

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7209433号
(P7209433)

(45)発行日 令和5年1月20日(2023.1.20)

(24)登録日 令和5年1月12日(2023.1.12)

(51)国際特許分類 F I
G 0 6 F 16/953(2019.01) G 0 6 F 16/953

請求項の数 13 (全17頁)

(21)出願番号	特願2021-516412(P2021-516412)	(73)特許権者	512015127
(86)(22)出願日	令和2年8月11日(2020.8.11)		バイドゥ オンライン ネットワーク テクノロジー(ペキン)カンパニー リミテッド
(65)公表番号	特表2022-520683(P2022-520683 A)		中華人民共和国 ペキン ハイディエン ディストリクト シャンディー テンス ストリート ナンバー 10 バイドゥ キャンパス 3エフ
(43)公表日	令和4年4月1日(2022.4.1)	(74)代理人	110002468 弁理士法人後藤特許事務所
(86)国際出願番号	PCT/CN2020/108438	(72)発明者	謝 達
(87)国際公開番号	WO2021/139154		中華人民共和国 100085 北京市海 淀区上地十街10号百度大厦三層
(87)国際公開日	令和3年7月15日(2021.7.15)	(72)発明者	鄭 志洵
審査請求日	令和3年3月22日(2021.3.22)		中華人民共和国 100085 北京市海 淀区上地十街10号百度大厦三層
(31)優先権主張番号	202010024434.0		最終頁に続く
(32)優先日	令和2年1月10日(2020.1.10)		
(33)優先権主張国・地域又は機関	中国(CN)		

(54)【発明の名称】 データプリフェッチ方法、デバイス、電子機器、コンピュータ可読記憶媒体及びコンピュータプログラム製品

(57)【特許請求の範囲】

【請求項1】

ユーザが入力したクエリプレフィックスを取得することと、
 事前にトレーニングされた言語モデルに基づいて前記クエリプレフィックスを判別して、
 前記クエリプレフィックスのパープレキシティを取得することと、
 前記パープレキシティが所定の閾値よりも小さいか否かを判断することと、
 前記パープレキシティが前記所定の閾値よりも小さい場合に、前記クエリプレフィックスに基づいてプリフェッチ要求を送信することと、を含み、
 前記クエリプレフィックスに特殊文字が含まれる場合に、
 前記クエリプレフィックスのパープレキシティ補正係数を取得することをさらに含み、
 前記パープレキシティが所定の閾値よりも小さいか否かを判断することは、
 前記パープレキシティ補正係数を用いて前記パープレキシティを補正することと、
 補正後のパープレキシティが前記所定の閾値よりも小さいか否かを判断することと、を含み、
 前記パープレキシティが前記所定の閾値よりも小さい場合に、前記クエリプレフィックスに基づいてプリフェッチ要求を送信することは、
 前記補正後のパープレキシティが前記所定の閾値よりも小さい場合に、前記クエリプレフィックスに基づいて前記プリフェッチ要求を送信することを含む、
 データプリフェッチ方法。

【請求項2】

前記クエリプレフィックスのパープレキシティ補正係数を取得することは、
下式を採用して前記クエリプレフィックスのパープレキシティ補正係数 R_e を計算することを含み、

【数 1】

$$R_e = \frac{f}{N} \log_2 \text{count}(sw)$$

ここで、 N は、前記クエリプレフィックスの文の長さを表し、 $\text{count}(sw)$ は、前記クエリプレフィックスに含まれる特殊文字の個数を表し、 f は、所定の係数を表す、請求項 1 に記載の方法。

10

【請求項 3】

事前にトレーニングされた言語モデルに基づいて前記クエリプレフィックスを判別して、前記クエリプレフィックスのパープレキシティを取得することは、

前記クエリプレフィックスを分割して、複数の分割単語を取得することと、

各前記分割単語を事前にトレーニングされた語意モデルにそれぞれ入力して、各前記分割単語の単語埋め込みを生成し、かつ各前記分割単語の単語埋め込みに基づいて前記クエリプレフィックスの単語埋め込みを決定することと、

前記クエリプレフィックスの単語埋め込みを前記事前にトレーニングされた言語モデルに入力して、前記クエリプレフィックスのパープレキシティを取得することと、を含む請求項 1 に記載の方法。

20

【請求項 4】

各前記分割単語の単語埋め込みに基づいて前記クエリプレフィックスの単語埋め込みを決定することは、

各前記分割単語の単語埋め込みを結合して、前記クエリプレフィックスの単語埋め込みを取得することを含む、請求項 3 に記載の方法。

【請求項 5】

前記クエリプレフィックスに基づいてプリフェッチ要求を送信する後に、

サーバが前記プリフェッチ要求に基づいて返された前記クエリプレフィックスに対応するプリフェッチ結果を受信することをさらに含む、請求項 1 に記載の方法。

30

【請求項 6】

ユーザが入力したクエリプレフィックスを取得する第 1 の取得モジュールと、事前にトレーニングされた言語モデルに基づいて前記クエリプレフィックスを判別して、前記クエリプレフィックスのパープレキシティを取得する判別モジュールと、

前記パープレキシティが所定の閾値よりも小さいか否かを判断する判断モジュールと、

前記パープレキシティが前記所定の閾値よりも小さい場合に、前記クエリプレフィックスに基づいてプリフェッチ要求を送信する送信モジュールと、を含み、

前記クエリプレフィックスに特殊文字が含まれる場合に、前記クエリプレフィックスのパープレキシティ補正係数を取得する第 2 の取得モジュールをさらに含む、

前記判断モジュールは、

前記パープレキシティ補正係数を用いて前記パープレキシティを補正する補正ユニットと、

40

補正後のパープレキシティが前記所定の閾値よりも小さいか否かを判断する判断ユニットと、を含み、

前記送信モジュールは、具体的には、前記補正後のパープレキシティが前記所定の閾値よりも小さい場合に、前記クエリプレフィックスに基づいて前記プリフェッチ要求を送信する、

データプリフェッチデバイス。

【請求項 7】

前記第 2 の取得モジュールは、具体的には、

50

下式を採用して前記クエリプレフィックスのパープレキシティ補正係数 R_e を計算し、
【数 2】

$$R_e = \frac{f}{N} \log_2 \text{count}(sw)$$

ここで、 N は、前記クエリプレフィックスの文の長さを表し、 $\text{count}(sw)$ は、前記クエリプレフィックスに含まれる特殊文字の個数を表し、 f は、所定の係数を表す、請求項 6 に記載のデバイス。

