(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4845420号
(P4845420)
(45)発行日　平成23年12月28日(2011.12.28)　(24)登録日　平成23年10月21日(2011.10.21)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 350C
G06F 17/30 370Z

請求項の数 8　外国語出願　(全 16 頁)

(21)出願番号　特願2005-134488 (P2005-134488)
(22)出願日　平成17年5月2日 (2005.5.2)
(65)公開番号　特開2005-322244 (P2005-322244A)
(43)公開日　平成17年11月17日 (2005.11.17)
審査請求日　平成20年4月30日 (2008.4.30)
(31)優先権主張番号　10/837,540
(32)優先日　平成16年4月30日 (2004.4.30)
(33)優先権主張国　米国 (US)

(73)特許権者　500046438
マイクロソフト　コーポレーション
アメリカ合衆国　ワシントン州　９８０５２－６３９９　レッドモンド　ワン　マイクロソフト　ウェイ
(74)代理人　100077481
弁理士　谷　義一
(74)代理人　100088915
弁理士　阿部　和夫
(72)発明者　ベンユー　チャン
アメリカ合衆国　９８０５２　ワシントン州　レッドモンド　ワン　マイクロソフト　ウェイ　マイクロソフト　コーポレーション内

最終頁に続く

(54)【発明の名称】情報の多様性および豊富さを向上させるよう検索結果のドキュメントを順位付ける方法およびシステム

(57)【特許請求の範囲】
【請求項１】
検索結果のドキュメントを順位付けるコンピューティングデバイスであって、
コンピュータ実行可能な命令を有するメモリと、
前記メモリに格納されたコンピュータ実行可能な命令を実行する中央処理装置と
を備え、前記コンピュータ実行可能な命令は、
問い合わせをユーザから受信するステップと、
前記受信された問い合わせに対して検索結果としてのドキュメントを識別するステップと、
前記識別されたドキュメントのキーワードを識別するステップと、
前記検索結果の識別されたドキュメントの各組に対して、組のうちのドキュメントが有する、前記組のうちのもう一方のドキュメントに対する類似度を計算するステップであって、前記類似度は、次の式
【数１】

$$aff(d_i, d_j) = \frac{\vec{d}_i \cdot \vec{d}_j}{\|\vec{d}_i\|}$$

に基づいて計算され、ただし、ａｆｆ（ｄ$_i$，ｄ$_j$）はドキュメントｄ$_i$のドキュメントｄ$_j$に対する類似度であり、ドキュメントｄ$_i$は前記組のうちのドキュメントを表して、ドキュメントｄ$_j$は前記組のうちのもう一方のドキュメントを表して、

【数２】

$\vec{d}_i$

はドキュメントｄ$_i$のベクトルを表して、

【数３】

$\vec{d}_j$

はドキュメントｄ$_j$のベクトルを表して、および

【数４】

$\left\|\vec{d}_i\right\|$

はドキュメントｄ$_i$のベクトルの長さを表して、各ベクトルは、対応するドキュメントの識別された少なくとも１つのキーワードに対する少なくとも１つのエントリを有する、ステップと、

前記検索結果の識別された各ドキュメントに対して、類似度の行列から得られる正規化された類似度の行列の要素に基づいて、前記識別されたドキュメントに対する情報の豊富さを計算するステップであって、前記情報の豊富さは、次の式

【数５】

$$InfoRich(d_i) = \sum_{all\, j \neq i} InfoRich(d_j) \cdot \tilde{M}_{ji}$$

に基づいて計算され、ただし、ＩｎｆｏＲｉｃｈ（ｄ$_i$）はドキュメントｄ$_i$の情報の豊富さであり、ドキュメントｄ$_i$は前記検索結果の識別されたドキュメントを表して、ドキュメントｄ$_j$は前記検索結果における別のドキュメントを表して、

【数６】

$\tilde{M}_{ji}$

は前記正規化された類似度の行列の要素であり、前記類似度の行列は前記検索結果における識別されたドキュメントおよび別のドキュメントの計算された類似度を含む、ステップと、

