(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4904496号
(P4904496)
(45)発行日　平成24年3月28日(2012.3.28)　(24)登録日　平成24年1月20日(2012.1.20)

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

請求項の数 8　(全 27 頁)

(21)出願番号　特願2006-304301 (P2006-304301)
(22)出願日　平成18年11月9日 (2006.11.9)
(65)公開番号　特開2008-123111 (P2008-123111A)
(43)公開日　平成20年5月29日 (2008.5.29)
審査請求日　平成21年9月25日 (2009.9.25)

特許法第３０条第１項適用　２００６年５月１９日　社団法人　情報処理学会発行の「情報処理学会研究報告　２００６－ＮＬ－１７３」に発表

(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】文書類似性導出装置及びそれを用いた回答支援システム

(57)【特許請求の範囲】
【請求項１】
文からなる文書の文を形態素解析する手段と、
　形態素解析された文書から、当該文書に出現する索引語のＴＦ／ＩＤＦによる重みを要素としたＴＦ／ＩＤＦベクトルを求める手段と、
　形態素解析された文書から、当該文書に出現する体言について当該体言が出現する文中で共起した用言の頻度を要素とした共起ベクトルを求める手段とを含み、
　第１の文書のＴＦ／ＩＤＦ文書ベクトル及び共起ベクトルを求め、
　第２の文書のＴＦ／ＩＤＦ文書ベクトル及び共起ベクトルを求め、
　求めた第１の文書のＴＦ／ＩＤＦ文書ベクトル及び共起ベクトルと第２の文書のＴＦ／ＩＤＦ文書ベクトル及び共起ベクトルから第１の文書と第２の文書の類似性を求める文書類似性導出装置。
【請求項２】
文からなる文書の文章を形態素解析する手段と、
　形態素解析された文書から、当該文書に出現する索引語のＴＦ／ＩＤＦによる重みを要素としたＴＦ／ＩＤＦベクトルを求める手段と、
　形態素解析された文書から、当該文書に出現する索引語が出現する文章の文タイプを決定し、それぞれの文タイプの頻度を要素とした文タイプベクトルを求める手段とを含み、
　第１の文書のＴＦ／ＩＤＦ文書ベクトル及び文タイプベクトルを求め、
　第２の文書のＴＦ／ＩＤＦ文書ベクトル及び文タイプベクトルを求め、

求めた第１の文書のＴＦ／ＩＤＦ文書ベクトル及び文タイプベクトルと第２の文書のＴＦ／ＩＤＦ文書ベクトル及び文タイプベクトルから第１の文書と第２の文書の類似性を求める文書類似性導出装置。

【請求項３】

文からなる文書の文を形態素解析する手段と、

形態素解析された文書から、当該文書に出現する索引語のＴＦ／ＩＤＦによる重みを要素としたＴＦ／ＩＤＦベクトルを求める手段と、

形態素解析された文書から、当該文書に出現する体言について当該体言が出現する文中で共起した用言の頻度を要素とした共起ベクトルを求める手段と、

形態素解析された文書から、当該文書に出現する索引語が出現する文章の文タイプを決定し、それぞれの文タイプの頻度を要素とした文タイプベクトルを求める手段とを含み、

第１の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルを求め、

第２の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルを求め、

求めた第１の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルと第２の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルから第１の文書と第２の文書の類似性を求める文書類似性導出装置。

【請求項４】

前記請求項１ないし３のいずれかに記載の文書類似性導出装置の各手段を含み、

ＴＦ－ＩＤＦベクトルと共起ベクトル及び／又は文タイプベクトルである文書ベクトルを第１の文書について求め、

複数文書からなる第２の文書群の各文書の文書ベクトルを求め、

求めた第２の文書群の各文書の文書ベクトルから平均文書ベクトルを求め、

求めた第２の文書群の平均文書ベクトルと第１の文書の文書ベクトルから第１の文書と第２の文書群の類似性を求める文書－文書群類似性導出装置。

【請求項５】

前記請求項１ないし３のいずれかに記載の文書類似性導出装置の各手段を含み、

ＴＦ－ＩＤＦベクトルと共起ベクトル及び／又は文タイプベクトルである文書ベクトルを比較対象となる比較対象文書について求め、

第ｎの文書の索引ＴＦ－ＩＤＦの文書ベクトル、共起ベクトル及び文タイプベクトルを求め、

比較対象文書の文書ベクトルと第ｎの文書の文書ベクトルから比較対象文書と第ｎの文書の類似性を求め、

ｎは１ないしＮまであり、各第ｎの文書と比較対象文書の類似性の中から類似性の高い第ｎの文書を特定する高類似性文書特定装置。

【請求項６】

前記請求項１ないし３のいずれかに記載の文書類似性導出装置の各手段を含み、

ＴＦ－ＩＤＦベクトルと共起ベクトル及び／又は文タイプベクトルである文書ベクトルを比較対象文書について求め、

複数文書からなる第ｎの文書群の各文書の文書ベクトルを求め、

求めた第ｎの文書群の各文書の文書ベクトルから平均文書ベクトルを求め、

求めた第ｎの文書群の平均文書ベクトルと第１の文書の文書ベクトルから第１の文書と第ｎの文書群の類似性を求め、

ｎは１ないしＮまであり、各第ｎの文書群と比較対象文書の類似性の中から類似性の高い第ｎの文書群を特定する高類似性文書群特定装置。

【請求項７】

文からなる文書の文を形態素解析する手段と、

形態素解析された文書から、当該文書に出現する索引語のＴＦ／ＩＤＦによる重みを要素としたＴＦ／ＩＤＦベクトルを求める手段と、

形態素解析された文書から、当該文書に出現する体言について当該体言が出現する文中で共起した用言の頻度を要素とした共起ベクトルを求める手段と、

形態素解析された文書から、当該文書に出現する索引語が出現する文章の文タイプを決定し、それぞれの文タイプの頻度を要素とした文タイプベクトルを求める手段としてコンピュータを機能させ、
第１の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルを求め、
第２の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルを求め、
求めた第１の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルと第２の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルから第１の文書と第２の文書の類似性をコンピュータに求めさせる文書類似性導出プログラム。
【請求項８】
文からなる文書の文を形態素解析するステップと、
形態素解析された文書から、当該文書に出現する索引語のＴＦ／ＩＤＦによる重みを要素としたＴＦ／ＩＤＦベクトルを求めるステップと、
形態素解析された文書から、当該文書に出現する体言について当該体言が出現する文中で共起した用言の頻度を要素とした共起ベクトルを求めるステップと、
形態素解析された文書から、当該文書に出現する索引語が出現する文章の文タイプを決定し、それぞれの文タイプの頻度を要素とした文タイプベクトルを求めるステップとを含み、
第１の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルを求めるステップと、
第２の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルを求めるステップと、
求めた第１の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルと第２の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルから第１の文書と第２の文書の類似性を求めるステップとをさらに含む文書類似性導出方法。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、文書の類似性を求める文書類似性導出装置に関する。
【背景技術】
【０００２】
近年インターネットやパソコンの普及により、アフターサービスの一環としてパソコン技術サポートの必要性が高まっている。多くのパソコン技術サポートセンターでは、主に電話で技術サポートを行う従来型のコールセンターに加えて、インターネット経由でＥ－ｍａｉｌでの問い合わせを受け付けるメールコールセンターがたくさん設置されてきている。
【０００３】
メールコールセンターで行われている技術サポートは、すべて無料サポートである。質問メールは、夕刻から深夜にかけて多く送付されてくる。問い合わせメールの受信から回答の発信までは所定時間内（例えば２４時間以内）に完了することが求められている。このような制約があるため、企業にとって、正確かつ迅速なサポートを行うには人件費などのコストが膨大なものになりつつある。そこで、メールコールセンターの自動化が強く求められている。
そこで、問い合わせに適した回答作成を支援する回答支援装置が、特開２００１－２７３３０８号公報に開示されている。
【０００４】
この特開２００１－２７３３０８号公報の回答支援装置は、予め想定された問い合わせの内容とこの問い合わせに対する回答作成を支援する支援情報とが対応づけられて格納される支援情報記憶手段と、予め想定された問い合わせの内容とこの問い合わせに対する回

答作成者の回答作成者情報とが対応づけられて格納される回答作成者情報記憶手段と、入力される問い合わせの内容により前記支援情報記憶手段を検索して得られた支援情報および当該問い合わせの内容を、当該内容により前記回答作成者情報記憶手段を検索して得られた回答作成者情報に対応する回答作成者に送付する情報送付手段とを有するものである。

【特許文献１】特開２００１－２７３３０８号公報

【発明の開示】

【発明が解決しようとする課題】

【０００５】

前記背景技術の回答支援装置によれば、消費者からの問い合わせに対して適切な回答作成者に問い合わせが転送されると共に、回答作成者には問い合わせに係る支援情報を得ることができるため迅速に回答することができる。

【０００６】

しかしながら、この背景技術の回答支援装置は、具体的には、消費者がリストボックスやチェックボックスなどの選択形式の問い合わせに対して対応する回答作成者及び支援情報を特定するものであり、消費者が自ら作成した問い合わせ文章に対応することができないという課題を有する。なお、支援情報とは、具体的には、消費者への問い合わせに係るマニュアル、仕様書のことである。

【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、文章による問い合わせに対してその種類を特定し、適切な回答を支援する回答支援システムを提供することを目的とする。また、この回答支援システムで用いる文書間の類似性を求める新たな手法を提供することも目的とする。

【課題を解決するための手段】

【０００８】

消費者などの質問者からの質問に対して、システムが少数の回答候補を提示して、それらから回答者が最適なものを選択し、最終査読して回答を仕上げるような支援システムを作成し、使用するのが現実的である。このようなシステムでは、最終査読以外は自動的に処理されることになり、コストの大幅な削減が実現できる。発明者は鋭意努力によりシステムＡＣＣＥＳＳ（Ａｕｔｏｍａｔｅｄ　Ｃａｌｌ－ＣＥｎｔｅｒ　Ｓｅｒｖｉｃｅ　Ｓｙｓｔｅｍ）を作成した。

【０００９】

実際のメールコールセンターでの約三年間の実務により収集した三万件以上の最終査読済みの「質問応答」データの中から、約一万件を使って「質問応答データベース」を構築した。

【００１０】

実際の質問には、同一内容のものや似た内容のものが多い。したがって、ユーザからの質問が来る度に一々回答を作成するのは無駄である。そのため、既に回答した質問応答データから「質問応答データベース」を構築し、それを再利用できるようにすると、大幅なコストダウンができる。

