(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5213098号
(P5213098)
(45)発行日　平成25年6月19日(2013.6.19)　(24)登録日　平成25年3月8日(2013.3.8)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 330C
G06F 17/30 170A
G06F 17/30 210D

請求項の数 8　外国語出願　(全 12 頁)

(21)出願番号　特願2007-165692(P2007-165692)
(22)出願日　平成19年6月22日(2007.6.22)
(65)公開番号　特開2009-3814(P2009-3814A)
(43)公開日　平成21年1月8日(2009.1.8)
審査請求日　平成22年6月9日(2010.6.9)

(73)特許権者　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100120868
弁理士　安彦　元
(72)発明者　呉　友政
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　柏岡　秀紀
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

審査官　打出　義尚

最終頁に続く

(54)【発明の名称】質問応答方法及びシステム

(57)【特許請求の範囲】
【請求項１】
ユーザから入力された質問文を構成する単語から、キーワードと、質問の種別を規定するアンサータイプとを特定する質問文解析ステップと、
上記質問文解析ステップにおいて抽出したキーワードを検索クエリーとし、当該キーワードに関連するウェブページを検索する検索ステップと、
上記検索ステップにおいて検索した各ウェブページから、上記アンサータイプに基づいて解答候補を順次抽出する解答候補抽出ステップと、
上記解答候補抽出ステップにおいて抽出した各解答候補に基づいて、候補選択スニペットをクラスタに割り当て、これをトレーニングデータとする、クラスタリングステップと、
上記トレーニングデータを解析することにより上記クラスタを分類し、更に上記トレーニングデータの解析と同一解析条件の下で上記質問文を解析し、当該質問文の解析結果と最も類似するクラスタを順に抽出し、これを応答として出力する分類ステップとを有し、
上記分類ステップは、上記トレーニングデータと上記質問文との単語重複度を示すＳＢＦＳ（similarity-based feature set）、上記トレーニングデータと上記質問文とのブーリアン重複度を示すＢＭＦＳ(Boolean match-based feature set)、上記トレーニングデータを構成する解答候補の前後を構成する文字を含めた文字列と上記質問文との類似度を示すＷＷＦＳ(window-based word feature set)に基づいて、当該質問文の解析結果と最も類似するクラスタを順に抽出し、

上記ＳＢＦＳは、上記キーワードのbi-gramの一致度に基づくものであり、
上記ＢＭＦＳは、解答候補が質問文のbi-gramと一致したbi-gramを有するか否かに基づくものであり、
上記ＷＷＦＳは、以下のＩＳＦ値により重み付けされていること
$ISF(w_j,C_i)=(N(w_j,C_i)+0.5)/(N(w_j)+0.5)$
ここで、$N(w_j)$は、単語w_jが含まれているウェブページのスニペットの総数、$N(w_j,C_i)$は、単語w_jが含まれているクラスタC_iにおけるスニペットの数
を特徴とする質問応答方法。
【請求項２】
上記分類ステップは、ＳＶＭ(Support Vector Machine)を利用することにより、質問文の解析結果と最も類似するクラスタを順に抽出すること
を特徴とする請求項１記載の質問応答方法。
【請求項３】
ユーザから入力された質問文を構成する単語から、キーワードと、質問の種別を規定するアンサータイプとを特定する質問文解析手段と、
上記質問文解析手段により抽出されたキーワードを検索クエリーとし、当該キーワードに関連するウェブページを検索する検索手段と、
上記検索ステップにより検索された各ウェブページから、上記アンサータイプに基づいて解答候補を順次抽出する解答候補抽出手段と、
上記解答候補抽出手段により抽出された各解答候補に基づいて、候補選択スニペットをクラスタに割り当て、これをトレーニングデータとする、クラスタリング手段と、
上記トレーニングデータを解析することにより上記クラスタを分類し、更に上記トレーニングデータの解析と同一解析条件の下で上記質問文を解析し、当該質問文の解析結果と最も類似するクラスタを順に抽出し、これを応答として出力する分類手段とを備え、
上記分類手段は、上記トレーニングデータと上記質問文との単語重複度を示すＳＢＦＳ（similarity-based feature set）、上記トレーニングデータと上記質問文とのブーリアン重複度を示すＢＭＦＳ(Boolean match-based feature set)、上記トレーニングデータを構成する解答候補の前後を構成する文字を含めた文字列と上記質問文との類似度を示すＷＷＦＳ(window-based word feature set)に基づいて、当該質問文の解析結果と最も類似するクラスタを順に抽出し、
上記分類手段は、上記ＳＢＦＳを、上記キーワードのbi-gramの一致度に基づくものとし、
上記ＢＭＦＳを、解答候補が質問文のbi-gramと一致したbi-gramを有するか否かに基づくものとし、
上記ＷＷＦＳを、以下のＩＳＦ値により重み付けすること
$ISF(w_j,C_i)=(N(w_j,C_i)+0.5)/(N(w_j)+0.5)$
ここで、$N(w_j)$は、単語w_jが含まれているウェブページのスニペットの総数、$N(w_j,C_i)$は、単語w_jが含まれているクラスタC_iにおけるスニペットの数
を特徴とする質問応答システム。
【請求項４】
上記分類手段は、ＳＶＭ(Support Vector Machine)を利用することにより、質問文の解析結果と最も類似するクラスタを順に抽出すること
を特徴とする請求項３記載の質問応答システム。
【請求項５】
上記質問文解析手段と、上記検索手段と、上記解答候補抽出手段と、上記クラスタリング手段と、上記分類手段とを備える制御装置と、当該制御装置に対して通信網を介して情報を送受信可能な複数のユーザ用端末装置とを備え、
上記ユーザ用端末装置は、ユーザからの上記質問文の入力を受け付け、これを通信網を介して上記制御装置における上記質問文解析手段へと送信するとともに、当該制御装置における上記分類手段から出力される上記応答を上記通信網を介して受信し、これをユーザに