前記検索結果のドキュメントについて計算された情報の豊富さに基づいて順位付けされた検索結果のドキュメントの表示を表示装置に出力するステップと

を含むことを特徴とするコンピューティングデバイス。

【請求項２】

前記ドキュメントを識別するステップは、

検索エンジンサービスへ前記問い合わせを送信するステップと、

検索結果としてのドキュメントを受信するステップと

を含むことを特徴とする請求項１に記載のコンピューティングデバイス。

【請求項３】

前記ドキュメントは、ウェブページであることを特徴とする請求項１に記載のコンピューティングデバイス。

【請求項４】

コンピュータシステムにおいて、ドキュメントの集まり内のドキュメントの情報の豊富さを中央処理装置およびメモリにより計算する方法であって、

前記集まり内の各ドキュメントが有する、前記集まり内の別のドキュメントに対する類似度を前記中央処理装置が識別するステップであって、前記類似度は、前記集まり内のドキュメントであるドキュメントの組のうちの１つのドキュメント、および前記集まり内の別のドキュメントである前記組のうちのもう一方のドキュメントを含むドキュメントの各組に対して識別されて、１つのドキュメントの情報が別のドキュメントの情報にどの程度含まれているかを示す、ステップと、

前記集まり内の各ドキュメントに対して、類似度の行列から得られる正規化された類似度の行列の要素に基づいて前記情報の豊富さを前記中央処理装置が決定するステップと、

前記集まり内のドキュメントの前記決定された情報の豊富さを前記メモリに前記中央処理装置が格納するステップと、

前記格納された情報の豊富さに基づいて前記ドキュメントを前記中央処理装置が順位付けするステップであって、各ドキュメントに対して前記決定された情報の豊富さは、次の式

【数７】

$$InfoRich(d_i) = \sum_{all\, j \neq i} InfoRich(d_j) \cdot \tilde{M}_{ji}$$

として定義され、ただし、ＩｎｆｏＲｉｃｈ（ｄ$_{i}$）はドキュメントｄ$_{i}$の情報の豊富さであり、ドキュメントｄ$_{i}$は前記集まり内のドキュメントを表して、ドキュメントｄ$_{j}$は前記集まり内の別のドキュメントを表して、

【数８】

$$\tilde{M}_{ji}$$

は前記正規化された類似度の行列の要素であり、前記類似度の行列は、前記集まり内のドキュメントの、前記集まり内の別のドキュメントに対する識別された類似度を含み、前記集まり内のドキュメントはウェブページである、ステップと

を備えたことを特徴とする方法。

【請求項５】

ウェブページの集まり内のドキュメントの情報の豊富さを計算するコンピューティングデバイスであって、

コンピュータ実行可能な命令を有するメモリと、

前記メモリに格納されたコンピュータ実行可能な命令を実行する中央処理装置と

を備え、前記コンピュータ実行可能な命令は、

前記集まり内の各ウェブページが有する、前記集まり内の別のウェブページに対する類似度を識別するステップであって、前記類似度は、前記ウェブページの集まり内のウェブページであるウェブページの組のうちの１つのウェブページ、および前記ウェブページの集まり内の別のウェブページである前記組のうちのもう一方のウェブページを含むウェブページの各組に対して識別されて、１つのウェブページの情報が別のウェブページの情報にどの程度含まれているかを示す、ステップと、

前記集まり内の各ウェブページに対して、類似度の行列から得られる正規化された類似度の行列の要素に基づいて前記情報の豊富さを決定するステップと、

前記集まり内のウェブページの前記決定された情報の豊富さを前記メモリに格納するステップと、

前記格納された情報の豊富さに基づいて前記ウェブページを順位付けするステップであって、各ウェブページに対して前記決定された情報の豊富さは、次の式

【数９】

$$InfoRich(d_i) = \sum_{all\, j \neq i} InfoRich(d_j) \cdot \tilde{M}_{ji}$$

として定義され、ただし、ＩｎｆｏＲｉｃｈ（ｄ$_{i}$）はウェブページｄ$_{i}$の情報の豊富さであり、ウェブページｄ$_{i}$は前記集まり内のウェブページを表して、ウェブページｄ$_{j}$は前記集まり内の別のウェブページを表して、