【００１１】

「ユーザからの質問」には、「質問」とは思われないものも含まれている。このような「質問」に対しては、最終査読者が「査読」するのではなく、異なる観点からの対応が必要になるものもある。

【００１２】

ユーザからの質問は自由記述であるので、質問文にはミスタイプ、かな漢字変換の誤り、文法的不完全さなどが多く見られる。このような現象を前提として処理しなければならない。しかし、最終査読済みの質問応答データではそれらは修正・訂正されているので、質問応答データベースを検索して、うまく活用することには大きな利点がある。

【００１３】

質問応答データベースは、二層にカテゴライズした構造としている。それぞれのカテゴリにはそれぞれを特徴づけるタグを付与している。最下層のカテゴリには、実際の質問応答データが格納されている。二層にカテゴライズした理由は、最下層でうまくマッチするものがないとき、いわゆるシソーラスにおける上位概念を利用するという考えに似ている。なお、ここで二層を示したが三層以上であってもよい。

【００１４】

質問応答データベースの再利用に関しては、パソコンユーザから送られてきた問い合わせメールから、その質問がどの質問カテゴリに属するのかを統計的処理などにより推定することにより行う。推定された質問カテゴリごとのカテゴリ回答文を用いて作成した回答候補を最終査読者に提示することで、回答作成を効率化することになる。

【００１５】

質問カテゴリ推定の精度を検証するために、システムを作成し、評価実験を行った。１４５個の質問カテゴリを用いて実験したところ、８６％の割合で正解の質問カテゴリを上位３位以内に推定した。蓄積されている三万件以上の「質問応答」データの中から、約一万件を使って「質問応答データベース」を構築してみたため、残りの約二万件の「質問応答」データを「質問応答データベース」に加えることにより、正解の質問カテゴリを発見する精度は大幅に改善できる。

【００１６】

（１）文書間の類似性（図１、図２参照）

本発明に係る文書類似性導出装置は、文からなる文書の文を形態素解析する手段と、形態素解析された文書から、当該文書に出現する索引語のＴＦ／ＩＤＦによる重みを要素としたＴＦ／ＩＤＦベクトルを求める手段と、形態素解析された文書から、当該文書に出現する体言について当該体言が出現する文中で共起した用言の頻度を要素とした共起ベクトルを求める手段と、形態素解析された文書から、当該文書に出現する索引語が出現する文章の文タイプを決定し、それぞれの文タイプの頻度を要素とした文タイプベクトルを求める手段とを含み、第１の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルを求め、第２の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルを求め、求めた第１の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルと第２の文書のＴＦ／ＩＤＦ文書ベクトル、共起ベクトル及び文タイプベクトルから第１の文書と第２の文書の類似性を求めるものである。

【００１７】

このように本発明においては、ＴＦ－ＩＤＦベクトルに加え、少なくとも共起ベクトル及び文タイプベクトルの一方を用いて文書間の類似性を求めているので、より文書内の意味内容を反映した類似性を求めることができるという効果を奏する。

【００１８】

文書は一以上の文からなる。したがって、第１の文書が１つの文からなり、第２の文書が複数の文からなる場合、第１の文書が複数の文からなり、第２の文書が１つの文からなる場合、第１の文書及び第の２文書ともに複数の文からなる場合、第１の文書及び第２の文書ともに１つの文からなる場合がある。つまり、本発明の文書類似性導出装置により、文間、複数の文からなる文書間、文－複数の文からなる文書間の類似性を求めることができる。

【００１９】

共起ベクトルは、図２（上部）に示すように、文書に出現する体言について当該体言が出現する文中で共起した用言の頻度を算出し、その算出した頻度を要素として共起ベクトルを求める。図２（上部）では、例として、形態素解析により判明した「ＡＡＡ」という体言について形態素解析により判明した「ａａ」という用言が３回文書中に出現したことを算出している。

【００２０】

同様に、図２（下部）は文タイプベクトルも説示しており、文書に出現する索引語が出現する文章の文タイプを決定し、それぞれの文タイプの頻度を要素として文タイプベクト

ルを求めている。図２（下部）では、例として、形態素解析により判明した「ＡＡＡ」という体言について文タイプの決定処理を経て「ＱＵＥＳＴＩＯＮ」の文タイプが３回文書中に出現したことを算出している。

【００２１】

（２）文書群と文書の類似性（図３参照）

本発明に係る文書－文書群類似性導出装置は、前記文書類似性導出装置の各手段を含み、ＴＦ－ＩＤＦベクトルと共起ベクトル及び／又は文タイプベクトルである文書ベクトルを第１の文書について求め、複数文書からなる第２の文書群の各文書の文書ベクトルを求め、求めた第２の文書群の各文書の文書ベクトルから平均文書ベクトルを求め、求めた第２の文書群の平均文書ベクトルと第１の文書の文書ベクトルから第１の文書と第２の文書群の類似性を求めるものである。

このように本発明においては、文書間の類似性だけでなく、文書群と文書の類似性を求めることもできるという効果を有する。

【００２２】

（３）高い類似性を有した文書の特定

本発明に係る高類似性文書特定装置は、前記文書類似性導出装置の各手段を含み、ＴＦ－ＩＤＦベクトルと共起ベクトル及び／又は文タイプベクトルである文書ベクトルを比較対象となる比較対象文書について求め、第ｎの文書の索引ＴＦ－ＩＤＦの文書ベクトル、共起ベクトル及び文タイプベクトルを求め、比較対象文書の文書ベクトルと第ｎの文書の文書ベクトルから比較対象文書と第ｎの文書の類似性を求め、ｎは１ないしＮまであり、各第ｎの文書と比較対象文書の類似性の中から類似性の高い第ｎの文書を特定するものである。

このように本発明においては、複数の文書と比較対象文書の類似性を求め、高い類似性を有する文書を特定するので、比較対象文書の内容によく類似した文書を得ることができるという効果を有する。

【００２３】

（４）高い類似性を有した文書群の特定

本発明に係る高類似性文書群特定装置は、前記文書類似性導出装置の各手段を含み、ＴＦ－ＩＤＦベクトルと共起ベクトル及び／又は文タイプベクトルである文書ベクトルを比較対象文書について求め、複数文書からなる第ｎの文書群の各文書の文書ベクトルを求め、求めた第ｎの文書群の各文書の文書ベクトルから平均文書ベクトルを求め、求めた第ｎの文書群の平均文書ベクトルと第１の文書の文書ベクトルから第１の文書と第ｎの文書群の類似性を求め、ｎは１ないしＮまであり、各第ｎの文書群と比較対象文書の類似性の中から類似性の高い第ｎの文書群を特定するものである。

【００２４】

（５）回答支援システム（図４参照）

本発明に係る回答支援システムは、前記高類似性文書群特定装置を含み、前記各第ｎの文書群は類似する質問文からなり、比較対象文書も質問文であり、各第ｎの文書群の質問内容に対応する回答文を関連付けて予め記録し、前記高類似性文書群特定装置により類似性の高いとされた第ｎの文書群に関連付けられている回答文を出力するものである。

【００２５】

後説する実施形態では、この回答支援システムを具体例として示したものである。特に、実施形態では、各第ｎの文書群を複数層のツリー構造にてデータベース化している。また、高類似性文書群特定装置は、このように各第ｎの文書群内の文書が相互に類似性が高くなるように、比較対象文書が属すべき最も類似性の高い第ｎの文書群を特定することにも用いることができ、それが後説する質問応答データベース構築支援システムとなる。

【００２６】

図４は回答支援システムの発明原理図である。第１文書群から第Ｎ文書群までがあり、それぞれの文書群に対して予め平均文書ベクトルを求めて記録しておき、また、それぞれの文書群に対して共通の回答文を求めておく。そして、対象文書の文書ベクトルを求めて

、対象文書の文書ベクトルと記録している各文書群の平均文書ベクトルから対象文書と各文書との類似性を求め、最も高い類似性を有する第ｎ文書群を特定し、この第ｎ文書群の回答文を最適な回答文として使用者に出力する。なお、現在対象文書となっている文書も第ｎ文書群に振り分けられ、新しく振り分けられた文書を含めて再度第ｎ文書群の平均文書ベクトルを求める。新しく文書が振り分けられる度にしてもよいし、所定文書数蓄積された場合、所定期間毎に平均文書ベクトルを求めてもよい。同様に、新しい振り分けも所定文書数蓄積された場合、所定期間毎に実行してもよい。

【００２７】

これまで装置又はシステムとして本発明を把握してきたが、所謂当業者であれば明らかであるように、プログラム又は方法としても把握することができる。

これら前記の発明の概要は、本発明に必須となる特徴を列挙したものではなく、これら複数の特徴のサブコンビネーションも発明となり得る。

【発明を実施するための最良の形態】

【００２８】

［１．システム概要］

システムは、ユーザから質問メールを受け取ると、自動的に処理を始める。処理の結果、すなわち、回答候補は質問者への返答メールの形に整形されて最終査読者の査読を待つ。

【００２９】

質問文の解析は、形態素解析のみを行い、その結果を言語データベースおよび知識データベースに照らして、質問応答データベース検索の準備をし、自動的に検索を実施する。言語データベースは、言語的素性を持つ辞書のようなものであり、知識データベースはパソコンに関する事典のようなものである。

質問文の解析として、形態素解析のみを活用する理由は、依存構造解析などの信頼性に起因するものである。

【００３０】

最終査読者の査読済み質問応答データは、質問応答データベースの更新に供される。すなわち、質問応答データベースは査読済み回答文の返信をトリガとしてその質問応答データにより更新される。

【００３１】

システム画面の例を図５に示す。いくつかの操作機能が用意されており、最終査読者が効率的な査読を行えるよう配慮されている。最終査読者がいくつかの回答候補のいずれかが妥当であると判断し所定のボタンを押下すると、その選択された回答候補が質問者に自動的にメール返信される。微細な加工が必要な場合には、この画面の上で加工し、それが質問者に回答メールとして自動的に返信される。つまり、最終査読者は選択した回答候補の文章を適宜修正し、メールを送信することができる。

【００３２】

図５のシステム画面の構成は図６に説示する通りである。質問文入力フォームに質問文を入力し、カテゴリ判定ボタンを使用者が押下することで質問文と質問カテゴリの類似度が算出される。算出された類似度順にカテゴリのリストボックスに質問カテゴリを表示する。質問カテゴリ又は質問カテゴリに属する質問文を使用者が選択することで、テキストボックスに回答文が表示される。

