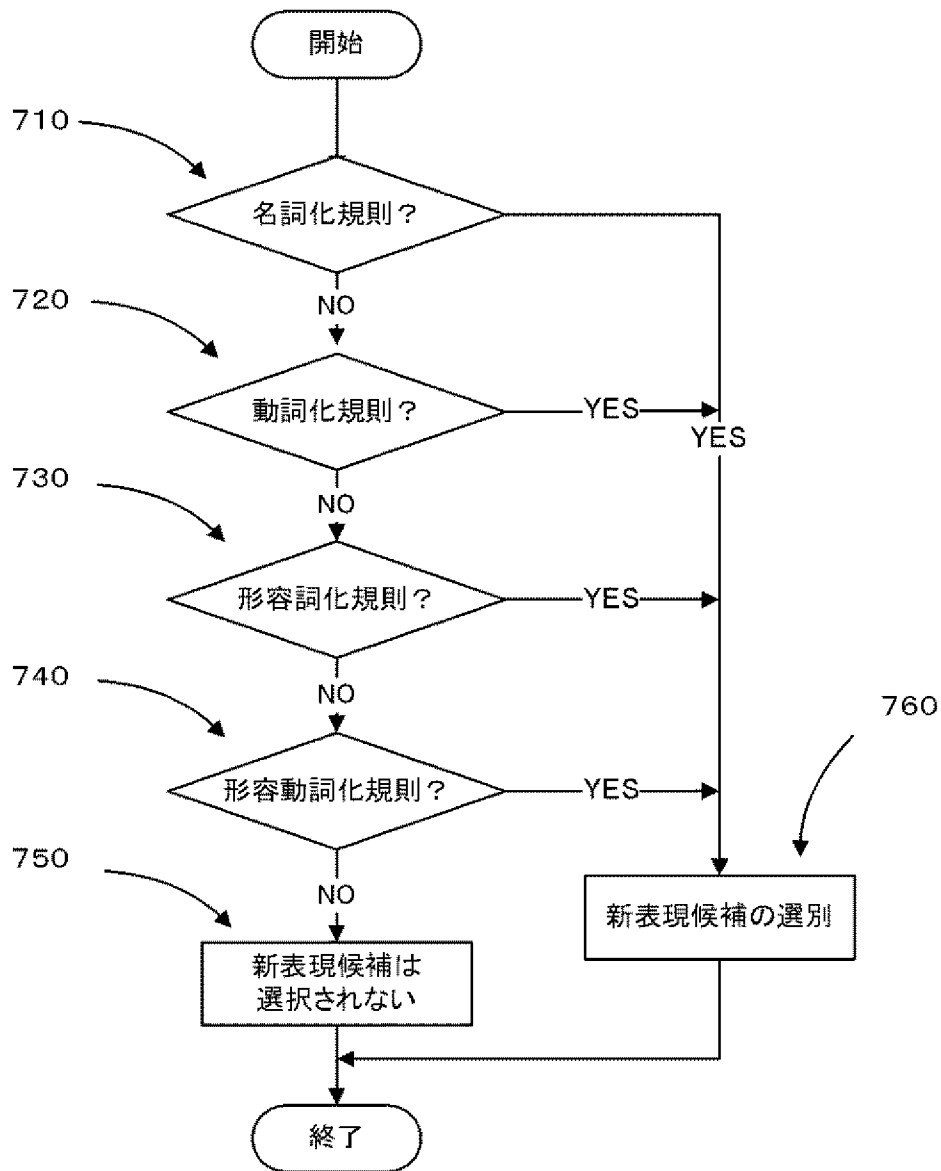
[図5]



[図6]



[図7]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2006/314000

A. CLASSIFICATION OF SUBJECT MATTER

G06F17/21(2006.01) i, G06F17/28(2006.01) i, G06F17/30(2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F17/21-17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2006
Kokai Jitsuyo Shinan Koho	1971-2006	Toroku Jitsuyo Shinan Koho	1994-2006

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Shinsuke MORI et al., "n Glam Tokei ni yoru Corpus kara no Michigo Chushutsu", IEICE Technical Report NLC 95-8, 20 July, 1995 (20.07.95), Vol.95, No.168, pages 7 to 12	1-11
Y	JP 11-85761 A (AI Soft Kabushiki Kaisha), 30 March, 1999 (30.03.99), Claims (Family: none)	1-11
Y	Maki NAGAO et al., "Daikibo Nihongo Text no n glam Tokei no Tsukurikata to Goku no Jido Chushutsu 93-NL-96-1", 09 July, 1993 (09.07.93), Vol.93, No.61, pages 1 to 8	1-11

☒ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
16 August, 2006 (16.08.06)

Date of mailing of the international search report
22 August, 2006 (22.08.06)

Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2006/314000

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	JP 2004-62262 A (Hitachi, Ltd.), 26 February, 2004 (26.02.04), Par. No. [0015] (Family: none)	2, 7, 10

A. 発明の属する分野の分類（国際特許分類（I P C））
 Int.Cl. G06F17/21(2006.01)i, G06F17/28(2006.01)i, G06F17/30(2006.01)i

B. 調査を行った分野
 調査を行った最小限資料（国際特許分類（I P C））

Int.Cl. G06F17/21-17/30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1 9 2 2 - 1 9 9 6 年
日本国公開実用新案公報	1 9 7 1 - 2 0 0 6 年
日本国実用新案登録公報	1 9 9 6 - 2 0 0 6 年
日本国登録実用新案公報	1 9 9 4 - 2 0 0 6 年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
Y	森信介他, n グラム統計によるコーパスからの未知語抽出, 電子情報通信学会技術研究報告N L C 9 5 - 8, 1 9 9 5. 0 7. 2 0, V o l. 9 5, N o. 1 6 8, p. 7 - p. 1 2	1 - 1 1
Y	J P 1 1 - 8 5 7 6 1 A (エー・アイ・ソフト株式会社) 1 9 9 9. 0 3. 3 0, 特許請求の範囲（ファミリーなし）	1 - 1 1

☒ C 欄の続きにも文献が列挙されている。

☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー	の日の後に公表された文献
「A」特に関連のある文献ではなく、一般的技術水準を示すもの	「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの	「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）	「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「O」口頭による開示、使用、展示等に言及する文献	「&」同一パテントファミリー文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願	

国際調査を完了した日
 1 6 . 0 8 . 2 0 0 6

国際調査報告の発送日
 2 2 . 0 8 . 2 0 0 6

国際調査機関の名称及びあて先
 日本国特許庁（I S A / J P）
 郵便番号100-8915
 東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）	5 M	9 4 5 9
和田 財太		
電話番号 03-3581-1101 内線 3599		

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
Y	長尾真他, 大規模日本語テキストのnグラム統計の作り方と語句の自動抽出93-NL-96-1, 1993. 07. 09, Vol. 93, No. 61, p. 1-p. 8	1-11
Y	JP 2004-62262 A (株式会社日立製作所) 2004. 02. 26, 【0015】 (ファミリーなし)	2, 7, 10