【数１０】

$$\tilde{M}_{ji}$$

は前記正規化された類似度の行列の要素であり、前記類似度の行列は、前記集まり内のウ

ェブページの、前記集まり内の別のウェブページに対する識別された類似度を含む、ステップと
を含むことを特徴とするコンピューティングデバイス。
【請求項６】
前記集まり内の別のウェブページに対する類似度を識別するステップは、類似度の図式を生成するステップであって、前記類似度の図式は、ドキュメントをノードとして表現し、類似度の値をノードの間の有向のエッジの重みとして表現する、ステップを含むことを特徴とする請求項５に記載のコンピューティングデバイス。
【請求項７】
前記ウェブページを順位付けするステップは、
検索エンジンサービスへ問い合わせを送信するステップと、
検索結果としてのウェブページを受信するステップと
を含むことを特徴とする請求項５に記載のコンピューティングデバイス。
【請求項８】
前記ウェブページを順位付けするステップは、前記問い合わせに対する前記検索結果の各ドキュメントの適合度にさらに基づくことを特徴とする請求項７に記載のコンピューティングデバイス。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、一般には、検索エンジンサービスに送信された検索要求により識別される検索結果のドキュメントを順位付ける技術に関する。
【背景技術】
【０００２】
ＧｏｏｇｌｅやＯｖｅｒｔｕｒｅなど多数の検索エンジンサービスにより、インターネットを介してアクセス可能な情報検索が提供されている。これらの検索エンジンは、ユーザに、ユーザが関心を持つであろうウェブページなどの表示ページを検索することを可能にする。ユーザが、検索語を含む検索要求を送信した後、検索エンジンサービスは、これらの検索語に関連する可能性があるウェブページを識別する。関連するウェブページを速やかに識別するために、検索エンジンサービスは、キーワードのウェブページへのマッピングを保持する。このマッピングは、各ウェブページのキーワードを抽出するために、ウェブ（すなわちワールドワイドウェブ）を「クロール（ｃｒａｗｌｉｎｇ）」することによって作成される。ウェブをクロールするために、検索エンジンサービスは、ルートウェブページ（ｒｏｏｔ　ｗｅｂ　ｐａｇｅ）のリストを使用して、ルートウェブページを介してアクセス可能なすべてのウェブページの識別を行うことができる。様々な周知の情報検索技術、例えば、ヘッドラインの単語、ウェブページのメタデータに与えられている単語、およびハイライトされている単語などを識別することなどを用いて、あらゆる特定のウェブページのキーワードを抽出する。検索エンジンサービスは、各ウェブページが検索要求にどの程度関連しているかを示す適合度の得点を、各組合せの緊密性、およびウェブページの人気度（例えば、ＧｏｏｇｌｅのＰａｇｅＲａｎｋ）などに基づいて計算することができる。次いで、検索エンジンサービスは、ユーザに、この適合度に基づいた順序で、各ウェブページへのリンクを表示する。より一般的には、検索エンジンは、あらゆるドキュメントの集まりにおける情報検索を可能にする。例えば、ドキュメントの集まりには、すべての米国特許、すべての連邦の法廷意見、会社のすべての保管文書などが含まれる。
【０００３】
ウェブベースの検索エンジンサービスによって提供される検索結果のうちで、最も高い順位のウェブページを、人気のある同一の話題にすべて方向付けることがある。例えば、ユーザが「スピルバーグ（Ｓｐｉｅｌｂｅｒｇ）」という検索語で検索要求を送信する場合、検索結果のうちで最も高い順位のウェブページは、おそらくスティーブン・スピルバ

ーグ（Ｓｔｅｖｅｎ　Ｓｐｉｅｌｂｅｒｇ）に関係するはずである。しかし、ユーザが、スティーブン・スピルバーグに関心が無く、代わりに、同姓の数学の教授のホームページの所在位置を突き止めることに関心がある場合、ウェブページの順位は、ユーザの役には立たない。教授のホームページは、検索結果に含まれているかも知れないが、ユーザは、教授のホームページへのリンクの所在位置を突き止めるために、検索結果のウェブページへのリンクからなるページを何ページも綿密に見る必要がある場合がある。一般に、所望のドキュメントを検索結果の最初のページ上で識別しない場合、ユーザが、所望のドキュメントの所在位置を突き止めることは困難になることがある。さらに、ユーザは、関心のあるドキュメントを探すために、検索結果を複数のページにわたり見る必要がある場合、挫折することがある。