【００３３】

ここで、質問文入力フォームへの質問文の入力は、例えば、メーラの本文表示からテキストデータを貼り付けることで入力する。ただし、この例に限定されない。メーラに回答支援システムの機能を具備させてもよいし、逆に、回答支援システムにメーラの機能を具備させてもよい。さらには、メールシステムを用いることなく、質問者からの質問文を他の通信方法で取得する方式を適用することもできる。例えば、ＨＴＴＰ、ＦＴＰを用いることができる。

【００３４】

［１．１　形態素解析］
文を形態素に分割して品詞を見分ける形態素解析については、自然言語処理の基礎技術の一つであり、所謂当業者であれば適宜適用が可能であるため、ここでは詳述しない。形態素解析エンジンとしては、例えば、ＭｅＣａｂ、ＣｈａＳｅｎ、ＫＡＫＡＳＩなどがある。
【００３５】
［１．２　システム構成］
回答者が使用する回答者コンピュータ１００、２００上に回答支援システムを構築する。回答者コンピュータ１００、２００に回答支援プログラムがインストールされ、回答支援システムが構築される。本実施形態では、このように一つのコンピュータにより回答支援システムが構築されているが、クライアント・サーバ型で構築することもできる。例えば、クライアントでは、ユーザからの質問文をクライアントが受けてサーバに送信し、サーバで処理されて複数の回答候補をクライアントに返信する構成である。
【００３６】
本実施形態の回答支援システムを構築したコンピュータの属するネットワーク構成の一例を図７に示す。ＬＡＮ上に回答者コンピュータ１００、回答者コンピュータ２００、サーバ３００、プリンタ（サーバ）４００及びネットワーク機器５００が接続され、相互に通信可能となっている。また、ネットワーク機器５００は外部ネットワークとも接続し、他のコンピュータとＬＡＮ上のコンピュータを通信可能としている。ここでは、質問者であるユーザからメールが送信されるとして、メールサーバが送信するメールがネットワーク機器５００を介して回答者コンピュータに送信される。回答者コンピュータが複数ある場合のメールの振り分け処理などは、コールセンタに構築されたシステムの一機能として実装され、周知・慣用技術であるためここでは詳述しない。
【００３７】
回答支援システムが構築される回答者コンピュータ１００は、例えば、ＣＰＵ(Central Processing Unit)１０１、ＲＡＭ１０２、ＲＯＭ１０３、外部記憶装置であるＨＤ(hard disk)１０４、ＣＤ－ＲＯＭからデータを読み出すＣＤ－ＲＯＭドライブ１０５、入力装置であるマウス１１１及びキーボード１１２、出力装置であるディスプレイ１２１とスピーカー１２２、並びに、ネットワークに接続するためのＬＡＮインタフェース１３１からなる構成をとる。
回答者コンピュータ１００の構成の一例を示したが、回答者コンピュータ２００、サーバ３００、ユーザコンピュータ６００も同様の構成である。
【００３８】
［２．質問応答データベース］
質問応答データベースも回答者コンピュータ１００、２００にそれぞれ構築するものとする。ここで、別途データベースサーバとして構築し、複数の回答者コンピュータが共通に使用する構成にすることもできる。
【００３９】
［２．１　質問応答データベースの構築手法］
質問応答メールデータとは、パソコンユーザから送られてきた問い合わせメールとそれに対する査読済み回答文のペアのことである。
質問応答メールデータの中には、同一データないしは類似データが多数存在する。したがって、問い合わせメールの内容または意味が同一または類似で、それらの回答文の文章表現も同じまたは類似である場合、それらを「類似データ」とみなす。
【００４０】
メールコールセンターの質問応答データベースを構築するにあたって、実際には、１０１３５件の質問応答メールデータを使用した。これらに対して、「質問カテゴリ」を作成し、類似データの「質問カテゴリ分類」を行った。質問カテゴリは、上に述べたように二層構造にし、上位層をブランチカテゴリ、下位層をリーフカテゴリと呼ぶ。類似データは同一リーフカテゴリに分類し、さらに相関関係があるリーフカテゴリは同一ブランチカテ

ゴリに分類する。
【００４１】
［２．２　質問応答データベースの構成］
構築したメールコールセンターの質問応答データベースは二段階（二層）のツリー構造である。リーフカテゴリはブランチカテゴリに属する場合もあるし、直接ルートカテゴリに属する場合もある。ルートカテゴリはブランチカテゴリの上位カテゴリであるが、ツリー構造のルートノードであるので、「層」とはみなさない。すなわち、全体を三層構造とは呼ばないことにしている。
【００４２】
ブランチカテゴリはデータを持たず、リーフカテゴリは同一データないしは類似データを持つ。ルートと各カテゴリの相関関係は下記の通りである：
・ルートカテゴリ→ブランチカテゴリ→リーフカテゴリ
・ルートカテゴリ→リーフカテゴリ
この概略を図８に示す。
【００４３】
構築したメールコールセンターの質問応答データベースでは、１０１３５件の質問応答メールデータの内、利用対象外データ３５９８件を除き、計６５３７件に対して、８３個のブランチカテゴリおよび６３４個のリーフカテゴリが設定された。利用対象外データとは、いわゆるすなおな形・内容のものではなかったものなどであり、再利用にはむかないものなどである。説示中にでてきた数字はある検証実験で得られたものである。
【００４４】
［２．３　質問応答データベース構築支援システム］
質問応答データベースの構築には、多くの工数を要する。したがって、当初は、１７０５の質問応答データについて人手で質問応答データベースを構築した。
質問応答データベース構築の効率をあげるため、その後、質問応答データベース構築支援システムを作成して活用した。上に述べた６５３７件の質問応答データは、この質問応答データベース構築支援システムを使用して構築したものである。そのスクリーンショットを図９に示す。なお、質問応答データベースは質問応答データベース構築支援システムを用いることなく、全て人手により構築してもよい。
【００４５】
［２．３．１　システム構成］
メールコールセンターの質問応答データベース構築支援システムの特徴は下記の通りである.
・分類する質問メールと既存の質問カテゴリの類似度を計算する
・操作しやすいＧＵＩインタフェースを提供する
－質問メール、質問カテゴリに既存の質問メールの内容表示および質問カテゴリの表示
－分類する質問メールに対して、類似度順で質問カテゴリの提示
－分類する質問メールに対して、属する質問カテゴリの選択・作成・削除
－質問カテゴリごとの回答文テンプレートの作成(後記参照)
【００４６】
ここでは、メールコールセンターの質問応答データベース構築支援システムのシステム構成、システム用データベース、質問カテゴリ判定システムおよびＧＵＩインタフェースについて述べる。
メールコールセンターの質問応答データベース構築支援システムの構成は図１０で示す。
メールコールセンターの質問応答データベース構築支援システムでは、分類する質問メールが下記の４つのステップを通して質問カテゴリに分類される(回答文の作成に関しては後説)。
【００４７】
１）分類する質問メールを質問カテゴリ判定システムを通して、既存の質問カテゴリと

の類似度を計算する。

【００４８】

２）ＧＵＩインタフェースで１）で計算した類似度順ですべての既存の質問カテゴリを提示する。

【００４９】

３）分類する質問メールが提示された既存のリーフカテゴリに属すると判断される場合、そのリーフカテゴリに分類する。自動的に処理することもできるが、本実施形態ではＧＵＩを介して使用者からの承認を経て分類している。具体的には２）で提示したリスト形式で表示した質問カテゴリの使用者からの指定を受け付け、さらに、分類の承認を受け付ける。

【００５０】

４）分類する質問メールが提示された既存のリーフカテゴリに属しないと判断される場合、分類する質問メールに対して、質問カテゴリの作成基準に従って、新しいリーフカテゴリまたはブランチカテゴリの作成を行う。分類する質問メールを新しく作成したリーフカテゴリに分類する。自動的に処理することもできるが、本実施形態ではＧＵＩを介して使用者からの承認を経て分類している。具体的には２）で質問カテゴリが表示されない場合、質問カテゴリが表示された場合でも適当な質問カテゴリでないときに、使用者から新しいリーフカテゴリまたはブランチカテゴリの作成の指示を受け付ける。

【００５１】

［２．３．２　システム用データベース］

メールコールセンターの質問応答データベース構築支援システムで、質問メールの分類を行う際に、分類する質問メールデータを質問カテゴリ判定システムを通して、既存の質問カテゴリとの類似度の計算を行うため、事前に、メールコールセンターの質問応答データベース構築支援システム用の質問カテゴリを用意する必要がある。

ここで、メールコールセンターの質問応答データベース構築支援システム用データベースとして、１７０５件の質問メールを利用して作成した質問カテゴリを利用する(［２．１　質問応答データベースの構築手法］を参照)。

【００５２】

［２．３．３　質問カテゴリ判定システム］

メールコールセンターの質問応答データベース構築支援システムでは、質問カテゴリ判定システムを利用して、質問メールが属するリーフカテゴリの判定を行う。

判定手法として、質問メールと質問カテゴリをベクトル空間上の点で表し、ベクトル間の類似度を定義する。

【００５３】

質問メールの文書ベクトルに関して、質問メールと質問メールが属する質問カテゴリの類似度が大きくなるように、ベクトルの要素を決定する。判定手法では、ＴＦ－ＩＤＦの重みづけによる文書ベクトルを拡張し、体言と用言の共起および文の特徴を考慮することで、質問メールの内容をより正確に反映する文書ベクトルを用いる。

【００５４】

質問メールの文書ベクトルは、下記の３種類である。

・ＴＦ－ＩＤＦによる文書ベクトル
・体言と用言の共起を考慮した文書ベクトル
・文タイプを考慮した文書ベクトル

また、質問カテゴリに属する質問メールの文書ベクトルを平均化したものを質問カテゴリの文書ベクトルとし、判定する質問メールの文書ベクトルとの重みづき余弦尺度によって、両方の類似度を求める。類似度の計算結果によって、質問メールが属する質問カテゴリを判定する。

詳細には、後記［３．特徴ベクトル］で説示する。つまり、特徴ベクトルは回答支援システムで質問文に対する適切な回答文を特定するだけでなく、質問応答データベース構築支援システムでも使用する。

