【公報種別】特許法第１７条の２の規定による補正の掲載
【部門区分】第６部門第３区分
【発行日】平成21年8月27日(2009.8.27)

【公開番号】特開2008-140357(P2008-140357A)
【公開日】平成20年6月19日(2008.6.19)
【年通号数】公開・登録公報2008-024
【出願番号】特願2007-70697(P2007-70697)
【国際特許分類】
Ｇ０６Ｆ　12/00　　(2006.01)
Ｇ０６Ｆ　17/30　　(2006.01)
【ＦＩ】
Ｇ０６Ｆ　12/00　　５２０Ａ
Ｇ０６Ｆ　17/30　　１７０Ａ
Ｇ０６Ｆ　17/30　　４１４Ｂ

【手続補正書】
【提出日】平成21年7月9日(2009.7.9)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
記憶装置に入力された複数の電子文書を含む文書集合中の個々の電子文書から単語あるいは連続する複数の文字からなる索引語を切り出して記憶装置に格納する工程と、
それぞれの索引語が各電子文書中に出現する回数を取得する工程と、
前記文書集合中の各電子文書にそれぞれ異なる文書識別番号を割り当てる工程と、
各々の索引語について得られた文書識別番号と索引語出現回数の組であるポスティングをデータ圧縮する工程とをコンピュータに実行させると共に、
前記ポスティングをデータ圧縮する工程は、各ポスティング中の電子文書の文書識別番号を可変長のバイト列で表現し、ｗを１以上８以下の与えられた整数とするとき、前記バイト列中のｗビットによって１から（$2^{W}-1$）までの索引語出現回数を表現し、2^{W}以上の索引語出現回数を追加のバイト列を用いて表現する
ことを特徴とするプログラム。
【請求項２】
請求項１記載のプログラムにおいて、ｘを、ｘ＋ｗが１以上８以下となる与えられた整数とする場合、前記ポスティングを表現するバイト列中のｘビットに追加情報を書き込む
ことを特徴とするプログラム。
【請求項３】
請求項１又は２記載のプログラムにおいて、
前記ポスティングをデータ圧縮する工程は、各バイト中の特定の位置のビットにより電子文書の文書識別番号を表現するバイト列の最終バイトか否かを表現すると共に、前記追加のバイト列は、各バイト中の前記特定の位置のビットにより電子文書の文書識別番号を表現する最終バイトではないことを表現し、前記特定の位置のビットに隣接するビットにより当該追加のバイト列の最終バイトか否かを表現する
ことを特徴とするプログラム。
【請求項４】
複数の電子文書を含む文書集合を記憶装置に入力する工程と、

前記入力された文書集合中の個々の電子文書から単語あるいは連続する複数の文字からなる索引語を切り出す工程と、
それぞれの索引語が各電子文書中に出現する回数を取得する工程と、
前記文書集合中の各電子文書にそれぞれ異なる文書識別番号を割り当てる工程と、
各々の索引語について得られた文書識別番号と索引語出現回数からなるポスティングをデータ圧縮する工程と、
データ圧縮されたポスティングを含む転置インデックスを出力する工程とを有し、
前記ポスティングをデータ圧縮する工程では、各ポスティング中の電子文書の文書識別番号を可変長のバイト列で表現し、ｗを１以上８以下の与えられた整数とするとき、索引語出現回数が１から（$2^{W}-1$）までのときは当該索引語出現回数を前記バイト列中のｗビットによって表現し、索引語出現回数が2^{W}以上のときは当該索引語出現回数を追加のバイト列を用いて表現する
ことを特徴とする転置インデックス作成方法。
【請求項５】
請求項４記載の転置インデックス作成方法において、ｘを、ｘ＋ｗが１以上８以下となる与えられた整数とする場合、前記ポスティングを表現するバイト列中のｘビットに追加情報を書き込む
ことを特徴とする転置インデックス作成方法。
【請求項６】
請求項４又は５記載の転置インデックス作成方法において、
前記ポスティングをデータ圧縮する工程は、各バイト中の特定の位置のビットにより電子文書の文書識別番号を表現するバイト列の最終バイトか否かを表現すると共に、前記追加のバイト列は、各バイト中の前記特定の位置のビットにより電子文書の文書識別番号を表現する最終バイトではないことを表現し、前記特定の位置のビットに隣接するビットにより当該追加のバイト列の最終バイトか否かを表現する
ことを特徴とする転置インデックス作成方法。
【請求項７】
それぞれに識別番号が割り当てられた複数の電子文書からなる文書集合の全ての索引語について、前記索引語が出現する電子文書の識別番号と前記電子文書中での当該索引語の出現回数の組であるポスティングを格納した転置インデックスを用いた検索手段による検索方法であって、
前記転置インデックスは、各ポスティング中の電子文書の識別番号が可変長のバイト列としてコンピュータにより作成され、ｗを１以上８以下の与えられた整数とするとき、コンピュータは、前記バイト列中のｗビットによって１から（$2^{W}-1$）までの出現回数を表現し、2^{W}以上の出現回数を追加のバイト列を用いて表現し、前記電子文書の識別番号を表現するバイト列は、各バイト中の特定の位置のビットにより当該バイト列の最終バイトか否かを表現し、前記追加のバイト列は、各バイト中の前記特定の位置のビットにより文書識別番号の最終バイトではないことを表現し、前記特定の位置のビットに隣接するビットにより当該追加のバイト列の最終バイトか否かを表現しているとき、
入力部から第１の索引語と第２の索引語を受け付ける工程と、
前記転置インデックス中の前記第１の索引語の転置リストに含まれる電子文書の識別番号を取得する工程と、
前記取得した識別番号が前記第２の索引語の転置リストに含まれているかどうかを２分探索によって判定する工程と、
前記取得した識別番号のうち前記第２の索引語の転置リストに含まれている識別番号またはその識別番号に対応する電子文書を出力する工程とを有し、
前記２分探索に際しては、前記第２の索引語の転置リストの所望の読み位置のバイトに移動し、当該バイトが前記電子文書の識別番号を表現するバイト列の最終バイトか否かを判定し、最終バイトでない場合には最終バイトが見つかるまで前記転置リストの先頭の方向に１バイトずつ読み進み、前記電子文書の識別番号を表現するバイト列の最終バイト中

の出現回数を表現するｗビットの情報から前記出現回数を表現する追加のバイト列が存在するか否かを判定し、追加のバイト列が存在しないと判定されたときは当該バイトの次のバイトをポスティングの先頭としてポスティングを読み取り当該ポスティング中の識別番号を取得し、追加のバイト列が存在すると判定されたときは追加のバイト列の最終バイトが見つかるまで前記転置リストの下流側に１バイトずつ読み進み、追加のバイト列の最終バイトの次のバイトをポスティングの先頭としてポスティングを読み取り当該ポスティング中の識別番号を取得する処理を実行する
　ことを特徴とする検索方法。
【手続補正２】
【補正対象書類名】明細書
【補正対象項目名】発明の名称
【補正方法】変更
【補正の内容】
【発明の名称】プログラム、転置インデックスの格納方法及び検索方法