【０００４】

【特許文献１】米国特許出願第＿＿＿＿号明細書、"Method and System for Calculating Importance of a Block Within a Display Page"、＿＿日出願

【発明の開示】

【発明が解決しようとする課題】

【０００５】

最も高い順位のドキュメントのうちで、より多様な話題を提供する、ドキュメントを順位付けるための技術を有することが望ましく、さらに、このような最も高い順位の各ドキュメントは、その話題に関係する情報内容を非常に豊富に有することが望ましい。

【課題を解決するための手段】

【０００６】

システムは、情報の豊富さおよび話題の多様性に基づいて検索結果のドキュメントを順位付ける。順位付けのシステムにより、検索結果のドキュメントをそれらの関連性に基づいてグループ化し、ドキュメントを同様な話題に関連付けることを意味する。順位付けのシステムにより、最も高い順位のドキュメントには、各トピックに及ぶ少なくとも１つのドキュメントを含み得るように、ドキュメントを順位付ける。次いで、順位付けのシステムにより、最も高い順位のドキュメントのうちの１つとして、グループ内で最も豊富な情報を有するドキュメントを、各グループから選択する。

【発明を実施するための最良の形態】

【０００７】

情報の豊富さおよび話題の多様性に基づいて検索結果のドキュメントを順位付ける方法およびシステムを提供する。一実施形態では、順位付けのシステムは、検索結果に含まれる各ドキュメントの情報の豊富さを決定する。情報の豊富さは、ドキュメントがその話題に関係する情報をどの程度含むのかを示す尺度である。情報の豊富さが高いドキュメント（例えばウェブページ）は、同一の話題に関係し、情報の豊富さがより低いドキュメントの情報を含む情報を含むことがありそうであろう。順位付けのシステムは、検索結果のドキュメントをそれらの関連性に基づいてグループ化し、ドキュメントは同様な話題に関連付けられることを意味する。順位付けのシステムにより、最も高い順位のドキュメントには、各話題に及ぶ少なくとも１つのドキュメントを含むことができ、すなわち、各グループから１つのドキュメントを含むことができ得るように、ドキュメントを順位付ける。順位付けのシステムは、グループ内でドキュメントに属する情報の豊富さが最も高いドキュメントを、各グループから選択する。ドキュメントを順位付けの順序でユーザに提示する場合、ユーザは、おそらく、検索結果の最初のページで、人気のある単一の話題だけよりもむしろ、様々な話題に及ぶドキュメントを見つけるであろう。例えば、検索要求が、検索語の「スピルバーグ」を含む場合、検索結果の最初のページ上のある１つのドキュメントは、スティーブン・スピルバーグに関係しているとともに、最初のページ上の別のドキュメントは、スピルバーグ教授に関係していることが可能である。このようにして、検索結果の最初のページで多様性のある話題に及ぶドキュメントをユーザに提示する可能性は高くなり、また、関心のある話題が検索要求に関係する最も人気のある話題ではない場合でも、ユーザが挫折する可能性は低くなる。さらに、順位付けのシステムは、情報の豊富

さがより低いドキュメントよりも、情報の豊富さがより高いドキュメントを順位付けるので、ユーザは、検索結果の最初のページで提示されるドキュメントのうちから所望の情報を見つける可能性が高くなる。