【請求項 8】

前記判別モジュールは、

前記クエリプレフィックスを分割して、複数の分割単語を取得する分割ユニットと、
各前記分割単語を事前にトレーニングされた語意モデルにそれぞれ入力して、各前記分割単語の単語埋め込みを生成する生成ユニットと、

各前記分割単語の単語埋め込みに基づいて前記クエリプレフィックスの単語埋め込みを決定する決定ユニットと、

前記クエリプレフィックスの単語埋め込みを前記事前にトレーニングされた言語モデルに入力して、前記クエリプレフィックスのパープレキシティを取得する判別ユニットと、を含む請求項 6 に記載のデバイス。

【請求項 9】

前記決定ユニットは、具体的には、各前記分割単語の単語埋め込みを結合して、前記クエリプレフィックスの単語埋め込みを取得する、請求項 8 に記載のデバイス。

【請求項 10】

サーバが前記プリフェッチ要求に基づいて返された前記クエリプレフィックスに対応するプリフェッチ結果を受信する受信モジュールをさらに含む、請求項 6 に記載のデバイス。

【請求項 11】

少なくとも 1 つのプロセッサと、

前記少なくとも 1 つのプロセッサに通信可能に接続されているメモリと、を含み、

前記メモリには、前記少なくとも 1 つのプロセッサに実行可能で、前記少なくとも 1 つのプロセッサによって実行されると、前記少なくとも 1 つのプロセッサに請求項 1 ~ 5 のいずれか 1 項に記載の方法を実行させることができる命令が記憶されている、電子機器。

【請求項 12】

請求項 1 ~ 5 のいずれか 1 項に記載の方法をコンピュータに実行させるコンピュータ命令が記憶されている非一時的コンピュータ可読記憶媒体。

【請求項 13】

プロセッサによって実行されると、請求項 1 ~ 5 のいずれか一項に記載の方法が実現されるコンピュータプログラムを含む、コンピュータプログラム製品。

【発明の詳細な説明】

【技術分野】

【0001】

本願は、コンピュータの技術分野に関し、特にスマート検索の技術分野に関する。

【背景技術】

【0002】

検索プリフェッチ機能は、ユーザが実際に検索をクリックする前にクライアントにより開始した、クエリプレフィックスに対するプリフェッチ要求であり、事前にプリフェッチ結果を取得してユーザに提示することにより、ユーザに検索速度がより速い体験を与え、ユーザへの喜びを増加させることができる。

【0003】

現在、プリフェッチ要求を発行する方式は、主に、ユーザが入力したクエリプレフィックスに対する補完マッチングに基づいてプリフェッチ要求を発行することである。例えば、

10

20

30

40

50

ユーザが「劉徳」というクエリプレフィックスを入力するとき、「劉徳」というクエリプレフィックスを「劉徳華」に補完してプリフェッチ要求発行をトリガーする。これにより、補完マッチングに基づいてプリフェッチ要求を発行する方式により、ユーザがクエリプレフィックスを入力する過程に、大量のプリフェッチ要求を出すことを引き起こし、これらのプリフェッチ要求のプリフェッチ成功率が低く、機器コストを無駄にするとともにシステムの安定性に一定の影響を与える。

【発明の概要】

【0004】

本願の実施例は、データプリフェッチ方法、デバイス、電子機器及びコンピュータ可読記憶媒体を提供する。

10

【0005】

第1の様態では、本願の実施例に係るデータプリフェッチ方法は、ユーザが入力したクエリプレフィックスを取得することと、事前にトレーニングされた言語モデルに基づいて前記クエリプレフィックスを判別して、前記クエリプレフィックスのパープレキシティを取得することと、前記パープレキシティが所定の閾値よりも小さいか否かを判断することと、前記パープレキシティが前記所定の閾値よりも小さい場合に、前記クエリプレフィックスに基づいてプリフェッチ要求を送信することと、を含む。

【0006】

このように、クエリプレフィックスのパープレキシティの判別に基づくプリフェッチ要求の送信は、現在の、補完マッチングに基づくプリフェッチ要求の送信に比べて、パープレキシティの高いプリフェッチ要求を直接フィルタリングして除去することにより、プリフェッチの成功率を向上させ、かつ過剰なプリフェッチ要求によるバックエンドサーバの機器コストをさらに低減することができるとともに、システム安定性に影響を与えることと、過剰なプリフェッチ結果の提示によりユーザに視覚体験上の干渉を与えることとを避け、ユーザ体験を向上させることができる。

20

【0007】

好ましくは、前記クエリプレフィックスに特殊文字が含まれる場合に、前記方法は、前記クエリプレフィックスのパープレキシティ補正係数を取得することをさらに含み、前記パープレキシティが所定の閾値よりも小さいか否かを判断することは、前記パープレキシティ補正係数を用いて前記パープレキシティを補正することと、補正後のパープレキシティが前記所定の閾値よりも小さいか否かを判断することと、を含み、前記パープレキシティが前記所定の閾値よりも小さい場合に、前記クエリプレフィックスに基づいてプリフェッチ要求を送信することは、前記補正後のパープレキシティが前記所定の閾値よりも小さい場合に、前記クエリプレフィックスに基づいて前記プリフェッチ要求を送信することを含む。

30

【0008】

このように、この補正プロセスにより、特殊文字による、対応するクエリプレフィックスのパープレキシティへの影響を低減することにより、プリフェッチの成功率をさらに向上させることができる。

40

【0009】

好ましくは、前記クエリプレフィックスのパープレキシティ補正係数を取得することは、下式を採用して前記クエリプレフィックスのパープレキシティ補正係数 R_e を計算することを含み、

【数1】

$$R_e = \frac{f}{N} \log_2 \text{count}(sw)$$

50

ここで、Nは、前記クエリプレフィックスの文の長さを表し、count(sw)は、前記クエリプレフィックスに含まれる特殊文字の個数を表し、fは、所定の係数を表す。

【0010】

このように、特殊文字の個数を用いて対応するパープレキシティ補正係数を算出することにより、クエリプレフィックスのパープレキシティに対する最適化を実現することができる。

【0011】

好ましくは、事前にトレーニングされた言語モデルに基づいて前記クエリプレフィックスを判別して、前記クエリプレフィックスのパープレキシティを取得することは、

前記クエリプレフィックスを分割して、複数の分割単語を取得することと、

各前記分割単語を事前にトレーニングされた語意モデルにそれぞれ入力して、各前記分割単語の単語埋め込みを生成し、かつ各前記分割単語の単語埋め込みに基づいて前記クエリプレフィックスの単語埋め込みを決定することと、