表示すること
を特徴とする請求項３又は４記載の質問応答システム。
【請求項６】
ユーザから入力された質問文を構成する単語から、キーワードと、質問の種別を規定するアンサータイプとを特定する質問文解析ステップと、
上記質問文解析ステップにおいて抽出したキーワードを検索クエリーとし、当該キーワードに関連するウェブページを検索する検索ステップと、
上記検索ステップにおいて検索した各ウェブページから、上記アンサータイプに基づいて解答候補を順次抽出する解答候補抽出ステップと、
上記解答候補抽出ステップにおいて抽出した各解答候補に基づいて、候補選択スニペットをクラスタに割り当て、これをトレーニングデータとする、クラスタリングステップと、
上記トレーニングデータを解析することにより上記クラスタを分類し、更に上記トレーニングデータの解析と同一解析条件の下で上記質問文を解析し、当該質問文の解析結果と最も類似するクラスタを順に抽出し、これを応答として出力する分類ステップとをコンピュータに実行させ、
上記分類ステップは、上記トレーニングデータと上記質問文との単語重複度を示すＳＢＦＳ（similarity-based feature set）、上記トレーニングデータと上記質問文とのブーリアン重複度を示すＢＭＦＳ(Boolean match-based feature set)、上記トレーニングデータを構成する解答候補の前後を構成する文字を含めた文字列と上記質問文との類似度を示すＷＷＦＳ(window-based word feature set)に基づいて、当該質問文の解析結果と最も類似するクラスタを順に抽出し、
上記ＳＢＦＳは、上記キーワードのbi-gramの一致度に基づくものであり、
上記ＢＭＦＳは、解答候補が質問文のbi-gramと一致したbi-gramを有するか否かに基づくものであり、
上記ＷＷＦＳは、以下のＩＳＦ値により重み付けされていること
$ISF(w_j,C_i)=(N(w_j,C_i)+0.5)/(N(w_j)+0.5)$
ここで、$N(w_j)$は、単語w_jが含まれているウェブページのスニペットの総数、$N(w_j,C_i)$は、単語w_jが含まれているクラスタC_iにおけるスニペットの数
を特徴とするプログラム。
【請求項７】
上記分類ステップは、ＳＶＭ(Support Vector Machine)を利用することにより、質問文の解析結果と最も類似するクラスタを順に抽出すること
を特徴とする請求項６記載のプログラム。
【請求項８】
請求項７項記載のプログラムが記録されていることを特徴とする記録媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
ユーザから入力された質問文に対して応答を出力可能な質問応答方法及びシステム、プログラム並びに記録媒体に関する。
【背景技術】
【０００２】
近年におけるインターネットの普及に伴い、ユーザは、検索エンジンを利用して所望のウェブページをサーチし、そこから知見を得ることが可能となってきている。特にユーザが知りたい情報について検索エンジンを介してウェブページを検索する場合、検索クエリーとして、キーワードを入力することにより、当該キーワードに関連するスニペットを持つウェブページが自動抽出され、これを検索リストとして表示されることになる。ユーザは、かかる検索リストに表示されたウェブページ一覧から所望のウェブページにアクセスし、知りたい情報を取得することが可能となる。
【０００３】