【００５５】
［２．３．４　ＧＵＩ］
メールコールセンターの質問応答データベース構築支援システムをツールとしてユーザが使用する際、容易に利用できるようにシステム用のＧＵＩインタフェースを作成した。
図９で表示したボタンを押すことで、［２．３．１　システム構成］冒頭で説明したインタフェース機能を実現することが可能である。
【００５６】
初期起動時、図９のウィンドウ左側に、分類するメール、リーフカテゴリ及びブランチカテゴリのリストを表示する。
分類する質問メール一件を選択してクリックすることで、図９のウィンドウ中央に選択した質問メールの内容が表示される。
【００５７】
図９で選択した質問メールに対して、「類似度計算」ボタンを押すことで、図９のウィンドウ左側のリーフカテゴリおよびブランチカテゴリが類似度順で再表示される。類似度順は色付で表示され、色が濃いほど類似度が高いことを示す。つまり、類似度を色の濃淡で顕示している。
【００５８】
リーフカテゴリをクリックして、リーフカテゴリに分類した質問メールの一覧が表示される。ここでリーフカテゴリを選択すると、そのリーフカテゴリに属する質問メールのリストが表示され、その中の質問メールをクリックすることで、図９のウィンドウ右側に質問メールの内容が表示され、使用者は参照することが可能である。ここで、分類する質問メールがどのリーフカテゴリに属するかを判断する。属するリーフカテゴリが存在する場合、使用者が「振り分け」ボタンを押すことで、属するリーフカテゴリに分類する。属するリーフカテゴリ存在しない場合、「（新）カテゴリ作成」ボタンを押して、属するリーフカテゴリを作成する。
【００５９】
［３．カテゴリ回答文作成］
リーフカテゴリに分類された同一データないしは類似データの回答文は同じまたは類似であるため、リーフカテゴリに属する質問メールに対して、共通回答文であるリーフカテゴリ回答文を作成する。未知の問い合わせメールの回答文を作成する際、その質問メールの属するリーフカテゴリのカテゴリ回答文を用いて、回答文の作成を行う。
リーフカテゴリ回答文には、定型回答文とテンプレート回答文の２種類を用意する。定型回答文は機種関連情報などを含んでいない場合のためのものであり、回答文を作成する際にそのまま出力する。一方、テンプレート回答文は、ハードウェアやソフトウェアなどの多種類の機種関連情報に関するスロットが用意されており、それらの機種関連情報を機種関連情報データベースから抽出し、スロットに入れ、回答文を作成し出力する。
【００６０】
本章では、構築したメールコールセンターの質問応答データベースを利用し回答文の作成への応用について述べる。
メールコールセンターの質問応答データベースでは、リーフカテゴリは類似質問メールの集合である。類似質問メールとは、お問い合わせメールの内容または意味が類似して、そのお問い合わせメールに対して送信した最終査読データである回答文が類似または同じである質問メールのことを指す。
【００６１】
そのため、リーフカテゴリに属するすべての類似質問メールに対して、共通の回答文を持つと考えればよい。その共通の回答文はリーフカテゴリ回答文と定める。１個のリーフカテゴリに対して、一種類のリーフカテゴリ回答文を持つ。また、リーフカテゴリに属する類似質問メールを参照して、リーフカテゴリ回答文を作成することが可能である。
【００６２】
前説したようにリーフカテゴリ回答文は下記の２種類がある。リーフカテゴリ回答文に

ＯＳや機種などの情報によって変更する部分を含んでいない定型回答文と、リーフカテゴリ回答文にＯＳや機種などの情報によって変更する部分を含んでいるテンプレート回答文である。

次節からそれぞれのリーフカテゴリ回答文について説明を行う。

【００６３】

［３．１　回答文作成］

リーフカテゴリ回答文を利用して、お問い合わせメールの回答文を作成する方法について述べる。

お問い合わせメールの回答文を作成する流れを図１１に示す。

お問い合わせメールは質問カテゴリ判定システムを通して、属するリーフカテゴリを判定する。属するリーフカテゴリを決定すれば、リーフカテゴリ回答文を利用して回答文を作成する。

【００６４】

［３．１．１　定型回答文の利用］

お問い合わせメールが属するリーフカテゴリのリーフカテゴリ回答文が定型回答文である場合、そのリーフカテゴリ回答文をお問い合わせメールの回答文として出力する。

【００６５】

［３．１．２　テンプレート回答文の利用］

お問い合わせメールが属するリーフカテゴリのリーフカテゴリ回答文がテンプレート回答文である場合、以下の４つのステップでお問い合わせメールの回答文を作成する。

【００６６】

１）お問い合わせメールにある機種名を抽出する。

【００６７】

２）１）で抽出した機種名を利用して機種ＤＢでテンプレート回答文にあるスロットに入る機種情報をマッチングする。

【００６８】

３）２）でマッチングした機種情報をスロットに入れ換える。

【００６９】

４）回答文を出力する。

【００７０】

テンプレート回答文を利用してお問い合わせメールの回答文を作成する場合、事前に機種ＤＢを作成しておく。本研究では、メールコールセンターでサポートする全機種に対して、テンプレート回答文を作成する際に利用したスロット項目に対応する機種情報を集めて、機種ＤＢを作成した。

ここでは、機器名、機器情報について説示したが、当然他の情報であってもよい。

【００７１】

［４．特徴ベクトル］

これより、分類済みの質問応答データベースを用いて、新たに入力として与えられた質問文がどのカテゴリに属するのかを判定する手法について説示する。

本実施形態では、質問文とカテゴリをベクトル空間上の点で表す。また、ベクトル間の類似度を定義する。質問文と、その質問文が属するカテゴリとの類似度が大きくなるようにベクトルの要素を決定し、類似度を定義することで、類似度によって質問文のカテゴリを推定するものである。

【００７２】

本手法では、ＴＦ－ＩＤＦの重みづけによる文書ベクトルを拡張し、体言・用言の共起と、文の特徴を考慮することで、質問文の内容をより正確に反映した文書ベクトルを用いる。

つまり、以下の特徴ベクトルを複合的に用いる。

・ＴＦ(Term Frequency)／ＩＤＦ(Inverse Document Frequency)による特徴ベクトル
・体言と用言の共起を考慮した特徴ベクトル

・文タイプを考慮した特徴ベクトル

【００７３】

また、カテゴリに属する質問文の文書ベクトルを平均化したものをカテゴリの文書ベクトルとし、重み付きの余弦尺度によって類似度を求める。余弦尺度は、ふたつのベクトルの類似度を、ベクトルがなす角の余弦によって考えるもので、同じベクトル同士はそのなす角が０で余弦は１となり、完全に異なる要素を持つベクトル同士は直交して余弦は０になるというものである。ベクトル同士のなす角θの余弦は以下の式で表せる。

【００７４】

【数１】

$$\cos\vartheta = \frac{\vec{v}\cdot\vec{v'}}{|v|\cdot|v'|} \qquad \cdots\ (1)$$

$(\vec{v}\cdot\vec{v'}$は内積$)$

ベクトルが正規化済みであるならば、これはベクトル同士の内積に等しい。以降、基本的にベクトルはすべて正規化済みであると仮定する。つまり、余弦尺度は内積によって求められる。

【００７５】

本実施形態においては、文書を特徴付ける語として、名詞(未知語を含む)、動詞、形容詞などの自立語の原型と品詞情報の組を用いることとする。語の解析は、日本語係り受け解析器ＣａｂｏＣｈａと、形態素解析器ＭｅＣａｂを用いて、形態素解析と、文節区切りまでを行った。半角と全角の同じ文字や、アルファベットの大文字小文字などを区別しないようにあらかじめ前処理した文を形態素解析器にかけ、連続する数字・アルファベット・記号はつなげて名詞とした。ただし、連続する名詞を複合名詞とすることは、学習データが不十分になると考え、行っていない。数詞と助数詞の連続については、数詞を実際の数字の並びではなく、数クラスに置き換える処理を行った。なお、括弧内の文で、二文節以上の文は、括弧内の文であるという情報は保持しつつ、別の一文として切り離して扱うようにしている。

【００７６】

［４．１　ＴＦ／ＩＤＦによる特徴ベクトル］

システムでは、問い合わせメール中に出現する語のＴＦ／ＩＤＦによる重みを要素とした文書ベクトルを拡張したもので質問文を表現する。

ＴＦ－ＩＤＦ重み付けはテキストの自動索引づけにおいて、索引語の重みを計算する手法である。ＴＦ（Ｔｅｒｍ　Ｆｒｅｑｕｅｎｃｙ）とは、ある文書ｄにおける索引語ｔの生起頻度であり、ｔｆ（ｄ，ｔ）と表記する。またＩＤＦ（Ｉｎｖｅｒｓｅ　Ｄｏｃｕｍｅｎｔ　Ｆｒｅｑｕｅｎｃｙ）は文書の数Ｎと、索引語ｔが一回以上生起する文書の数ｄｆｒｅｑ（ｔ）によって次のように定義される。

【００７７】

【数２】

$$idf(t) = \log\frac{N}{dfreq(t)} + 1 \qquad \cdots\ (2)$$

索引語ｔの文書ｄにおける重みｗ（ｔ，ｄ）として、ＴＦとＩＤＦの積をもちいるのがＴＦ－ＩＤＦ重み付けである。重み付けにＴＦを用いるのは、文書中で繰り返し生起する語はその文書において重要な概念であると考えるためである。しかし、多くの文書に生起

する語は、文書を特定する性質を持たず、索引語として適していない。そこで、語がどのくらい特定性を持つかをＩＤＦによって重み付けに反映させている。

【００７８】

［４．２　体言と用言の共起を考慮した特徴ベクトル］

ＴＦ／ＩＤＦによる重み付けは、通常、ある語が特定の文書を特徴付ける尺度を表現するものであり、文の構造を反映しない。したがって、

・「電源を切る。」

・「電源を入れる。」

という二つの文に対して、「電源」という語は同じ重みが与えられる。だが実際には、目的とする質問文のカテゴリ判定においては、この二つは違う特徴を持つものとして認識すべきである。これは、語の出現頻度だけを考えていては、とらえにくい特徴である。そこで、ＴＦ／ＩＤＦによる重み付けに加えて、体言に対する用言の一文での共起の度合を重みとして用いることを考える。それぞれの体言について、一文中で共起した用言の頻度を要素とする特徴ベクトルを用いる。文書ベクトルの要素として、ＴＦ－ＩＤＦ重みと一緒に保持しておく。これにより、ふたつの語を比べた際に、共起ベクトルの余弦尺度による類似度を用いることを考える。

通常、ＴＦ－ＩＤＦのみによる文書ベクトルＶとＶ´の類似度ｓｉｍ（Ｖ、Ｖ´）は、余弦尺度、つまり内積によって求める。全文書中の語の数、すなわち文書ベクトルの次元をｎとすると、以下のように表される。