前記クエリプレフィックスの単語埋め込みを前記事前にトレーニングされた言語モデルに入力して、前記クエリプレフィックスのパープレキシティを取得することと、を含む。

【0012】

このように、事前にトレーニングされた語意モデルにより、クエリプレフィックスに対応する、中国語の語意の理解力がより強い単語埋め込みを生成することができ、該単語埋め込みに基づいてクエリプレフィックスのパープレキシティを判別することにより、判別精度を向上させることができる。

【0013】

好ましくは、各前記分割単語の単語埋め込みに基づいて前記クエリプレフィックスの単語埋め込みを決定する前記ことは、

各前記分割単語の単語埋め込みを結合して、前記クエリプレフィックスの単語埋め込みを取得することを含む。

【0014】

好ましくは、前記クエリプレフィックスに基づいてプリフェッチ要求を送信することの後に、前記方法は、

サーバが前記プリフェッチ要求に基づいて返された前記クエリプレフィックスに対応するプリフェッチ結果を受信することをさらに含む。

【0015】

このように、ユーザがクエリプレフィックスを入力する過程に、取得されたプリフェッチ結果をユーザに提示し、ユーザに検索速度がより速い体験を与えることができる。

【0016】

第2の様態では、本願の実施例に係るデータプリフェッチデバイスは、

ユーザが入力したクエリプレフィックスを取得する第1の取得モジュールと、

事前にトレーニングされた言語モデルに基づいて前記クエリプレフィックスを判別して、

前記クエリプレフィックスのパープレキシティを取得する判別モジュールと、

前記パープレキシティが所定の閾値よりも小さいか否かを判断する判断モジュールと、

前記パープレキシティが前記所定の閾値よりも小さい場合に、前記クエリプレフィックスに基づいてプリフェッチ要求を送信する送信モジュールと、を含む。

【0017】

好ましくは、前記デバイスは、

前記クエリプレフィックスに特殊文字が含まれる場合に、前記クエリプレフィックスのパープレキシティ補正係数を取得する第2の取得モジュールをさらに含み、

前記判断モジュールは、

前記パープレキシティ補正係数を用いて前記パープレキシティを補正する補正ユニットと、

補正後のパープレキシティが前記所定の閾値よりも小さいか否かを判断する判断ユニットと、を含み、

前記送信モジュールは、具体的には、前記補正後のパープレキシティが前記所定の閾値よ

10

20

30

40

50

りも小さい場合に、前記クエリプレフィックスに基づいて前記プリフェッチ要求を送信する。

【0018】

好ましくは、前記第2の取得モジュールは、具体的には、下式を採用して前記クエリプレフィックスのパープレキシティ補正係数を計算し、

【数2】

$$Re = \frac{f}{N} \log_2 \text{count}(sw)$$

10

【0019】

ここで、Nは、前記クエリプレフィックスの文の長さを表し、count(sw)は、前記クエリプレフィックスに含まれる特殊文字の個数を表し、fは、所定の係数を表す。

【0020】

好ましくは、前記判別モジュールは、前記クエリプレフィックスを分割して、複数の分割単語を取得する分割ユニットと、各前記分割単語を事前にトレーニングされた語意モデルにそれぞれ入力して、各前記分割単語の単語埋め込みを生成する生成ユニットと、各前記分割単語の単語埋め込みに基づいて前記クエリプレフィックスの単語埋め込みを決定する決定ユニットと、前記クエリプレフィックスの単語埋め込みを前記事前にトレーニングされた言語モデルに入力して、前記クエリプレフィックスのパープレキシティを取得する判別ユニットと、を含む。

20

【0021】

好ましくは、前記決定ユニットは、具体的には、各前記分割単語の単語埋め込みを結合して、前記クエリプレフィックスの単語埋め込みを取得する。

【0022】

好ましくは、前記デバイスは、サーバが前記プリフェッチ要求に基づいて返された前記クエリプレフィックスに対応するプリフェッチ結果を受信する受信モジュールをさらに含む。

30

【0023】

第3の態様では、本願の実施例に係る電子機器は、少なくとも1つのプロセッサと、前記少なくとも1つのプロセッサに通信可能に接続されているメモリと、を含み、前記メモリには、前記少なくとも1つのプロセッサに実行可能で、前記少なくとも1つのプロセッサによって実行されると、前記少なくとも1つのプロセッサに上記データプリフェッチ方法を実行させることができる命令が記憶されている。

【0024】

第4の態様では、本願の実施例に係る非一時的コンピュータ可読記憶媒体には、上述したデータプリフェッチ方法をコンピュータに実行させるコンピュータ命令が記憶されている。

40

【0025】

上記好ましい形態の他の効果について、以下、具体的な実施例と組み合わせて説明する。

【図面の簡単な説明】

【0026】

図面は、本解決手段をよりよく理解するためのものであり、本願を限定するものではない。

【0027】

【図1】本願の実施例に係るデータプリフェッチ方法のフローチャートである。

【図2】本願の実施例に係るパープレキシティの判別プロセスのフローチャートである。

【図3】本願の具体的な実施例に係るプリフェッチ要求の発行の論理ブロック図である。

【図4】本願の実施例に係るデータプリフェッチ方法を実現するデータプリフェッチデバ

50

イスのブロック図である。

【図5】本願の実施例に係るデータプリフェッチ方法を実現する電子機器のブロック図である。

【発明を実施するための形態】

【0028】

以下、図面を参照しながら、本願の例示的な実施例を説明し、理解し易くするために、様々な詳細を説明するが、これらは単なる例示的なものであると考えるべきである。したがって、本願の範囲及び精神から逸脱することなく、ここで説明された実施例に対して様々な変更及び修正を行うことができることを、当業者は認識すべきである。同様に、明確さと簡潔さのために、以下の説明では、公知の機能及び構造についての説明を省略する。

10

【0029】

関連技術における、補完マッチングに基づいてプリフェッチ要求を発行する方式の成功率が低いという問題を解決するために、本願の実施例は、事前にトレーニングされた言語モデルを導入してユーザが入力したクエリプレフィックスを判別して、該クエリプレフィックスのパープレキシティを取得するとともに、該クエリプレフィックスのパープレキシティが所定の閾値よりも低い場合に、該クエリプレフィックスに基づいてプリフェッチ要求を送信するデータプリフェッチ方法を提供する。これにより、クエリプレフィックスのパープレキシティを、プリフェッチ要求を発行するか否かの重要な根拠とすることにより、パープレキシティの高いプリフェッチ要求を直接フィルタリングして除去することができ、プリフェッチの成功率を向上させることができる。