ところで、現在におけるウェブページの検索方法では、検索リストに表示されたウェブページ一覧から、所望の情報が記載されている、真のウェブページをユーザ自身が順次アクセスしながら見つけ出す必要があり、労力の負担が増大し、また検索に要する時間が長期化してしまうという問題点があった。

【０００４】

このため、このような検索エンジンを介して検索リストを表示する代替として、ユーザ自身が知りたい情報を自然な文章として端末を介して入力し、かかる質問文に対する応答を直接出力する質問応答システムが従来から望まれていた。このため、かかる質問応答システムに関する研究も従来より行われていた。

【０００５】

従来の質問応答システムは、４つのカテゴリーに分類することができる。

【０００６】

先ず、質問文の全てのキーワードと、解答候補との間で類似性を示す距離を求め、これに基づいて解答候補から正解を選び出すモデルが提案されている。しかし、このモデルでは、質問と解答候補が依拠する文章とが表面上一致していないだけで、正解を出すことができなくなるという問題点があった。

【０００７】

また、質問文を最初に予め定義したカテゴリーに分類し、これをオフラインの下で学習したアンサーパターンを利用して正解を抽出するモデルも提案されている。しかし、このモデルは、予め定義した何種類かの質問のタイプに対しては高い正確性を出すことが可能であるが、オープンドメインな質問応答のための質問のタイプを定義するのが困難であり、あらゆるタイプの質問に対して対応することができない。

【０００８】

自然言語処理(ＮＬＰ：Natural Language Processing)に基づくモデルは、ユーザの質問を解析し、応答に相当する文を意味的な表現へと繋げ、そして意味的にマッチングするものを解答として見つけ出すものである。このモデルは、ＴＲＥＣ（Text REtrieval Conference）のワークショップにおいてよく実演されるものであるが、ＮＬＰツールの高パフォーマンスに大きく依存するものである。このため、処理時間が長時間に亘るとともに、作業量の増加が無視できない。

【０００９】

さらに、マシーンラーニングに基づくモデルも研究されている（例えば、特許文献１参照。）が、未だその有用性は確立されていない。

【００１０】

即ち、これらの開示技術は、手入力された質問と解答のペアが所定量必要になるという問題点がある。また、マシーンラーニング技術に特化した上記ペアを収集するのは多大な労力が必要になり、コスト増につながる。

【非特許文献１】Jun Suzuki,Yutaka Sasaki,Eisaku Maeda.SVM Answer Selection for Open-Domain Question Answering,In Proc.of Coling-2002,pp974～980(2002).

【発明の開示】

【発明が解決しようとする課題】

【００１１】

そこで、本発明は、上述した問題点に鑑みて案出されたものであり、オープンドメインな質問応答システムを実現する上で、その解答の正答率を向上させることが可能な質問応答方法及びシステム、プログラム並びに記録媒体を提供することにある。

【課題を解決するための手段】

【００１２】

本発明を適用した質問応答方法は、ユーザから入力された質問文を構成する単語から、キーワードと、質問の種別を規定するアンサータイプとを特定する質問文解析ステップと、上記質問文解析ステップにおいて抽出したキーワードを検索クエリーとし、当該キーワードに関連するウェブページを検索する検索ステップと、上記検索ステップにおいて検索