【００７９】

【数３】

$$sim(\vec{V},\vec{V'}) = \vec{V}\cdot\vec{V'} = \sum_{i=1}^{n}\left\{\omega_i\cdot\omega_i'\right\} \quad \cdots \quad (3)$$

$(\omega_i\cdot\omega_i'$はTF-IDFによる重み$)$

ここで、共起ベクトルの類似度を重みに加える。要素にTF-IDFによる重みと体言・用言の共起ベクトルを持つ文書ベクトル$Ｖ_c$、$Ｖ´_c$の類似度ｓｉｍ（Ｖ、Ｖ´）を、以下のように定義する。

【００８０】

【数４】

$$sim(\vec{V_c},\vec{V_c'}) = \sum_{i=1}^{n}\left\{(\omega_i\cdot\omega_i')(\vec{c_i}\cdot\vec{c_i'})\right\} \quad \cdots \quad (4)$$

$(\vec{c_i}\cdot\vec{c_i'}$は共起ベクトル$)$

上式は、ある語ｉについて、ＴＦ－ＩＤＦの重みが大きいほど、また、語ｉに同じような共起の傾向があるほど、文書ベクトルの類似度が高くなる。上記の「電源」の例の場合、それぞれに共起している用言は「切る」、「入れる」であるので、共起ベクトルの類似度は０である。したがって文書ベクトル全体の類似度も０となり、ふたつの文は似ていないと判断される。

【００８１】

［４．３　文タイプを考慮した特徴ベクトル］

パソコンユーザから送られてきた問い合わせメールの内容をより正確に反映した特徴ベクトルを作成するため、文中の語がどのような意味の文に出現するのか、という傾向につ

いて考える。そのために、まず問い合わせメールを分析してそれぞれの文タイプごとの特徴を調べ、分析結果をもとに文タイプ同定のルールを作成する。

【００８２】

［４．３．１　質問文の分析］

ここでは、質問メールを分析することにより、次のように少数の文タイプを設定した。

・Question：「～できますか?」「～を教えて下さい」など、質問を述べてある文。

・Problem：「～ができません」「～する方法がわかりません」など、問題を述べてある文。

・Intention：「～したい」「～しようと思う」など、質問者の意図・希望が述べてある文。

・Situation：問題発生の手順・状況などについて述べてある文。

・Think：「～だと思います」など、質問者の考えが述べてある文。

・Other case：「HDDでの再生は問題ありません」など、別の状況では問題が発生しない場合が述べてある文。

・About ：「～について」などの、質問内容を端的に表している文。質問、回答の一行目に述べられることがある.

・Message：エラーメッセージや、ダイアログなど、画面に表示された文字列の内容を述べてある文。

・etc：その他の情報

【００８３】

［４．３．２　質問文の分析結果］

上記の文タイプを集計した結果を以下に示す。

・Question　324
・Problem　648
・Intention　87
・Situation　398
・Think　37
・Other case　80
・About　368
・Message　96
・etc　34

【００８４】

質問について述べてある文や、パソコンの不具合・問題について述べてある文など、上記９種類の文タイプを設定し、約一週間分の問い合わせメール、３２３件２０７２文を分析して、文末表現や機能語から、文タイプを同定するルールを作成し、各文タイプ中での各単語の頻度を要素とする特徴ベクトルを用いる。上の表の右端の数字は、参考までに記したものであり、それぞれの文タイプについて、分析の際に出現した回数である。

【００８５】

ほとんどの質問には、QUESTIONかPROBLEMのどちらかが含まれ、どちらも出現しない場合は３２３件中に３件だけであった. その3件中のすべてにINTENTIONが含まれていた。さらに、それぞれのタイプについての分析を以下に示す。

・Question：ほとんどの場合文末が記号「?」か助詞「か」、あるいは「教えてください」「ご教示ください」「お願いします」などで終わる。その他のタイプはほとんどマッチしない。

・Problem：文末が自立の動詞・形容詞の基本形や、「～できません」「～しない」「～してしまう」などで終わる場合が全体の３分の２を占める。また、ＱＵＥＳＴＩＯＮ文の直前に多く出現する。

・Intention：ほとんどの場合、文末が「～したい」「～ほしい」「～しようと思っています」などで終わる。

・Situation：「～しました」のような過去形で終わる場合が多いが、そうでない場合も

多くある。PROBLEM文の直前に多く出現する。
・Think ：「～かと」を含むか、文末が「思う」「気がする」などで終わる。
・Other case：「～は」「～では」「～も」「～と」「～だと」などを含む文で、文末が「できる」「異常ない」「問題ない」「正常です」「発生しない」などで終わる。
・About ：質問、回答の一行目において、文末が名詞で終わる。
・Message：文の全部、あるいは一部が「」や''で括られていることが多い。その直後に「という」「と、」「って」などの語がつき、「表示されました」「出ました」「メッセージが出ました」などの文が続く。
・etc：「初心者です」「名前は～です」などの情報がある。これらについては、あらかじめ対応ルールを用意しておくのが難しく、また出現頻度も少ないため、今回は対応を見送ることにする。

【００８６】

［４．３．３　文タイプ同定ルール］

上記の分析結果をもとに、文タイプの同定ルールを作成した。 ルールは三段階に分けて適用される。まず、最初に適用するルールについて述べる。以下のそれぞれにあてはまる文に、重複を許してタイプを割り振っていく。

上記の分析結果をもとに、文タイプの同定ルールを作成した。ルールは三段階に分けて適用される。まず、最初に適用するルールについて述べる。以下のそれぞれにあてはまる文に、重複を許してタイプを割り振っていく。

【００８７】

・Question：文末が「が?」以外の疑問符で終わる．あるいは、文末が助詞「か」で終わる。あるいは、文末の5文節以内に「教えて」「教示」「教授」「お知らせ」「なぜ」「願い」を含む。
・Problem：文末が自立の動詞・形容詞の基本形で終わる。あるいは、文末が「でした」「が」でなく、格助詞「が」を含む文のうち、格助詞「が」と文末の間に他の助詞を含まない。あるいは、文末の3文節に「すみません」「すいません」「していません」「しておりません」を含まず、「なくなっています」「なくなった」「なくなり」「てしまった」「なります」「なりました」「まいました」「まいます」「ません」「ない」「しまう」「れる」「れます」を含む。
・Intention：文末の５文節に「(動詞)+たい」「ほしい」「(動詞)+(よ)うと」を含み、その後に動詞の「思う」「考える」が続く。
・Think ：文中に助詞の並び「かと」を含む、あるいは、文末の3文節に「思う」「思った」「思われ」「考えられ」「気がする」「気がします」を含む。
・Other case：助詞、あるいは助詞の並び「は」「では」「も」「と」「だと」を含む文で、文末が「できる」「動く」「作動(する)」「動作(する)」「起動(する)」の活用のうち、「基本形」「た」「ます」「ている」で終わるか、「異常」「問題」「不都合」の後に「ありません」「なかった」「ない」が続いて終わる。あるいは、助詞「は」「と」の後に、「正常に」「正しく」「普通に」「通常」「きちんと」「うまく」「ちゃんと」を含む文がくる。
・About ：質問、回答の一行目において、文末が名詞で終わる。
・Message：助詞「と」を含む文で、以降に「メッセージ」「ボックス」「ポップアップ」「表示」「エラー」動詞「出る」が出現する。助詞「と」の直前に、「」、（）、'、``、で括られた部分がある場合、複数の文にまたがっている場合でも、括弧などで括られた内部を全て``MESSAGE''と判断する。

【００８８】

次に、なにもタイプが割り振られなかった文に対して、SITUATIONか、PROBLEMを割り振る。次の三種類の場合を考える。
・質問文中にQUESTIONもPROBLEMも出現していない場合。ABOUTが出現している場合は、ABOUT中の語を含む文をPROBLEMとする。ABOUT中の語を含む文がない場合や、ABOUTが出現していない場合、タイプが割り振られていない一番最初の文をPROBLEMとする。残りはSITUA

TIONとする。
・質問文中にQUESTIONが出現している場合。QUESTIONの直前の文にタイプが割り振られていない場合、PROBLEMとする。残りはSITUATIONとする。
・それ以外の場合。タイプが割り振られていない文をすべてSITUATIONとする。
【００８９】
次に、重複した文タイプに対して、タイプ間の優先順位にもとづいたルールを適用してタイプを確定する。
文タイプが重複している文は、以下の優先順位で文タイプを決定する。
``MESSAGE''＞``ABOUT''＞``QUESTION''＞``PROBLEM''＞``OTHERCASE''＞``INTENTION''＞``THINK''
【００９０】
以上のルールを適用して、文タイプを決定する。質問文中に出現する語は、どのような文タイプ中で何度出現するのかという情報をベクトルとして持つことになる。
求めた文タイプを利用して、式（４）を次のように拡張する。要素にTF-IDFによる重み、体言・用言の共起ベクトル、文タイプベクトルを持つ文書ベクトルV_t、V'_tの類似度$sim(V_t, V'_t)$を、以下のように定義する。
【００９１】
【数５】

$$sim(\vec{V_t},\vec{V'_t}) = \sum_{i=1}^{n}\{(\omega_i\cdot\omega'_i)(\vec{c_i}\cdot\vec{c'_i})(\vec{t_i}\cdot\vec{t'_i})\} \qquad (\vec{t_i},\vec{t'_i}\text{は文タイプベクトル})$$

$$sim(\vec{V_t},\vec{V'_t}) = \sum_{i=1}^{n}\left\{(\omega_i\cdot\omega'_i)(\vec{c_i}\cdot\vec{c'_i})(\vec{t_i}\cdot\vec{t'_i})\right\} \quad \cdots \quad (5)$$