20

【0030】

図1を参照すると、本願の実施例に係るデータプリフェッチ方法は、電子機器に適用され、図1に示すように、以下のステップ101～104を含む。

【0031】

ステップ101では、ユーザが入力したクエリプレフィックスを取得する。

【0032】

本実施例では、上記クエリプレフィックスは、ユーザが入力操作を行うときに電子機器の検索ボックスに入力した、ユーザが現在入力しているクエリ内容を表示することができる。例えば、劉徳華の映画を検索しようとするとき、ユーザが現在入力している「劉徳華」は、クエリプレフィックスであってよい。一実施形態では、該検索ボックスは、電子機器にインストールされているクライアントアプリケーションプログラムの検索ボックスであってよい。該クライアントアプリケーションプログラムは、検索アプリケーションプログラム又は検索機能付きのアプリケーションプログラムであってよく、一般に、そのホームページには検索ボックス及び検索ボタンが設置されている。具体的には、ユーザが対応するアプリケーションプログラムアイコンをクリックした後に、アプリケーションプログラムのホームページを表示することができる。ユーザは、検索ボックスをクリックした後に、検索ボックスにクエリプレフィックスを入力することができる。

30

【0033】

理解できるように、本実施例の実行主体である電子機器は、ハードウェアであってもよく、ソフトウェアであってもよい。該電子機器がハードウェアである場合に、ウェブブラウザなどをサポートする様々な端末機器であってよく、例えば、スマートフォン、タブレットコンピュータ、電子書籍リーダー、ラップトップコンピュータ及びデスクトップコンピュータなどを含むが、これらに限定されない。該電子機器がソフトウェアである場合に、上記端末装置にインストールすることができ、かつ複数のソフトウェア又はソフトウェアモジュールとして実現するか、又は単一のソフトウェア又はソフトウェアモジュールとして実現することができる。ここで本実施例の実行主体を限定しない。

40

【0034】

ステップ102では、事前にトレーニングされた言語モデルに基づいて前記クエリプレフィックスを判別して、前記クエリプレフィックスのパープレキシティを取得する。

【0035】

50

本実施例では、上記事前にトレーニングされた言語モデルは、リカレントニューラルネットワーク (Recurrent Neural Network、RNN)、長期短期記憶 (Long Short-Term Memory、LSTM) ネットワーク、ゲート付きリカレントユニット (Gated Recurrent Unit、GRU) ネットワーク、BiLSTM ネットワークなどのいずれか1つとして選択されてよい。該GRU ネットワークは、LSTM ネットワークの効果の高い変形であり、該BiLSTM ネットワークは、順方向LSTM と逆方向LSTM を組み合わせて形成される。

【0036】

なお、該言語モデルのトレーニングプロセスに対して、関連方式を採用し、かつ歴史検索プロセスにおいて入力されたクエリプレフィックスの内容をトレーニングサンプルとしてトレーニングを行うことができ、本実施例は、これを限定しない。具体的なトレーニングプロセスには、パープレキシティ (perplexity、ppl) を用いて言語モデルの良否を判定することができ、tensorflow、paddle、PyTorchなどのトレーニングプラットフォームを用いることができる。事前にトレーニングされた言語モデルを用いてクエリプレフィックスを評価するとき、pplを、該クエリプレフィックスが通じるか否かの定量化の根拠として採用することができ、主に各単語に基づいて文の出現確率を推定し、かつ文の長さを正規化処理 (normalize) する。pplに関する式は、以下のとおりである：

【数3】

$$ppl = \sqrt[N]{\prod_{i=1}^N \frac{1}{p(w_i | w_1 w_2 \dots w_{i-1})}}$$

ここで、Nは、クエリプレフィックスの長さを表し、 $p(w_i)$ は、i番目の単語の確率を表し、 $p(w_i | w_1 w_2 \dots w_{i-1})$ は、最初のi-1個の単語に基づいて算出されたi番目の単語の確率を表す。一般的には、ppl値が小さいほど、クエリプレフィックスのパープレキシティが低く、通じる程度が高いことを示す。

【0037】

ステップ103では、上記パープレキシティが所定の閾値よりも小さいか否かを判断する。

【0038】

本実施例では、上記所定の閾値は、実際の状況に応じて事前に設定することができる。クエリプレフィックスのパープレキシティが所定の閾値よりも小さい場合に、該クエリプレフィックスの通じる程度が高いことを示し、該クエリプレフィックスに基づいてプリフェッチ要求を開始することができる。該クエリプレフィックスのパープレキシティが所定の閾値以上である場合に、該クエリプレフィックスのパープレキシティが高く、かつ通じる程度が低いことを示し、この場合に該クエリプレフィックスに基づいてプリフェッチ要求を開始する意味が大きくなり、対応するプリフェッチ要求を直接フィルタリングして除去することができる。

【0039】

ステップ104では、上記パープレキシティが所定の閾値よりも小さい場合に、上記クエリプレフィックスに基づいてプリフェッチ要求を送信する。

【0040】

本実施例では、プリフェッチ要求を送信することは、プリフェッチ要求を検索サーバなどのサーバに送信して、該サーバの検索サービスを呼び出して、現在のクエリプレフィックスに対応するプリフェッチ結果を取得することによってよい。

【0041】

好ましくは、上記ステップ104の後に、上記方法は、サーバが上記プリフェッチ要求に基づいて返されたクエリプレフィックスに対応するプリフェッチ結果を受信することをさ

10

20

30

40

50

らに含む。さらに、プリフェッチ結果を受信した後、電子機器は、プリフェッチ結果を表示して、取得されたプリフェッチ結果をユーザに提示し、ユーザに検索速度がより速い体験をユーザに与える。

【0042】

本願の実施例に係るデータプリフェッチ方法は、クエリプレフィックスのパープレキシティの判別に基づいてプリフェッチ要求を送信することにより、パープレキシティの高いプリフェッチ要求を直接フィルタリングして除去することができるため、プリフェッチの成功率を向上させ、かつ過剰なプリフェッチ要求によるバックエンドサーバの機器コストをさらに低減することができるとともに、システム安定性に影響を与えることと、過剰なプリフェッチ結果の提示によりユーザに視覚体験上の干渉を与えることとを避け、ユーザ体験を向上させることができる。