した各ウェブページから、上記アンサータイプに基づいて解答候補を順次抽出する解答候補抽出ステップと、上記解答候補抽出ステップにおいて抽出した各解答候補に基づいて、候補選択スニペットをクラスタに割り当て、これをトレーニングデータとする、クラスタリングステップと、上記トレーニングデータを解析することにより上記クラスタを分類し、更に上記トレーニングデータの解析と同一解析条件の下で上記質問文を解析し、当該質問文の解析結果と最も類似するクラスタを順に抽出し、これを応答として出力する分類ステップとを有し、上記分類ステップは、上記トレーニングデータと上記質問文との単語重複度を示すＳＢＦＳ（similarity-based feature set）、上記トレーニングデータと上記質問文とのブーリアン重複度を示すＢＭＦＳ(Boolean match-based feature set)、上記トレーニングデータを構成する解答候補の前後を構成する文字を含めた文字列と上記質問文との類似度を示すＷＷＦＳ(window-based word feature set)に基づいて、当該質問文の解析結果と最も類似するクラスタを順に抽出し、上記ＳＢＦＳは、上記キーワードのbi-gramの一致度に基づくものであり、上記ＢＭＦＳは、解答候補が質問文のbi-gramと一致したbi-gramを有するか否かに基づくものであり、上記ＷＷＦＳは、以下のＩＳＦ値により重み付けされていること

$ISF(w_j,C_i)=(N(w_j,C_i)+0.5)/(N(w_j)+0.5)$

ここで、$N(w_j)$は、単語w_jが含まれているウェブページのスニペットの総数、$N(w_j,C_i)$は、単語w_jが含まれているクラスタC_iにおけるスニペットの数、

を特徴とする。

【００１３】

本発明を適用した質問応答システムは、ユーザから入力された質問文を構成する単語から、キーワードと、質問の種別を規定するアンサータイプとを特定する質問文解析手段と、上記質問文解析手段により抽出されたキーワードを検索クエリーとし、当該キーワードに関連するウェブページを検索する検索手段と、上記検索ステップにより検索された各ウェブページから、上記アンサータイプに基づいて解答候補を順次抽出する解答候補抽出手段と、上記解答候補抽出手段により抽出された各解答候補に基づいて、候補選択スニペットをクラスタに割り当て、これをトレーニングデータとする、クラスタリング手段と、上記トレーニングデータを解析することにより上記クラスタを分類し、更に上記トレーニングデータの解析と同一解析条件の下で上記質問文を解析し、当該質問文の解析結果と最も類似するクラスタを順に抽出し、これを応答として出力する分類手段とを備え、上記分類手段は、上記トレーニングデータと上記質問文との単語重複度を示すＳＢＦＳ（similarity-based feature set）、上記トレーニングデータと上記質問文とのブーリアン重複度を示すＢＭＦＳ(Boolean match-based feature set)、上記トレーニングデータを構成する解答候補の前後を構成する文字を含めた文字列と上記質問文との類似度を示すＷＷＦＳ(window-based word feature set)に基づいて、当該質問文の解析結果と最も類似するクラスタを順に抽出し、上記分類手段は、上記ＳＢＦＳを、上記キーワードのbi-gramの一致度に基づくものとし、上記ＢＭＦＳを、解答候補が質問文のbi-gramと一致したbi-gramを有するか否かに基づくものとし、上記ＷＷＦＳを、以下のＩＳＦ値により重み付けすること

$ISF(w_j,C_i)=(N(w_j,C_i)+0.5)/(N(w_j)+0.5)$

ここで、$N(w_j)$は、単語w_jが含まれているウェブページのスニペットの総数、$N(w_j,C_i)$は、単語w_jが含まれているクラスタC_iにおけるスニペットの数

を特徴とする。

【００１４】

本発明を適用したプログラムは、ユーザから入力された質問文を構成する単語から、キーワードと、質問の種別を規定するアンサータイプとを特定する質問文解析ステップと、上記質問文解析ステップにおいて抽出したキーワードを検索クエリーとし、当該キーワードに関連するウェブページを検索する検索ステップと、上記検索ステップにおいて検索した各ウェブページから、上記アンサータイプに基づいて解答候補を順次抽出する解答候補抽出ステップと、上記解答候補抽出ステップにおいて抽出した各解答候補に基づいて、候