$(\vec{t_i}\cdot\vec{t'_i}$は文タイプベクトル$)$

式（５）は、語ｉが同じような文タイプに出現する傾向がある場合、文書ベクトルの類似度が大きくなることを表している。質問文が、どのようなことについて述べているのか、という傾向が似ているものを、類似度が高い、と評価する。
【００９２】
［４．４　カテゴリの平均ベクトルを用いた類似度計算］
未知の問い合わせメールがどの質問カテゴリに属するのかを計算するのに、各質問カテゴリ内の質問文の特徴ベクトルを平均化したものを便宜的に質問カテゴリのベクトルとする。その概念図を図１２に示す。
これらに対して、未知の問い合わせメールのベクトルとの類似度を計算する。これは、カテゴリ内の質問文の文書ベクトルを平均化することで、少数のノイズを取り除き、カテゴリ内で真に特徴的な語の情報のみを残すことができるからである。
【００９３】
【数６】

$$\vec{A} = \frac{\sum_{i=1}^{n}\vec{a_i}}{n} \cdots (6)$$

$\vec{A}$:n個の正規化前の質問文の文書ベクトル$\vec{a_1}\ldots\vec{a_n}$を持つカテゴリAの平均文書ベクトル

そして、質問ベクトルと各質問カテゴリとの距離を計算し、最も近い質問カテゴリをもつカテゴリに質問が属すると判断する。

この平均ベクトルを正規化したものと、未知の質問文の文書ベクトルとの類似度の計算結果を利用して、回答作成支援システムを作成する。

【００９４】

［４．５　具体例］

［４．５．１　文書ベクトル］

Ｗ：単語空間

Ｗ∋ｉ：ある体言と対応している

例：ｉ＝ＰＣ，ｉ’＝電源．．．

ここにおいて、ある文書ベクトルＶ内の単語ｉをＴＦ－ＩＤＦにおいて重み付けした値をＴＦ－ＩＤＦ（ｉ）と表すものとする。

【００９５】

この時、例として下に挙げる文章１、２における文章ベクトルは文中の体言ｉを軸とし

ｗ：ＴＦ－ＩＤＦ（ｉ）

ｃ：｛ｖ：ＴＦ－ＩＤＦ（ｖ），ｖ’：ＴＦ－ＩＤＦ（ｖ’），．．｝ｖ，ｖ’：体言ｉと文中で共起する用言、ｃ：ｖを軸として持つベクトル

ｔ：｛文タイプ１：文タイプ１中でのｉの出現回数，．．｝ｔ：文タイプを軸として持つベクトル

以上の３つの値をセットそして持ち、そのベクトルの要素ｗを正規化したものとする。

【００９６】

●文章１

買ってきたばかりのPCの電源が入りません。

何をしたらいいのでしょうか？

［文章ベクトルＶ］

｛ｉ１：［ｗ＝ＴＦ－ＩＤＦ（ＰＣ），ｃ＝｛買う：ＴＦ－ＩＤＦ（買う），入る：ＴＦ－ＩＤＦ（入る）｝，ｔ＝｛否定：１｝］

ｉ２：［ｗ＝ＴＦ－ＩＤＦ（電源），ｃ＝｛買う：ＴＦ－ＩＤＦ（買う），入る：ＴＦ－ＩＤＦ（入る）｝，ｔ＝｛否定：１｝］

ｉ３：［ｗ＝ＴＦ－ＩＤＦ（何），ｃ＝｛する：ＴＦ－ＩＤＦ（する）｝，ｔ＝｛疑問：１｝］｝＊ｉ１＝ＰＣ，ｉ２＝電源，ｉ３＝何

【００９７】

●文章２

PCが起動しないのですが、どうしたらよろしいですか？

［文章ベクトルＶ’］

｛ｉ１：［ｗ＝ＴＦ－ＩＤＦ（ＰＣ），ｃ＝｛ＴＦ－ＩＤＦ（起動）｝］，ｔ＝｛疑問：１｝｝｝＊ｉ１＝ＰＣ

【００９８】

［４．５．２　類似度］

文章ベクトルＶ、Ｖ’の類似度を計算しようとする時、ベクトル空間の次元数は（Ｖ∪Ｖ’）の単語空間の次元数に等しい。

よって、例におけるＶ■は軸としてＰＣしか持たない１次元のベクトル空間であるがここでは（Ｖ∪Ｖ’）の単語空間に拡張する。

また共起ベクトルｃ、文タイプベクトルｔも同様に拡張した文章ベクトルＶ’をＶ’’と表すとそれは以下のようになる。

【００９９】

[文章ベクトルＶ''］

｛ｉ１：［ｗ＝ＴＦ－ＩＤＦ（ＰＣ），ｃ＝｛買う：ＴＦ－ＩＤＦ（買う），入る：ＴＦ－ＩＤＦ（入る），起動：ＴＦ－ＩＤＦ（起動），する：ＴＦ－ＩＤＦ（する）｝，ｔ＝｛疑問：１，否定：０｝］

ｉ２：［ｗ＝ＴＦ－ＩＤＦ（電源），ｃ＝｛買う：ＴＦ－ＩＤＦ（買う），入る：ＴＦ－ＩＤＦ（入る），起動：ＴＦ－ＩＤＦ（起動），する：ＴＦ－ＩＤＦ（する）｝，ｔ＝｛疑問：０，否定：０｝］
ｉ３：［ｗ＝ＴＦ－ＩＤＦ（何），ｃ＝｛買う：ＴＦ－ＩＤＦ（買う），入る：ＴＦ－ＩＤＦ（入る），起動：ＴＦ－ＩＤＦ（起動），する：ＴＦ－ＩＤＦ（する）｝，ｔ＝｛疑問：０，否定：０｝］｝＊ｉ１＝ＰＣ，ｉ２＝電源，ｉ３＝何

ここにおいて類似度の計算は以下の式（５）に従う。また、ここにおける(t・t')は文タイプが一致すれば１一致しないならば０を返すものである。

【０１００】

［５．動作］

図１３は本実施形態に係る回答支援システムのブロック構成であり、図１４は本実施形態に係る回答支援システムの動作フローチャートである。なお、図１３に示したブロック構成は一例であり、所謂当業者で明らかであるように複数のモジュール構成をとることができる。そして、ここでは、動作主体を明示しているが、ハードウェア的視点から言えば、コンピュータ、プロセッサが動作主体である。

【０１０１】

質問文が内包された質問メールを質問者がユーザコンピュータ６００で作成し、回答者コンピュータ１００のアドレス宛に送信する。

回答者コンピュータ１００は複数のメールサーバを介してユーザコンピュータ６００からの質問メールをメーラで受信する。なお、回答者コンピュータ１００が直接アクセスするメールサーバが所定メールアドレスのメールを、登録された回答者コンピュータへ適宜振り分けする機能を有する構成であってもよい。

【０１０２】

使用者はメーラで受信した質問メールを本回答支援システムへ取り込む指示を行う。

入力部１は使用者から指示を受け付け、指示された質問メールを取り込む（Ｓ１００）。

前処理部２は全角（半角）文字変換やアルファベットの大文字（小文字）変換などの前処理を実行する。

形態素解析部３は前処理後の質問文を形態素解析する（Ｓ２００）。

【０１０３】

文書ベクトル作成部４のＴＦ－ＩＤＦ文書ベクトル部４１、共起ベクトル部４２及び文タイプ文書ベクトル部４３はそれぞれ取り込んだ質問メールの本文の各ベクトルを求める。

類似性算出部６は、各質問カテゴリの平均文書ベクトルを読み出し、この読み出した平均文書ベクトルと求めた質問メールの各ベクトルから式（５）を用いて類似度を求める（Ｓ３００、Ｓ４００）。

類似性算出部６が各質問カテゴリとの類似度を求めた後に、出力部８は各質問カテゴリを読み出し、類似度順に質問カテゴリをリスト表示する（Ｓ５００）。

【０１０４】

使用者からの質問カテゴリの選択を受け付け、回答文書特定部７が質問カテゴリの識別情報から質問カテゴリと関連付いて記録している回答文を読み出し、出力部８がそれを表示する（Ｓ６００）。

【０１０５】

出力部８が表示している回答文への修正を受け付ける。

使用者からの承認を受け付けると、メーラを介して回答文が質問者に返信される（Ｓ７００）。

【０１０６】

さらに、質問応答データベース構築支援システムを介して今回送信されてきた質問文、その回答文及び平均文書ベクトルが記録される（Ｓ８００）。ここで、既に質問カテゴリも決定されており、使用者から質問カテゴリの選択を受け付けることなく迅速に記録処理

がなされる。

【実施例】

【０１０７】

カテゴリ判定の精度を確かめるために、３種類の実験を行い、結果の評価する。

（カテゴリ判定実験）

実験データとして、分類済みの質問・回答データのうち、１カテゴリに３件以上の質問文を持つ６２９カテゴリをデータＡとして用いる。また、１カテゴリに１２件以上の質問文を持つ１４５カテゴリをデータＢとして用いる。データＡの総データ数は６５３６件で、８３個の上位カテゴリを持つ。データＢの総データ数は４０２３件で、５２個の上位カテゴリを持つ。これらのデータに対し、データを３分割してそのうちふたつを学習データとして用い、残りをテストデータとして３回テストを行った結果の平均をとる３分割交差検定を行い、質問文の正解カテゴリと、正解カテゴリの上位カテゴリを何位に判定したかを調べた。

質問文に対して、正解カテゴリと、正解の上位カテゴリをどれだけ上位に判定したかを評価とする。

【０１０８】

【表１】

データ	一位に判定	三位以内	上位カテゴリを一位に判定	上位カテゴリを三位以内
A(6536 件)	3043(46.6%)	4562(69.8%)	4463(68.3%)	5798(88.7%)
B(4023 件)	2505(62.3%)	3459(86.0%)	3202(79.6%)	3805(94.6%)

カテゴリ判定の実験結果である。

データＡでの判定結果は、データＢに比べると悪い。その理由として、学習データの不足が考えられる。データＡのカテゴリ数は６２９個であるが、その大半が１カテゴリ内に３個か４個程度のデータしか持っていない。学習データの数に比べて、カテゴリ数が非常に多いため、判定ミスが増加したものと考えられる。

【０１０９】

（ＴＦ－ＩＤＦによる重み付けと、提案手法との比較実験）

判定実験での実験データＢを用いて、単純なＴＦ－ＩＤＦによる重み付けだけを用いる文書ベクトルと、提案手法である、体言・用言の共起と、文タイプを考慮した文書ベクトルの類似度による判定精度を比較した。

質問文に対して、正解カテゴリと、正解の上位カテゴリをどれだけ上位に判定したかを評価とする。

【０１１０】

【表２】

実験データ B(総データ数 4023 件) による結果

文書ベクトル	一位に判定	三位以内	上位カテゴリを一位に判定	上位カテゴリを三位以内
TF-IDF	2343(58.2%)	3342(83.1%)	3148(78.3%)	3768(93.7%)
提案手法	2505(62.3%)	3459(86.0%)	3202(79.6%)	3805(94.6%)

【０１１１】

ＴＦ－ＩＤＦと提案手法の文書ベクトルによる比較実験の結果である。

提案手法の方が、若干精度が良いが、改善率としては一割程度である。その理由として、ベクトルを平均化した結果、カテゴリの文書ベクトルが非常に特徴的になってしまった。カテゴリの文書ベクトル同士の類似度を計算したところ、ほとんど０に近い値ばかりになった。特徴的な語があると、それに強く反応してしまい、他の要素があまり考慮されていない。