10

【0043】

具体的な実際の応用において、関連技術における、補完マッチングに基づいてプリフェッチ要求を発行する方式に比べて、本願の実施例のデータプリフェッチ方式は、プリフェッチ成功率を26.5%から45%まで向上させることができるため、速度体験効果を低減しない前提で40%のプリフェッチトラフィック配信を低減することができ、約1000台のサーバのコストを節減する。

【0044】

本願の実施例では、クエリプレフィックスに句読点などの特殊文字が含まれる場合に、該特殊文字は、クエリプレフィックスのパープレキシティに大きな影響を与えるため、特殊文字による影響を低減するために、判別して取得されたパープレキシティを最適化することができる。

20

【0045】

好ましくは、取得されたクエリプレフィックスに特殊文字が含まれる場合に、上記方法は上記クエリプレフィックスのパープレキシティ補正係数を取得することをさらに含む。さらに、上記ステップ103は、上記パープレキシティ補正係数を用いて上記パープレキシティを補正することと、補正後のパープレキシティが上記所定の閾値よりも小さいか否かを判断することと、を含む。その後、上記補正後のパープレキシティが上記所定の閾値よりも小さい場合に、上記クエリプレフィックスに基づいて上記プリフェッチ要求を送信する。このように、この補正プロセスにより、特殊文字による、対応するクエリプレフィックスのパープレキシティへの影響を低減することにより、プリフェッチの成功率をさらに向上させることができる。

30

【0046】

上記パープレキシティ補正係数と特殊文字との関係は、事前に設定することができる。一実施形態では、上記パープレキシティ補正係数は、特殊文字の有無に関連することができる。例えば、クエリプレフィックスに特殊文字が含まれれば、パープレキシティ補正係数、例えば、Xが存在し、このとき、該パープレキシティ補正係数Xを用いて、判別して取得された該クエリプレフィックスのパープレキシティを補正する必要がある。一方、クエリプレフィックスに特殊文字が含まれなければ、パープレキシティ補正係数が存在せず、判別して取得された該クエリプレフィックスのパープレキシティを補正する必要がある。

40

【0047】

別の実施形態では、上記パープレキシティ補正係数は、特殊文字の個数、位置などに関連することができる。例えば、上記パープレキシティ補正係数が特殊文字の個数に関連すれば、下式を採用してクエリプレフィックスのパープレキシティ補正係数 R_e を計算することができる。

【数4】

$$R_e = \frac{f}{N} \log_2 \text{count}(sw)$$

50

ここで、 N は、該クエリプレフィックスの文の長さ（例えば、文字の長さであると理解できる）を表し、 $\text{count}(sw)$ は、該クエリプレフィックスに含まれる特殊文字の個数を表し、 f は、所定の係数を表し、例えば、1などとして設定することができる。

【0048】

このように、事前にトレーニングされた言語モデルに基づいて判別して取得されたクエリプレフィックスのパープレキシティが ppl_0 であり、対応するパープレキシティ補正係数が R_e であれば、補正後のパープレキシティ PPL は、 $\text{PPL} = \text{ppl}_0 + R_e$ である。

【0049】

本願の実施例では、クエリプレフィックスに対する判別精度を向上させるために、事前にトレーニングされた言語モデルに基づいてクエリプレフィックスを判別する前に、まず、該クエリプレフィックスを処理して、対応する、中国語の語意の理解力がより強い単語埋め込み (word embedding) を生成し、かつ該単語埋め込みに基づいてパープレキシティを判別する。

10

【0050】

好ましくは、図2に示すように、上記ステップ102は、以下の3つのことを含んでよい。まず、取得されたクエリプレフィックスを分割して、複数の分割単語を取得し、該分割方式について、例えば、分割単語に基づくアイテム item 分割を用いて、 item_1 、 item_2 、 item_3 などを取得し、次に、各分割単語を事前にトレーニングされた語意モデルにそれぞれ入力して、各分割単語の単語埋め込みを生成し、各分割単語の単語埋め込みに基づいて該クエリプレフィックスの単語埋め込みを決定し、該事前にトレーニングされた語意モデルは、 EMLo 、 GPT 、 BERT 、 XLNet 、 ERNIE などを用いてもよく、例えば、図2に示す tensor_1 、 tensor_2 及び tensor_3 であり、最後に、該クエリプレフィックスの単語埋め込みを事前にトレーニングされた言語モデルに入力して、該クエリプレフィックスのパープレキシティを取得する。このように、クエリプレフィックスの単語埋め込みに基づいてクエリプレフィックスのパープレキシティを判別することにより、判別精度を向上させることができる。

20

【0051】

一実施形態では、各分割単語の単語埋め込みに基づいて、該クエリプレフィックスの単語埋め込みを決定するとき、各単語分割の単語埋め込みを結合して、該クエリプレフィックスの単語埋め込みを取得することができる。

30

【0052】

以下、図3を参照しながら本願の具体的な実施例に係るプリフェッチ要求の発行の論理を説明する。

【0053】

本願の具体的な実施例では、図3に示すように、対応するプリフェッチ要求の発行の論理は、以下の場合1～場合4を含む。

【0054】

場合1、 $1 \rightarrow 2 \rightarrow 3 \rightarrow 4$ であり、ここで、「 $1(\text{mod } 1)$ 」は、クライアントがユーザによって入力されたクエリプレフィックスを受信した後に、プリフェッチ要求を発行するように準備することを表し、「2」は、事前にトレーニングされた言語モデルに基づいてクエリプレフィックスを判別し、かつ取得されたパープレキシティが所定の閾値よりも小さいか否かを判断することを表し、「3」は、判断結果1、即ち、取得されたパープレキシティが所定の閾値以上であることを取得することを表し、「4」は、プリフェッチ要求を発行しないことを表す。

40

【0055】

場合2、 $1 \rightarrow 2 \rightarrow 3 \rightarrow 5 \rightarrow 6 \rightarrow 7$ であり、ここで、「 $1(\text{mod } 1)$ 」は、クライアントがユーザによって入力されたクエリプレフィックスを受信した後に、プリフェッチ要求を発行するように準備することを表し、「2」は、事前にトレーニングされた言語モデルに基づいてクエリプレフィックスを判別し、かつ取得されたパープレキシティが