補選択スニペットをクラスタに割り当て、これをトレーニングデータとする、クラスタリングステップと、上記トレーニングデータを解析することにより上記クラスタを分類し、更に上記トレーニングデータの解析と同一解析条件の下で上記質問文を解析し、当該質問文の解析結果と最も類似するクラスタを順に抽出し、これを応答として出力する分類ステップとをコンピュータに実行させ、上記分類ステップは、上記トレーニングデータと上記質問文との単語重複度を示すＳＢＦＳ（similarity-based feature set）、上記トレーニングデータと上記質問文とのブーリアン重複度を示すＢＭＦＳ(Boolean match-based feature set)、上記トレーニングデータを構成する解答候補の前後を構成する文字を含めた文字列と上記質問文との類似度を示すＷＷＦＳ(window-based word feature set)に基づいて、当該質問文の解析結果と最も類似するクラスタを順に抽出し、上記ＳＢＦＳは、上記キーワードのbi-gramの一致度に基づくものであり、上記ＢＭＦＳは、解答候補が質問文のbi-gramと一致したbi-gramを有するか否かに基づくものであり、上記ＷＷＦＳは、以下のＩＳＦ値により重み付けされていること

$ISF(w_j,C_i)=(N(w_j,C_i)+0.5)/(N(w_j)+0.5)$

ここで、$N(w_j)$は、単語w_jが含まれているウェブページのスニペットの総数、$N(w_j,C_i)$は、単語w_jが含まれているクラスタC_iにおけるスニペットの数

を特徴とする。

【発明の効果】

【００１５】

上述した構成からなる本発明では、後述する表１～３の結果から示されるように、オープンドメインな質問応答システムを実現する上で、その解答の正答率を向上させることが可能となる。

【発明を実施するための最良の形態】

【００１６】

以下、本発明を実施するための最良の形態として、ユーザから入力された質問文に対して応答を出力可能な質問応答システムに監視、図面を参照しながら詳細に説明をする。

【００１７】

本発明を適用した質問応答システム１は、図１に示すように、質問文を入力するユーザにより操作されるユーザ端末装置１１と、このユーザ端末装置１１により通信網１２を介してそれぞれアクセス可能なウェブサーバ１３と、このウェブサーバ１３を制御するための制御装置１４とを備えている。

【００１８】

ユーザ端末装置１１は、例えばパーソナルコンピュータ（ＰＣ）等が適用され、質問文を入力するためのマウスやキーボード等からなる操作部と、情報をユーザに対して表示するための、例えば液晶ディスプレイからなる表示部を備える。このユーザ端末装置１１は、ユーザからの質問文の入力を受け付けた場合に、これを通信網１２を介して制御装置１４へと送信する。

【００１９】

通信網１２は、例えばウェブサーバ１３とユーザ端末装置１１とを電話回線を介して接続されるインターネット網を始め、ＴＡ／モデムと接続されるＩＳＤＮ（Integrated Services Digital Network）／Ｂ（broadband）－ＩＳＤＮ等のように、情報の双方向送受信を可能とした公衆通信網等である。

【００２０】

また制御装置１４も同様にコンピュータで構成されるものであり、相互にバスで接続されたＣＰＵ(Central Processing Unit)や、メモリ、固定ディスクと、通信網１２を介してユーザ端末装置１１との間で情報を送受信するための通信インターフェースとを備えている。実際に、本発明に係る質問応答システム１を実行するためのプログラムは、この制御装置１４における固定ディスク等にインストールされることになる。また、このプログラムは、他のＣＤ－ＲＯＭ等に記録された記録媒体として具体化することも可能となる。

【００２１】

制御装置１４は、通信網１２を介してユーザ端末装置１１から受信した質問文を受けて、ウェブサーバ１３へアクセスし、後述するような処理を実行することにより、上記質問文に対する応答を作り出し、通信網１２を介してユーザ端末装置１１へと送信する。ユーザ端末装置１１は、送られてきた応答を液晶ディスプレイからなる表示部を介して表示する。