【０１１２】

（ＫＮＮ法と平均ベクトル法によるカテゴリ判定の比較実験）

判定実験での実験データＢを用いて、ＫＮＮ法によるカテゴリ判定と、平均ベクトル法によるカテゴリ判定の精度を比較した。平均ベクトル法では、未知の質問文の文書ベクトルを入力として、カテゴリの平均ベクトルとの類似度を用いてカテゴリを判定したが、ＫＮＮ法では入力ベクトルとすべての学習データ内の文書ベクトルとの類似度を求め、類似度が高い方からｋ個の文書ベクトルが属するカテゴリから、入力ベクトルの属するカテゴリを判定する。

【０１１３】

データＢでは、学習データ内のすべてのカテゴリが最低８個のデータを持っていることが保証されているので、ｋの値は８とした。

類似度の計算は、提案手法である体言・用言の共起と文タイプを考慮した文書ベクトルを用いて計算した。

質問文に対して、正解カテゴリと、正解の上位カテゴリをどれだけ上位に判定したかを評価とする。

【０１１４】

【表３】

実験データ B(総データ数 4023 件) による結果

	一位に判定	三位以内	上位カテゴリを一位に判定	上位カテゴリを三位以内
KNN 法	2402(59.7%)	3194(79.4%)	3243(80.6%)	3768(92.7%)
平均ベクトル法	2505(62.3%)	3459(86.0%)	3202(79.6%)	3805(94.6%)

【０１１５】

ＫＮＮ法と平均ベクトル法の比較実験の結果である。

上位カテゴリの一位判定において、ＫＮＮ法のほうがわずかに高い値を出しているものの、全体的には、特に三位以内での判定において、平均ベクトル法のほうが良い精度を出している。

ＫＮＮ法の判定ミスの理由として、類似度の高いｋ個のデータの中に、正解のカテゴリに属するデータがひとつもない場合が４０２３件中５８３件もあることが挙げられ、データの分布がかなりの範囲で重なっている。

【０１１６】

以上の前記各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】

【０１１７】

【図１】発明の原理図（文書の類似性）である。

【図２】発明の原理図（共起ベクトル、文タイプベクトル）である。

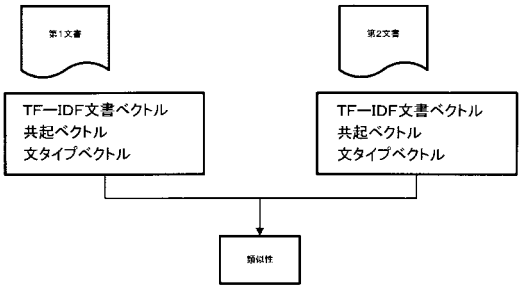
【図３】発明の原理図（文書－文書群の類似性）である。

【図４】発明の原理図（回答支援）である。

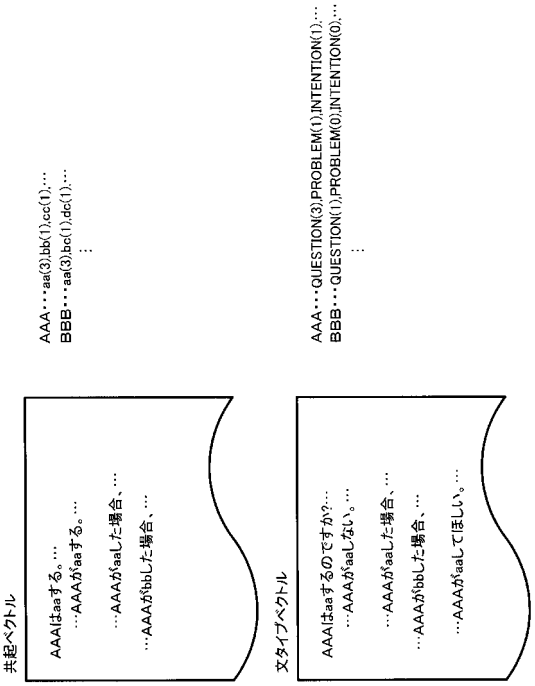
【図５】本発明の実施形態に係るシステム画面である。

【図６】図５のシステム画面の構成である。
【図７】本発明の実施形態に係る回答支援システムを構築したコンピュータの属するネットワーク構成の一例である。
【図８】本発明の実施形態に係る質問応答データベースのツリー構造である。
【図９】本発明の実施形態に係る質問応答データベース構築支援システムのスクリーンショットである。
【図１０】本発明の実施形態に係るメールコールセンターの質問応答データベース構築支援システムの構成である。
【図１１】本発明の実施形態に係る質問メールの回答文を作成する動作フローチャートである。
【図１２】本発明の実施形態に係る各質問カテゴリ内のベクトル概念図である。
【図１３】本発明の実施形態に係る回答支援システムのブロック構成である。
【図１４】本発明の実施形態に係る回答支援システムの動作フローチャートである。
【符号の説明】
【０１１８】
１　入力部
２　前処理部
３　形態素解析部
４　文書ベクトル作成部
４１　ＴＦ－ＩＤＦ文書ベクトル部
４２　共起ベクトル部
４３　文タイプ文書ベクトル部
５　文書ベクトル記憶部
６　類似性算出部
７　回答文書特定部
８　出力部
１００　回答者コンピュータ
１０１　ＣＰＵ
１０２　ＲＡＭ
１０３　ＲＯＭ
１０４　ＨＤ
１０５　ＣＤ－ＲＯＭドライブ
１１１　マウス
１１２　キーボード
１２１　ディスプレイ
１２２　スピーカー
１３１　ＬＡＮインタフェース
２００　回答者コンピュータ
３００　サーバ
４００　プリンタ
５００　ネットワーク機器
６００　ユーザコンピュータ

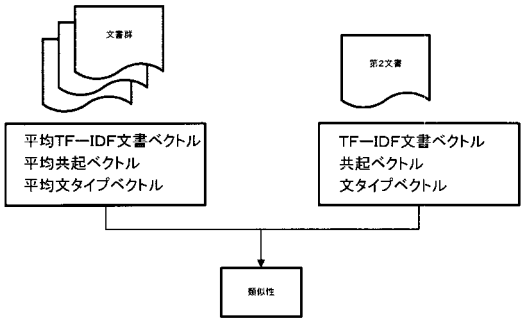
【図１】


第1文書
第2文書
TF−IDF文書ベクトル
共起ベクトル
文タイプベクトル
TF−IDF文書ベクトル
共起ベクトル
文タイプベクトル
類似性

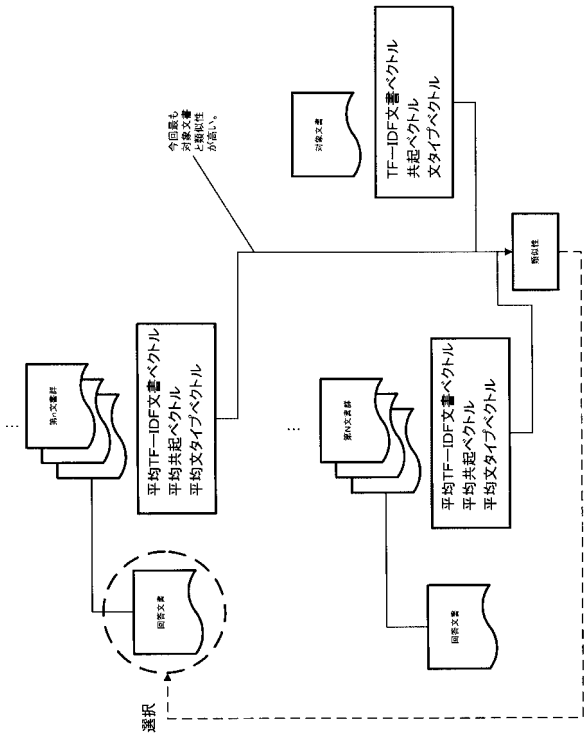
【図２】


AAA・・・aa(3),bb(1),cc(1),…
BBB・・・aa(3),bc(1),dc(1),…
…
AAA・・・QUESTION(3),PROBLEM(1),INTENTION(1),…
BBB・・・QUESTION(1),PROBLEM(0),INTENTION(0),…
…
共起ベクトル
AAAはaaする。…
…AAAがaaする。…
…AAAがaaLた場合、…
…AAAがbbLた場合、…
文タイプベクトル
AAAはaaするのですか?…
…AAAがaaLない。…
…AAAがaaLた場合、…
…AAAがbbLた場合、…
…AAAがaaLてほしい。…

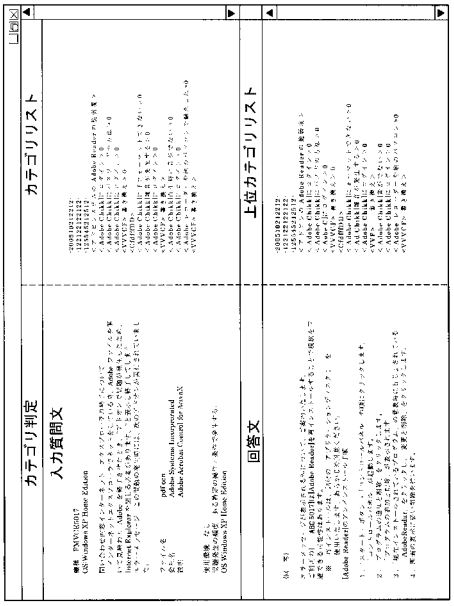
【図３】


文書群
第2文書
平均TF−IDF文書ベクトル
平均共起ベクトル
平均文タイプベクトル
TF−IDF文書ベクトル
共起ベクトル
文タイプベクトル
類似性

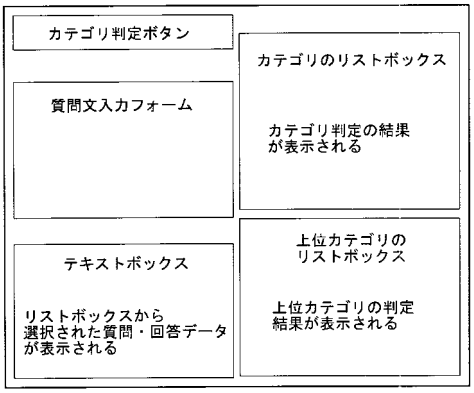
【図４】


今回長も対象文書と類似性が高い。
対象文書
TF−IDF文書ベクトル
共起ベクトル
文タイプベクトル
類似性
…
第n文書群
平均TF−IDF文書ベクトル
平均共起ベクトル
平均文タイプベクトル
回答文書
…
第N文書群
平均TF−IDF文書ベクトル
平均共起ベクトル
平均文タイプベクトル
回答文書
選択