【０００８】

一実施形態では、順位付けのシステムは、類似度の図式に基づいて、検索結果のドキュメントの情報の豊富さを計算する。類似度は、１つのドキュメントの情報が、別のドキュメントの情報に、どの程度含まれているかを示す尺度である。例えば、スピルバーグの映画のうちの１つを表面的に説明したドキュメントは、スピルバーグの映画のすべてについて詳しく説明したドキュメントに対して、高い類似度を有することがある。反対に、スピルバーグの映画のすべてについて詳しく説明したドキュメントは、スピルバーグの映画のうちの１つを表面的に説明したドキュメントに対して、比較的低い類似度を有する可能性がある。大きく異なる話題に関係するドキュメントは、互いに類似度を有していないはずである。他のすべてのドキュメントに対する各ドキュメントの類似度の集まりにより、類似度の図式を表現する。ドキュメントに対して高い類似度を有する他の多数のドキュメントがあるドキュメントは、高い情報の豊富さを有することがありそうであろう。理由は、そのドキュメントの情報は他の多数のドキュメントの情報を含むからである。さらに、高い類似度を有するこれらの他のドキュメントが、ドキュメント自体に比較的高い情報の豊富さをも有する場合、ドキュメントの情報の豊富さは、さらに高い。

【０００９】

一実施形態では、順位付けのシステムは、類似度の図式をやはり使用して、検索結果で高い順位のドキュメントの多様性を得る助けとなる。順位付けのシステムは、従来技術の順位付けの技術（例えば適合度）、情報の豊富さの技術、または他のいくつかの順位付けの技術に基づいたドキュメントの初期の順位の得点を有することができる。初めに、順位付けのシステムは、最も高い初期の順位の得点を有するドキュメントを、最も高い最終順位の得点を有するドキュメントとして選択する。次いで、順位付けのシステムは、選択したドキュメントに対して高い類似度を有する各ドキュメントの順位の得点を減少させる。それらのドキュメントの内容は、選択したドキュメントにおそらく含まれている冗長な情報であるので、順位付けのシステムは、順位の得点を減少させる。次いで、順位付けのシステムは、次に最も高い順位の得点を有するドキュメントを残りのドキュメントから選択する。順位付けのシステムは、新しく選択したドキュメントに対して高い類似度を有する各ドキュメントの順位の得点を減少させる。順位付けのシステムは、所望の個数のドキュメントが最終順位の得点を得る、すべてのドキュメントが最終順位の得点を得る、または、他の何らかの終了条件が満たされるまで、このプロセスを繰り返す。一実施形態では、多様性とは、ドキュメントの集まりにおける異なる話題の数を表現し、集まりにおけるドキュメントの情報の豊富さは、集まり全体に対するドキュメントの情報を提供する度合いを意味している。

【００１０】

情報の豊富さと多様性との組合せではなくて、情報の豊富さのみまたは多様性のみに基づいて、検索結果のドキュメントを順位付けることができることは当事業者には理解されよう。検索エンジンサービスは、例えば、同様の話題に関係するドキュメントのグループを識別し、そのグループ内の各ドキュメントの情報の豊富さを決定することにより、情報の豊富さのみを用いることがある。次いで、検索エンジンサービスは、決定した情報の豊富さをドキュメントの順位付けの計算に入れ、グループで最も高い情報の豊富さを有するドキュメントを、グループ内の他のドキュメントよりも高く順位付けることがありそうにすることがある。検索エンジンサービスは、例えば、同様の話題に関係するドキュメントのグループを識別し、各グループからの少なくとも１つのドキュメントを、その情報の豊富さに関わらず、検索結果として高く順位付け得ることによって、多様性のみを用いることがある。例えば、検索エンジンサービスは、検索結果の最初のページに、各グループから、グループのうちで最も高い適合度を有するドキュメントを表示するように選択することができる。

【００１１】
類似度の図式では、ドキュメントをノードとして表現し、類似度の値を、ノードの間の有向のエッジの重みとして表現する。順位付けのシステムでは、各ドキュメントを、ドキュメントの集まり内の他のすべてのドキュメントにマッピングする正方行列によって、類似度の図式を表現する。順位付けのシステムは、行列の要素の値を、対応するドキュメントの類似度に設定する。Ｍを行列とする場合、M_{ij}は、ドキュメントｊに対するドキュメントｉの類似度を表現する。順位付けのシステムは、各ドキュメントをベクトルとして表現することにより、ドキュメントの類似度を計算する。ベクトルは、ドキュメントの情報内容を表現する。例えば、各ベクトルは、ドキュメントの最も重要な２５個のキーワードを含むことができる。順位付けのシステムは、次の式に従って、類似度を計算することができる。