【００２２】

次に、本発明を適用した質問応答システム１の動作について説明をする。

【００２３】

図２は、質問応答システム１を実行する上でのフローチャートを示している。先ずステップＳ１において、ユーザからの質問文の入力を受け付ける。ちなみに、本発明は、オープンドメインの質問応答の実現を想定しているところ、ユーザは、言語や入力形式に支配されることなく、自然に質問したい内容を文章にし、これを入力していくことになる。このため、ユーザの入力すべき内容について、複雑なルールは特段存在せず、また高精度な言語解析技術も特段必要としない。

【００２４】

この質問文は、文書検索のクエリーとなりえるキーワードの集合と、質問の種別を規程するアンサータイプから構成されることになる。例えば、「いつ潜水艦が沈んだか？」という質問文が入力された場合においてキーワードは、「潜水艦」、「沈む」に相当し、アンサータイプは、「いつ」に相当するものとなる。即ち、このキーワードは、質問文中に含まれる名詞、動詞等を規程する単語であり、アンサータイプは、時、場所、主体、数量等、実際にユーザが知りたいカテゴリを示すものである。また、ステップＳ２は、質問からキーワードを抽出し、また疑問詞に基づいて質問の解答タイプを分類する。

【００２５】

次に、ステップＳ２へ移行し、上記ステップＳ１において抽出したキーワードを検索クエリーとし、当該キーワードに関連するウェブページを検索する。その結果、このステップＳ２においては、キーワードに関係するウェブページが順次検索されてくることになる。このステップＳ２においては、例えば検索エンジンにおけるスニペットの記載に基づいて検索を行うようにしてもよい。ちなみに、このステップＳ２において、上述した例の質問文が入力された場合には、「潜水艦」、「沈む」というキーワードに関係するウェブページが順次検索されてくることになる。

【００２６】

次にステップＳ３へ移行し、ステップＳ２において検索した各ウェブページから、アンサータイプに基づいて解答候補を順次抽出する。即ち、アンサータイプとして、時、場所、主体、数量等の何れかがステップＳ１において抽出されているため、これに関係する解答候補を抽出してくることになる。このステップＳ３においては、例えば検索エンジンにおけるスニペットの記載から解答候補を抽出するようにしてもよい。ちなみに、このステップＳ３において、上述した例の質問文が入力された場合には、「いつ」に相当する時を表すアンサータイプに基づく解答候補を抽出してくることになる。

【００２７】

次にステップＳ４へ移行し、クラスタリングを行う。このクラスタリングは、解答候補抽出ステップにおいて抽出した各解答候補に基づいて、候補選択スニペットをクラスタに割り当て、これをトレーニングデータとする。同一の解答候補を含むウェブスニペットは、同系列のスニペットとみなし、これを同一のクラスタに属させる。そして、この割り当ての結果を分類ステップのトレーニングデータとしてとして利用する。

【００２８】

即ち、検索エンジンのトップｍ位までのスニペット $\{s_1, s_2, \cdots, s_m\}$ からｎ個の解答候補 $\{c_1, c_2, \cdots, c_n\}$ を抽出する。これらのスニペットは、それぞれ解答候補 $\{c_i\}$ と少なくとも１の質問キーワード $\{q_i\}$ を保有している。そして、これらスニペット $\{s_1, s_2, \cdots, s_m\}$ は、ウェブサーチ結果のクラスタリングにより、ｎ個のクラスタ $\{C_1, C_2, \cdots, C_n\}$ へと割り当てられることになる。