50

所定の閾値よりも小さいか否かを判断することを表し、「3」は、判断結果2、即ち、取得されたパープレキシティが所定の閾値よりも小さいことを取得することを表し、「5」は、サーバにプリフェッチ要求を発行して、サーバの検索サービスを呼び出すことを表す。「6及び7」は、サーバがプリフェッチ結果をフィードバックすることを表す。

【0056】

場合3、プリフェッチ要求が成功し、 $8 \pmod{2}$ は、クライアントがサーバにフィードバックするプリフェッチの成功フラグを表す。

【0057】

場合4、プリフェッチ要求が失敗し、 $9 \rightarrow 10 \pmod{0}$ は、プリフェッチが失敗した後に、ユーザが直接開始した検索要求を表す。

【0058】

図4を参照すると、本願の実施例に係るデータプリフェッチデバイスは、電子機器に適用され、図4に示すように、該データプリフェッチデバイス40は、

ユーザが入力したクエリプレフィックスを取得する第1の取得モジュール41と、事前にトレーニングされた言語モデルに基づいて上記クエリプレフィックスを判別して、上記クエリプレフィックスのパープレキシティを取得する判別モジュール42と、上記パープレキシティが所定の閾値よりも小さいか否かを判断する判断モジュール43と、上記パープレキシティが上記所定の閾値よりも小さい場合に、上記クエリプレフィックスに基づいてプリフェッチ要求を送信する送信モジュール44と、を含む。

【0059】

好ましくは、上記デバイスは、

上記クエリプレフィックスに特殊文字が含まれる場合に、上記クエリプレフィックスのパープレキシティ補正係数を取得する第2の取得モジュールをさらに含む。

【0060】

上記判断モジュール43は、

上記パープレキシティ補正係数を用いて上記パープレキシティを補正する補正ユニットと、補正後のパープレキシティが上記所定の閾値よりも小さいか否かを判断する判断ユニットと、を含む。

【0061】

上記送信モジュール44は、具体的には、上記補正後のパープレキシティが上記所定の閾値よりも小さい場合に、上記クエリプレフィックスに基づいて上記プリフェッチ要求を送信する。

【0062】

好ましくは、前記第2の取得モジュールは、具体的には、

下式を採用して上記クエリプレフィックスのパープレキシティ補正係数 R_e を計算し、

【数5】

$$R_e = \frac{f}{N} \log_2 \text{count}(sw)$$

ここで、 N は、上記クエリプレフィックスの文の長さを表し、 $\text{count}(sw)$ は、上記クエリプレフィックスに含まれる特殊文字の個数を表し、 f は、所定の係数を表す。

【0063】

好ましくは、上記判別モジュール42は、

上記クエリプレフィックスを分割して、複数の分割単語を取得する分割ユニットと、各上記分割単語を事前にトレーニングされた語意モデルにそれぞれ入力して、各上記分割単語の単語埋め込みを生成する生成ユニットと、各上記分割単語の単語埋め込みに基づいて上記クエリプレフィックスの単語埋め込みを決定する決定ユニットと、

上記クエリプレフィックスの単語埋め込みを上記事前にトレーニングされた言語モデルに

10

20

30

40

50

入力して、上記クエリプレフィックスのパーレキシティを取得する判別ユニットと、を含む。

【0064】

好ましくは、上記決定ユニットは、具体的には、各上記分割単語の単語埋め込みを結合して、上記クエリプレフィックスの単語埋め込みを取得する。

【0065】

好ましくは、上記デバイスは、

サーバが上記プリフェッチ要求に基づいて返された上記クエリプレフィックスに対応するプリフェッチ結果を受信する受信モジュールをさらに含む。

【0066】

本実施例に係るデータプリフェッチデバイス40は、上記図1に示す実施例において実現された各プロセスを実現し、かつ同様の効果を達成することができ、重複を避けるため、ここでは説明を省略する。

【0067】

本願の実施例によれば、本願は、電子機器と可読記憶媒体をさらに提供する。

【0068】

図5に示すように、本願の実施例に係るデータプリフェッチ方法を実現する電子機器のブロック図である。電子機器は、ラップトップコンピュータ、デスクトップコンピュータ、ワークステーション、パーソナルデジタルアシスタント、サーバ、ブレードサーバ、大型コンピュータ、及びその他の適切なコンピュータなどの様々な形態のデジタルコンピュータを表すことを意図する。電子機器は、パーソナルデジタルアシスタント、セルラー電話、スマートフォン、ウェアラブルデバイス及びその他の類似するコンピューティングデバイスなどの様々な形態のモバイルデバイスを表すことができる。本明細書に示すコンポーネント、それらの接続及び関係と、それらの機能とは、例示的なものに過ぎず、本明細書で説明及び/又は要求された本願の実現を限定することを意図するものではない。

【0069】

図5に示すように、該電子機器は、1つ以上のプロセッサ501と、メモリ502と、高速インタフェース及び低速インタフェースを含む、各コンポーネントを接続するインタフェースと、を含む。各コンポーネントは、異なるバスを介して互いに接続され、かつ共通マザーボード上に取り付けられてもよく、必要に応じて他の方式で取り付けられてもよい。プロセッサは、電子機器内で実行された、外部入力/出力装置(例えば、インタフェースに結合された表示機器)上にGUIのグラフィック情報を表示するようにメモリ内又はメモリ上に記憶されている命令を含む命令を処理することができる。他の実施形態では、必要があれば、複数のプロセッサ及び/又は複数のバスを、複数のメモリと共に使用することができる。同様に、それぞれが必要な動作を提供する複数の電子機器(例えば、サーバレイ、ブレードサーバのグループ又はマルチプロセッサシステム)を接続することができる。図5において、プロセッサ501を例とする。

【0070】

メモリ502は、本願に係る非一時的コンピュータ可読記憶媒体である。上記メモリには、少なくとも1つのプロセッサに実行可能で、前記少なくとも1つのプロセッサに本願に係るデータプリフェッチ方法を実行させる命令が記憶されている。本願の非一時的コンピュータ可読記憶媒体には、本願に係るデータプリフェッチ方法をコンピュータに実行させるコンピュータ命令が記憶されている。

【0071】

メモリ502は、非一時的コンピュータ可読記憶媒体として、非一時的ソフトウェアプログラム、非一時的コンピュータ実行可能なプログラム及びモジュール、例えば、本願の実施例におけるデータプリフェッチ方法に対応するプログラム命令/モジュール(例えば、図4に示す第1の取得モジュール41、判別モジュール42、判断モジュール43及び送信モジュール44)を記憶することができる。プロセッサ501は、メモリ502内に記憶されている非一時的ソフトウェアプログラム、命令及びモジュールを実行することによ