【００１２】
【数１】

$$\mathrm{aff}(d_i, d_j) = \frac{\vec{d}_i \cdot \vec{d}_j}{\left\|\vec{d}_i\right\|} \qquad (1)$$

【００１３】
ただし、ａｆｆ（d_i，d_j）は、ドキュメントd_jに対するドキュメントd_iの類似度であり、d_iは、ドキュメントd_iのベクトルを表現し、d_jは、ドキュメントd_jのベクトルを表現し、

【００１４】
【数２】

$$\left\|\vec{d}_i\right\|$$

【００１５】
は、ベクトルd_iの長さを表現する。式１では、類似度を、d_jからd_iへの射影の長さに設定する。類似度は、異なる多くの方法で定義することができることは、当事業者には理解されよう。例えば、別のドキュメントに対するひとつのドキュメントの類似度を、その他のドキュメントのキーワードにおいて、そのひとつのドキュメントのキーワードが占める割合に基づいて定義することができる。他のドキュメントに対するひとつのドキュメントの類似度は、集合論の用語では、２つのドキュメントの積集合のキーワード数を、その他のドキュメント内のキーワード数で割ったものとして表現することができる。行列Ｍの各要素は、類似度の図式において、ひとつのドキュメントのノードからその他のドキュメントのノードへの有向のエッジを表現する。一実施形態では、順位付けのシステムは、類似度の閾値（例えば、２）未満の類似度の値を零に設定する。概念上、類似度が低い場合、類似度の図式において、ひとつのドキュメントのノードからその他のドキュメントのノードへの有向のエッジが存在しないことを意味する。類似度の行列は、次の式で表現する。

【００１６】
【数３】

$$M_{ij} = \begin{cases} aff(d_i, d_j), & add(d_i, d_j) \geq aff_t \quad \text{の場合} \\ 0 \quad , & add(d_i, d_j) < aff_t \quad \text{の場合} \end{cases} \qquad (2)$$

【００１７】
ただし、M_{ij}は、行列の要素であり、aff_tは、類似度の閾値である。ノードの間に多数のエッジを有するノードのグループは、単一の話題を表現する可能性がある。理由は、グループ内のドキュメントの多くが、互いに閾値の類似度よりも大きい類似度を有するからである。対照的に、ノードの間にリンクを有していないノードは、異なる話題に方向付けられたドキュメントを表現する。

【００１８】

順位付けのシステムは、類似度の図式に対して、エッジ（ｅｄｇｅ）解析のアルゴリズムを適用することによって、各ドキュメントの情報の豊富さを計算する。順位付けのシステムは、各行の値の合計が１になるよう類似度の行列を正規化する。正規化された類似度の行列は、次の式で表現する。

【００１９】

【数４】

$$\tilde{M}_{ij} = \begin{cases} M_{i,j} \Big/ \sum_{j=1}^{n} M_{ij}, & \sum_{j=1}^{n} M_{ij} \neq 0 \quad \text{の場合} \\ 0 \qquad\qquad , & \sum_{j=1}^{n} M_{ij} = 0 \quad \text{の場合} \end{cases} \qquad (3)$$

【００２０】

ただし、

【００２１】

【数５】

$\tilde{M}_{ij}$

【００２２】

は、正規化された類似度の行列の要素である。順位付けのシステムは、次の式に従って情報の豊富さを計算する。

【００２３】

【数６】

$$\mathrm{InfoRich}(d_i) = \sum_{\text{all } j \neq i} \mathrm{InfoRich}(d_j) \cdot \tilde{M}_{ji} \qquad (4)$$

【００２４】

ただし、ＩｎｆｏＲｉｃｈ（ｄ$_{i}$）は、ドキュメントｄ$_{i}$の情報の豊富さである。したがって、情報の豊富さは、帰納的に定義される。式４は、次の式により行列の形式で表現することができる。