【００２９】
仮にスニペットがＬ個の異なる解答候補を保有するものであれば、そのスニペットはＬ個の異なるクラスタに割り当てられることになる。また、異なるスニペットの解答候補が互いに同一であれば、これらのスニペットは同一のクラスタに割り当てられることになる。
【００３０】
最終的に、クラスタ $\{C_i\}$ は、解答候補 $\{c_i\}$ の数によって決定される。そしてクラスタC_iのクラスタ名は、解答候補c_iに基づくものとなる。これらクラスタ化された解答候補c_iが上述したトレーニングデータとなる。
【００３１】
次に、ステップＳ５へ移行し、トレーニングデータを解析することにより上記クラスタを分類する。
【００３２】
さらに次にステップＳ６へ移行し、トレーニングデータの解析と同一解析条件の下で上記質問文を解析し、当該質問文の解析結果と最も類似するクラスタを順に抽出し、これを応答として出力する。この分類ステップＳ６は、ユーザの質問から分析したテストベクトルを利用することにより、クラスタの一つに割り当て、質問に対する解答を、質問のクラスタの名称と仮定する。
【００３３】
以下、このステップＳ５とステップＳ６を一つの分類ステップとして説明をしていく。この分類ステップでは、ＳＶＭ(Support Vector Machine)を利用することにより、質問文の解析結果と最も類似するクラスタを順に抽出するようにしてもよい。
【００３４】
また、この分類ステップでは、トレーニングデータと質問文との単語重複度を示すＳＢＦＳ（similarity-based feature set）、トレーニングデータと質問文とのブーリアン重複度を示すＢＭＦＳ(Boolean match-based feature set)、トレーニングデータを構成する解答候補の前後を構成する文字を含めた文字列と上記質問文との類似度を示すＷＷＦＳ(window-based word feature set)に基づいて、当該質問文の解析結果と最も類似するクラスタを順に抽出するようにしてもよい。
【００３５】
ＳＢＦＳとしては、キーワードの重複度、キーワードの非重複度、キーワードのbi-gramの一致度、シーソラスの重複度、キーワードと解答候補との間の規格化距離の何れか１以上に基づくものであってもよい。
【００３６】
ＢＭＦＳは、人名が一致しているか否か、地域名が一致しているか否か、組織名が一致しているか否か、時を示す単語が一致しているか否か、数量を示す単語が一致しているか否か、語源が一致しているか否か、解答候補が質問文のbi-gramと一致したbi-gramを有するか否か、解答候補がネームディエンティティタイプを要求されているか否かの何れか１以上に基づくものであってもよい。
【００３７】
ＷＷＦＳは、以下のＩＳＦ値により重み付けされていてもよい。
$ISF(w_j,C_i)=(N(w_j,C_i)+0.5)/(N(w_j)+0.5)$
ここで、$N(w_j)$は、単語w_jが含まれているウェブページのスニペットの総数、$N(w_j,C_i)$は、単語w_jが含まれているクラスタC_iにおけるスニペットの数である。
【００３８】
最後にステップＳ７に移行し、実際に上記プロセスの下で得た回答をユーザ端末装置１１の表示部を介して表示する。
【００３９】
次に、本発明を用いた質問応答システム１による効果について説明をする。
【００４０】

中国語によるウェブの質問応答（QA）における３種類のデータセットとしてのCTREC04、CTREC05、CTEST05を利用し、本発明を適用したU(unsupervised)-SVMを検証することとした。CTREC04は、TREC2004FACTOIDのテスト問題から翻訳された１７８個の中国語による質問のセットからなる。CTREC05は、TRE2005FACTOIDのテスト問題から翻訳された２７９個の中国語による質問のセットからなる。CTEST05は、中国語で記載されたものを除く、TERCのテスト問題に類似する１７８個の中国語の質問のセットからなる。

【００４１】

実験は、３つの評価項目、即ち、top_1,top_5,mrr_5に基づいて評価を行った。評価結果を表１、２、３に示す。ここで、top_1は、解答の正確性がトップ１位である解答が含まれている割合を示している。top_5は、解答の正確性がトップ５位以内である解答の一つが含まれている割合を示している。mrr_5は、各質問に対する正解の平均相対ランク（１/ｎ）を示しており、ここで最高ランクｎ（ｎ≦５）としている。

【００４２】

各データセット（CTREC04、CTREC05、CTEST05）をU-SVMを用いて解析することにより得られた各評価項目（top_1,top_5,mrr_5）を表１に示す。

【００４３】

【表１】

	TREC04	TREC05	Test05
top_1	60.82%	57.33%	59.09%
mrr_5	71.31%	65.61%	67.34%
top_5	88.66%	80.00%	81.82%