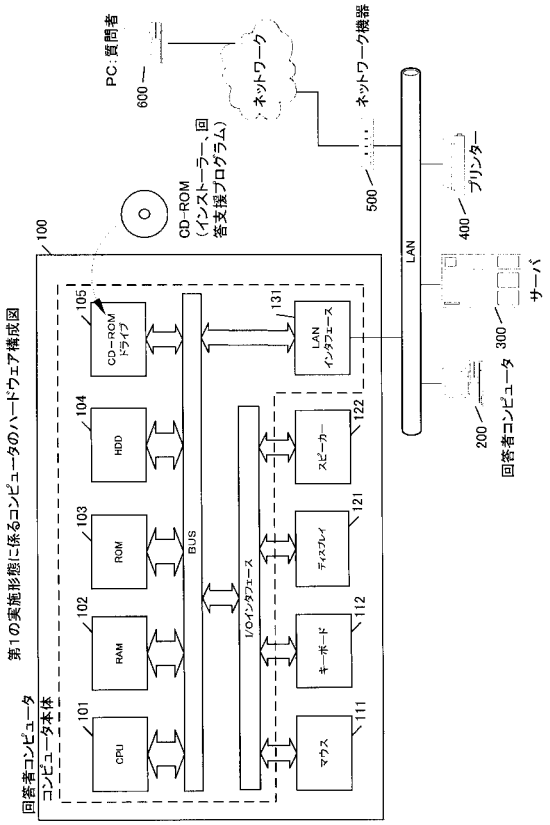
【図５】


カテゴリ判定
カテゴリリスト
入力質問文
上位カテゴリリスト
回答文

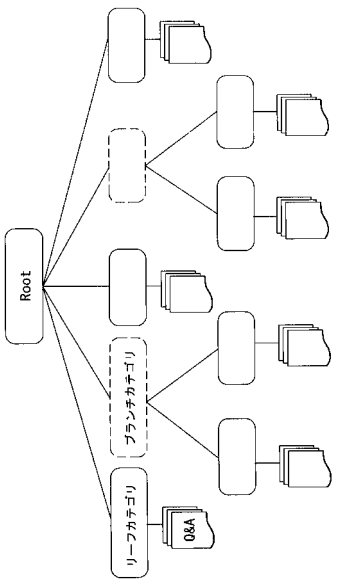
【図６】


カテゴリ判定ボタン
カテゴリのリストボックス
カテゴリ判定の結果が表示される
質問文入力フォーム
テキストボックス
リストボックスから選択された質問・回答データが表示される
上位カテゴリのリストボックス
上位カテゴリの判定結果が表示される

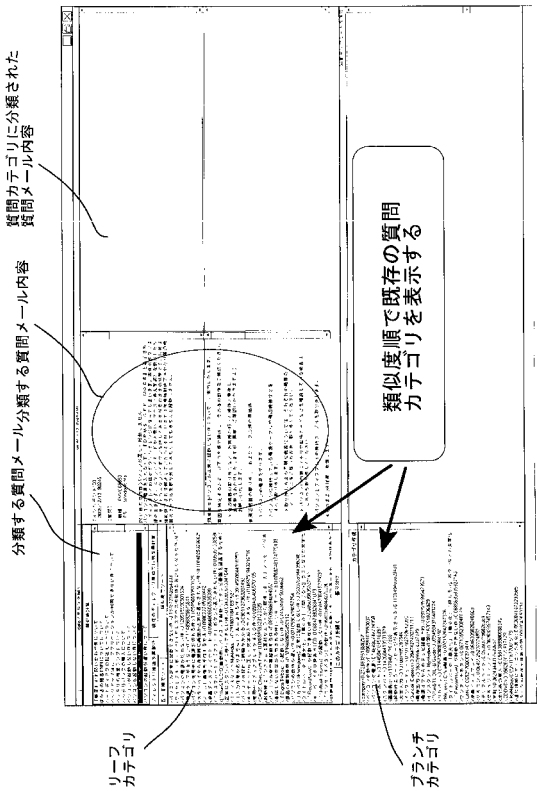
【図７】


第1の実施形態に係るコンピュータのハードウェア構成図
回答者コンピュータ
コンピュータ本体
100
CPU
101
RAM
102
ROM
103
HDD
104
CD-ROMドライブ
105
BUS
I/Oインタフェース
マウス
111
キーボード
112
ディスプレイ
121
スピーカー
122
LANインタフェース
131
CD-ROM
（インストーラー、回答支援プログラム）
PC：質問者
600
ネットワーク
ネットワーク機器
500
LAN
プリンター
400
サーバ
300
回答者コンピュータ
200

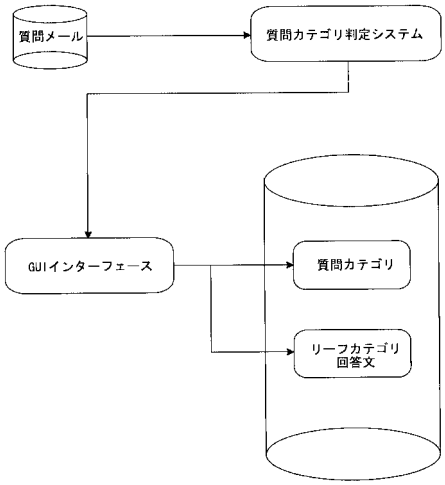
【図８】


Root
ブランチカテゴリ
リーフカテゴリ
Q&A

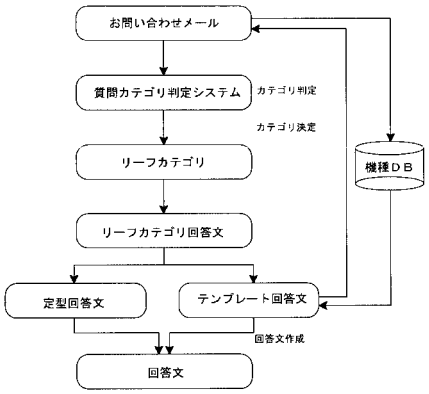
【図９】



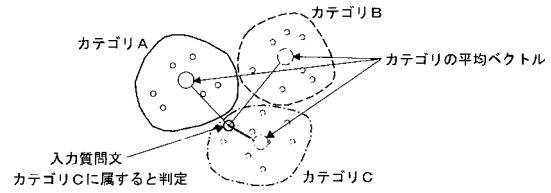
【図１０】



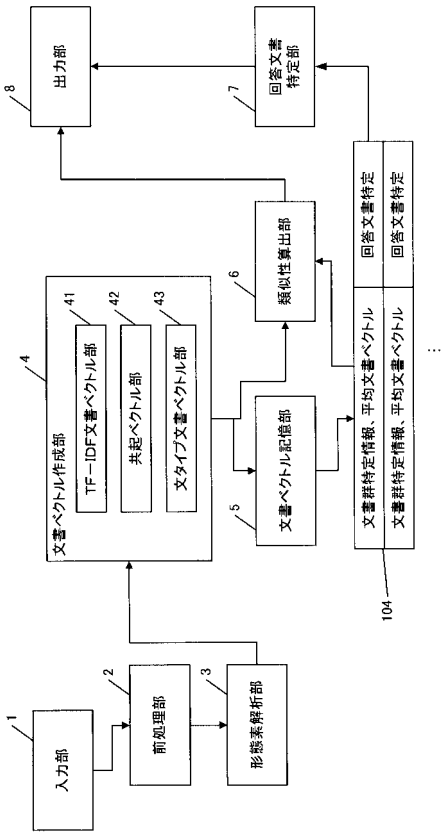
【図１１】



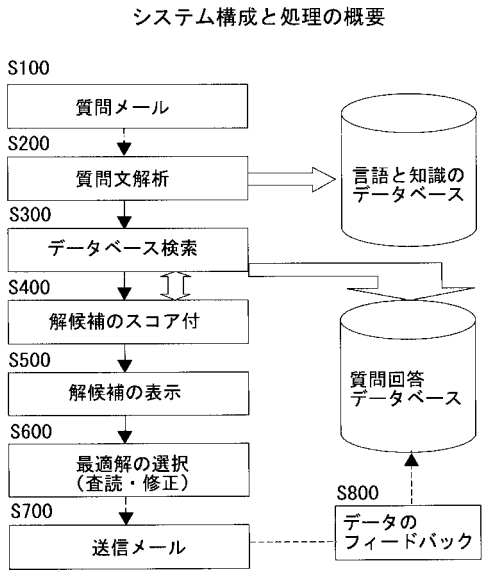
【図１２】



【図１３】



【図１４】


システム構成と処理の概要
S100
質問メール
S200
質問文解析
言語と知識の
データベース
S300
データベース検索
S400
解候補のスコア付
S500
解候補の表示
質問回答
データベース
S600
最適解の選択
（査読・修正）
S800
データの
フィードバック
S700
送信メール

フロントページの続き

(56)参考文献 特開２００２－２４５０６７（ＪＰ，Ａ）
特開２００１－３３１５１５（ＪＰ，Ａ）
特開平１０－０７８９７１（ＪＰ，Ａ）
特開平０６－１１０９２９（ＪＰ，Ａ）
大森信行、外３名，ｔｆ・ｉｄｆ法を用いた関連マニュアル群のハイパーテキスト化，情報処理学会研究報告（９７－ＮＬ－１２１），日本，社団法人情報処理学会，１９９７年 9月１２日，第９７巻，第８５号，ｐ．１１１－１１８
岩崎礼次郎、外１名，コールセンターにおける対話データを用いた営業日報の自動生成，第２２回 ことば工学研究会資料（ＳＩＧ－ＬＳＥ－Ａ５０３），日本，社団法人人工知能学会 ことば工学事務局，２００６年 3月１０日，ｐ．８７－９４
岡村潤、外３名，複数マニュアルの自動ハイパーテキスト化における類似度計算手法について，情報処理学会研究報告（９８－ＦＩ－５１），日本，社団法人情報処理学会，１９９８年 9月１８日，第９８巻，第８１号，ｐ．７１－７８
高木徹、外１名，単語出現共起関係を用いた文書重要度付与の検討，情報処理学会研究報告（９６－ＦＩ－４１），日本，社団法人情報処理学会，１９９６年 4月１８日，第９６巻，第３４号，ｐ．６１－６８

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １７／３０