10

20

30

40

50

り、サーバの様々な機能アプリケーション及びデータ処理を実行し、即ち、上記方法の実施例におけるデータプリフェッチ方法を実現する。

【0072】

メモリ502は、オペレーティングシステム、少なくとも1つの機能に必要なアプリケーションプログラムを記憶できるプログラム記憶領域と、電子機器の使用中に作成されたデータなどを記憶できるデータ記憶領域とを含んでよい。また、メモリ502は、高速ランダムアクセスメモリを含んでもよく、少なくとも1つの磁気ディスクメモリデバイス、フラッシュメモリデバイス、又は他の非一時的固体メモリデバイスなどの非一時的メモリを含んでもよい。幾つかの実施例では、メモリ502は、好ましくは、プロセッサ501に対して遠隔に配置されたメモリを含み、これらのリモートメモリは、ネットワークを介して電子機器に接続することができる。上記ネットワークの例は、インターネット、企業イントラネット、ローカルエリアネットワーク、モバイル通信ネットワーク及びそれらの組み合わせを含むが、これらに限定されない。

10

【0073】

データプリフェッチ方法の電子機器は、入力装置503及び出力装置504をさらに含んでよい。プロセッサ501、メモリ502、入力装置503及び出力装置504は、バス又は他の方式で接続することができ、図5において、バスによる接続を例とする。

【0074】

入力装置503は、入力された数字又は文字情報を受信するとともに、データプリフェッチ方法の電子機器のユーザ設定及び機能制御に関連するキー信号入力を生成することができ、入力装置は、例えば、タッチスクリーン、キーボード、マウス、トラックパッド、タッチパッド、ポインティングスティック、1つ以上のマウスボタン、トラックボール、ジョイスティックである。出力装置504は、表示機器、補助照明装置（例えば、LED）及び触覚フィードバック装置（例えば、振動モータ）などを含んでよい。該表示機器は、液晶ディスプレイ（LCD）、発光ダイオード（LED）ディスプレイ及びプラズマディスプレイなどを含んでよいが、これらに限定されない。幾つかの実施形態では、表示機器は、タッチスクリーンであってよい。

20

【0075】

本明細書で説明されたシステム及び技術の各実施形態は、デジタル電子回路システム、集積回路システム、専用ASIC（特定用途向け集積回路）、コンピュータハードウェア、ファームウェア、ソフトウェア、及び/又はそれらの組み合わせにおいて実現することができる。これらの様々な実施形態は、少なくとも1つのプログラマブルプロセッサを含むプログラマブルシステム上で実行及び/又は解釈できる1つ以上のコンピュータプログラムにおける実施を含んでもよく、該プログラマブルプロセッサは、専用又は汎用のプログラマブルプロセッサであってもよく、記憶システム、少なくとも1つの入力装置及び少なくとも1つの出力装置からデータと命令を受信し、かつデータと命令を該記憶システム、該少なくとも1つの入力装置及び該少なくとも1つの出力装置に伝送することができる。

30

【0076】

これらのコンピュータプログラム（プログラム、ソフトウェア、ソフトウェアアプリケーション、又はコードとも呼ばれる）は、プログラマブルプロセッサの機械命令を含み、かつこれらのコンピュータプログラムを、高レベルのプロセス及び/又はオブジェクト指向のプログラミング言語、及び/又はアセンブリ/機械言語を用いて実施することができる。本明細書で使用された用語「機械可読媒体」及び「コンピュータ可読媒体」とは、機械命令及び/又はデータをプログラマブルプロセッサに提供する任意のコンピュータプログラム製品、デバイス及び/又は装置（例えば、磁気ディスク、光ディスク、メモリ、プログラマブルロジック装置（PLD））を指し、機械可読信号としての機械命令を受信する機械可読媒体を含む。用語「機械可読信号」とは、機械命令及び/又はデータをプログラマブルプロセッサに提供する任意の信号を指す。

40

【0077】

ユーザとの対話を提供するために、本明細書で説明されたシステム及び技術をコンピュー

50

タ上で実施することができ、該コンピュータは、ユーザに情報を表示する表示装置（例えば、CRT（ブラウン管）又はLCD（液晶ディスプレイ）モニター）と、キーボードと、ポインティング装置（例えば、マウス又はトラックボール）とを有し、ユーザは、該キーボード及び該ポインティング装置により入力をコンピュータに提供することができる。他の種類の装置は、ユーザとの対話を提供することができ、例えば、ユーザに提供されるフィードバックは、任意の形式の感覚フィードバック（例えば、視覚的フィードバック、聴覚的フィードバック又は触覚フィードバック）であってよく、任意の形式（音響入力、音声入力又は触覚入力を含む）でユーザからの入力を受信することができる。

【0078】

本明細書で説明されたシステム及び技術は、バックエンドコンポーネントを含むコンピュータシステム（例えば、データサーバとして）、又はミドルウェアコンポーネントを含むコンピュータシステム（例えば、アプリケーションサーバ）、又はフロントエンドコンポーネントを含むコンピュータシステム（例えば、ユーザが本明細書で説明されたシステム及び技術の実施形態と対話できるグラフィカルユーザインタフェース又はウェブブラウザを有するユーザコンピュータ）、又はこのようなバックグラウンドコンポーネント、ミドルウェアコンポーネント又はフロントエンドコンポーネントの任意の組み合わせを含むコンピュータシステムにおいて実施することができる。システムのコンポーネントは、任意の形式又は媒体のデジタルデータ通信（例えば、通信ネットワーク）を介して互いに接続することができる。通信ネットワークの例は、ローカルエリアネットワーク（LAN）、広域ネットワーク（WAN）及びインターネットを含む。

【0079】

コンピュータシステムは、クライアント及びサーバを含んでよい。クライアントとサーバは、一般的に、互いに離れ、通常、通信ネットワークを介して対話する。クライアントとサーバとの関係は、対応するコンピュータ上で実行し、互いにクライアント-サーバの関係性を有するコンピュータプログラムによって生成される。