【００４４】

また、表２において、CTrec04と、CTrec05のテストデータをU-SVMとthe Retrieval-M（従来の検索手法）それぞれを用いて解析することにより得られた各評価項目（top_1,top_5,mrr_5）の相対比較を表２に示す。

【００４５】

【表２】

		Retrieval-M	U-SVM
Ctrec04	top_1	27.84%	53.61%
	mrr_5	43.67%	66.25%
	top_5	71.13%	88.66%
Ctrec05	top_1	34.00%	50.00%
	mrr_5	48.20%	62.38%
	top_5	71.33%	82.67%

【００４６】

さらに、Pattern-M（パターン重視による手法）とS-SVM（SVMを利用した教化学習法によるもの）に対するU-SVMのパフォーマンス性を比較するために、CTEST05のデータセットを用いて検証を行った。表３は、U-SVM、Pattern-M、S-SVMの各モデルを用いてCTEST05を解析することにより得られた各評価項目（top_1,mrr_5）の相対比較を示している。

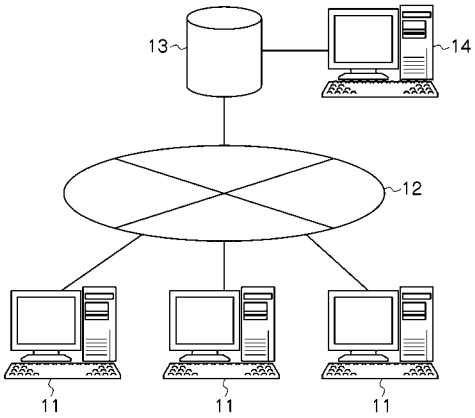
【００４７】

【表３】

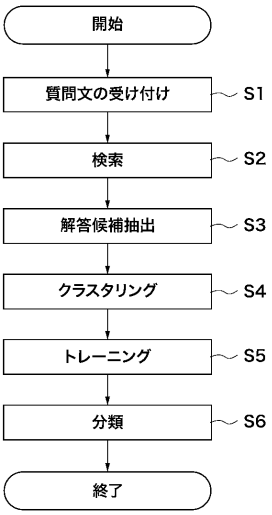
	S-SVM	Pattern-M	U-SVM
top_1	44.89%	53.14%	59.09%
mrr_5	56.49%	61.28%	67.34%
top_5	74.43%	73.14%	81.82%

【００４８】
上述した表１～３の結果から、各モデルによる正答率（パフォーマンスランキング）は、U-SVM＞Pattern-M＞S-SVM＞Retrieval-Mの順となった。
【００４９】
即ち、本発明では、上述した図２に示すフローに基づいて、解答を抽出していくため、正答率を向上させることが可能となる。
【図面の簡単な説明】
【００５０】
【図１】本発明を適用した質問応答システムの構成例を示す図である。
【図２】本発明を適用した質問応答システムの処理手順を示すフローチャートである。
【符号の説明】
【００５１】
１　質問応答システム
１１　ユーザ端末装置
１２　通信網
１３　ウェブサーバ
１４　制御装置

【図１】


13
14
12
11
11
11

【図２】


開始
質問文の受け付け
S1
検索
S2
解答候補抽出
S3
クラスタリング
S4
トレーニング
S5
分類
S6
終了

フロントページの続き

(56)参考文献 特開２００６－２４４１０２（ＪＰ，Ａ）
特開平０９－２３１２３８（ＪＰ，Ａ）
特開２００３－１５０６２４（ＪＰ，Ａ）
永田昌明、外２名，日本語自然文検索システム Ｗｅｂ Ａｎｓｗｅｒｓ，言語処理学会第１２回年次大会発表論文集，言語処理学会，２００６年 3月１３日，ｐ．３２０－３２３
佐々木裕，ＳＶＭを用いた学習型質問応答システムＳＡＩＱＡ－ＩＩ，情報処理学会論文誌，日本，社団法人情報処理学会，２００４年 2月１５日，第４５巻，第２号，ｐｐ．６３５－６４６

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １７／３０