【0080】

本願の実施例の技術手段によれば、パーレキシティの高いプリフェッチ要求を直接フィルタリングして除去することができるため、プリフェッチの成功率を向上させ、かつ過剰なプリフェッチ要求によるバックエンドサーバの機器コストをさらに低減することができる。とともに、システム安定性に影響を与えることと、過剰なプリフェッチ結果の提示によりユーザに視覚体験上の干渉を与えることとを避け、ユーザ体験を向上させることができる。

【0081】

なお、上記様々な形式のフローを使用して、ステップを改めて順序付けたり、追加したり、削除したりしてよい。例えば、本願において開示されている技術手段の所望の結果を実現できる限り、本願に記載された各ステップは、並列的に実行されてもよく、順次実行されてもよく、異なる順序で実行されてもよく、本明細書はこれを限定しない。

【0082】

当業者は、上記実施例の方法におけるフローの全部又は一部の実現が、コンピュータプログラムによって関連ハードウェアを制御して完成できることを理解すべきであり、上記プログラムはコンピュータ可読記憶媒体に記憶することができ、該プログラムが実行されると、上記各方法の実施例のフローを含んでよい。上記記憶媒体は、磁気ディスク、光ディスク、読み取り専用メモリ（Read-Only Memory、ROM）又はランダムアクセスメモリ（Random Access Memory、RAM）などであってよい。

【0083】

本開示の幾つかの実施例で説明されたこれらの実施例は、ハードウェア、ソフトウェア、ファームウェア、ミドルウェア、マイクロコード又はそれらの組み合わせによって実現できることを理解されたい。ハードウェアによって実現される場合、モジュール、ユニット、サブモジュール、サブユニットなどは、1つ以上の特定用途向け集積回路（Application Specific Integrated Circuits、ASIC）

10

20

30

40

50

、デジタルシグナルプロセッサ (Digital Signal Processing、DSP)、デジタル信号処理装置 (DSP Device、DSPD)、プログラマブルロジックデバイス (Programmable Logic Device、PLD)、フィールドプログラマブルゲートアレイ (Field-Programmable Gate Array、FPGA)、汎用プロセッサ、コントローラ、マイクロコントローラ、マイクロプロセッサ、本願に記載の機能を実行する他の電子ユニット又はそれらの組み合わせにおいて実現することができる。

【0084】

ソフトウェアによって実現される場合、本開示の幾つかの実施例に記載の機能を実行するモジュール (例えば、プロセス、関数など) により、本開示の幾つかの実施例に記載の技術を実現することができる。ソフトウェアコードは、メモリに記憶し、かつプロセッサにより実行することができる。メモリは、プロセッサ内又はプロセッサの外部で実現されてよい。

10

【0085】

上記具体的な実施形態は、本願の保護範囲を限定するものではない。当業者は、設計要件及びその他の要因に応じて、様々な修正、組み合わせ、サブ組み合わせ及び置換を行うことができることを、理解すべきである。本願の精神及び原則の範囲内で行われる修正、同等置換及び改善などは、いずれも本願の保護範囲に含まれるべきである。

【0086】

本願は、2020年1月10日に提出された中国特許出願第202010024434.0号の優先権を主張するものであり、その全ての内容は参照により本願に取り込まれるものとする。

20

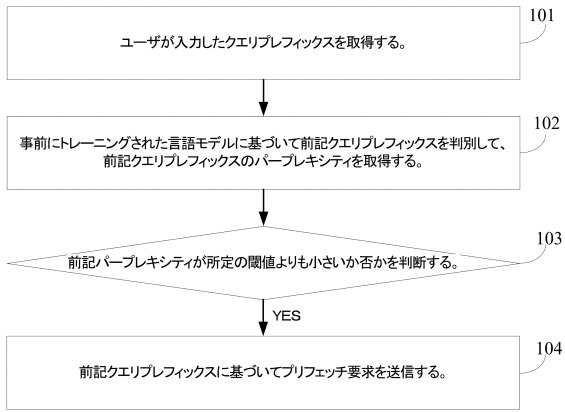
30

40

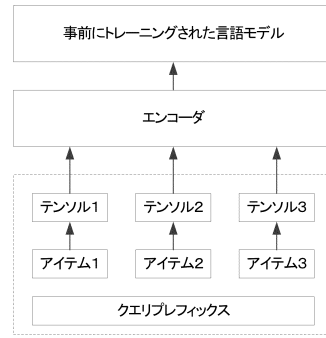
50

【図面】

【図 1】

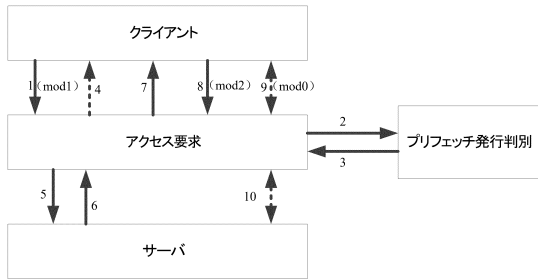


【図 2】

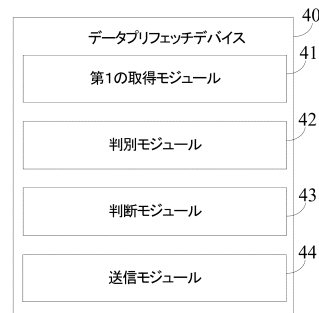


10

【図 3】

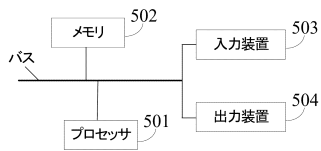


【図 4】



20

【図 5】



30

40

50

フロントページの続き

淀区上地十街10号百度大厦三層

(72)発明者 範 彪

中華人民共和国 100085 北京市海淀区上地十街10号百度大厦三層

審査官 原 秀人

(56)参考文献 特開2018-206361(JP, A)

国際公開第2019/198386(WO, A1)

米国特許出願公開第2019/0278870(US, A1)

中村 明 外, 複数モデルの統合によるLDAトピックモデルの高精度化とテキスト入力支援への応用, 情報処理学会論文誌 論文誌ジャーナル Vol. 50 No. 4 [CD-ROM], 日本, 社団法人情報処理学会, 2009年04月15日, Vol. 50 No. 4, pp. 1375--1389

鷹合 基行 外, 読影レポートを対象とした予測入力システム, 第72回(平成22年)全国大会講演論文集(2) 人工知能と認知科学, 日本, 社団法人情報処理学会, 2010年03月20日, pp. 2-777~2-778

(58)調査した分野 (Int.Cl., DB名)

G06F 16/00 - 16/958