【００２５】

【数７】

$$\lambda = \tilde{M}^{T}\lambda \qquad (5)$$

【００２６】

ただし、λ＝［ＩｎｆｏＲｉｃｈ（ｄ$_{i}$）］$_{n \times 1}$は、正規化された類似度の行列

【００２７】

【数８】

$\tilde{M}^{T}$

【００２８】

の固有ベクトルである。正規化された類似度の行列

【００２９】

【数９】

$\tilde{M}$

【００３０】

は、通常、疎行列であるので、場合によっては、すべてが零である行が、行列の中に現れることがあり、いくつかのドキュメントは、いくつかのドキュメントに対して重要な類似度を有する他のドキュメントがないことを意味する。順位付けのシステムは、意味のある固有ベクトルを計算するために、ドキュメントの人気度に基づくドキュメントの順位とす

ることができる、ダンピング因子（ｄｕｍｐｉｎｇ　ｆａｃｔｏｒ）（例えば、８５）を使用する。ダンピング因子を用いた情報の豊富さは、次の式で表現する。

【００３１】

【数１０】

$$\mathrm{InfoRich}(d_i) = c \cdot \sum_{\text{all } j \neq i} \mathrm{InfoRich}(d_j) \cdot \tilde{M}_{ji} + \frac{(1-c)}{n} \qquad (6)$$

【００３２】

ただし、ｃは、ダンピング因子であり、ｎは、集まりの中のドキュメントの数である。式６は、次のように行列の形式で表現することができる。

【００３３】

【数１１】

$$\vec{\lambda} = c\tilde{M}^{T}\vec{\lambda} + \frac{(1-c)}{n}\vec{e} \qquad (7)$$

【００３４】

ただし、

【００３５】

【数１２】

$\vec{e}$

【００３６】

は、すべての成分が１に等しい単位ベクトルである。情報の豊富さの計算は、情報のフローおよびシンク（ｓｉｎｋ）モデルから類推することができる。このモデルでは、情報が、各反復でノードの間を流れる。ドキュメントd_iには、次によって表現される類似度を有するドキュメントの集合Ａ（d_i）がある。

【００３７】

Ａ（d_i）＝｛d_j｜∀ｊ≠ｉ，ａｆｆ（d_i，d_j）＞aff_t｝　　（８）

各反復において、情報を、以下の規則のうち１つに従って流すことができる。

１．　確率ｃ（すなわち、ダンピング因子）で、情報は、Ａ（d_i）のうちの１つのドキュメントに流れ込み、ドキュメントd_jに流れ込む確率は、ａｆｆ（d_i，d_j）に比例する。

２．　１－ｃの確率で、情報は、集まり内のあらゆるドキュメントに無作為に流れ込む。

【００３８】

上述のプロセスからマルコフ連鎖を帰納することができ、状態はドキュメントによって与えられ、推移（すなわち、フロー）行列は、

【００３９】

【数１３】

$$c\tilde{M}^{T} + \frac{(1-c)}{n}U \qquad (a)$$

【００４０】

によって与えられる。ただし、

【００４１】

【数１４】

$$U = \left[\frac{1}{n}\right]_{n \times n}$$

【００４２】

である。各状態の定常確率分布は、推移行列の主固有ベクトルによって与えられる。

【００４３】
一実施形態では、順位付けのシステムは、同一の話題に方向付けられる複数のドキュメントが、他の話題に方向付けられるドキュメントを排除して、すべてが高く順位付けられないよう、情報の豊富さと類似性のペナルティを組合せることによって、類似度の順位を計算する。類似性のペナルティの使用により、最も高い順位の複数のドキュメントの間における話題の多様性が増大する。順位付けのシステムは、切望の反復アルゴリズムを使用して、類似性のペナルティを計算することができ、ドキュメントの初期の類似度の順位を、ドキュメントの情報の豊富さに設定する。各反復では、アルゴリズムは、その次に最も高い類似度の順位を有するドキュメントを選択し、同一の話題に方向付けられるドキュメントの類似度の順位を、類似性のペナルティによって減少させる。したがって、ドキュメントを選択した後、同一の話題に方向付けられる他のすべてのドキュメントが減少された類似度の順位を有することにより、最も高い順位のドキュメントが様々な話題を表す機会を向上させるであろう。順位付けのシステムは、次の式に従って、ドキュメントの類似度の順位を減少させることができる。

【００４４】
【数１５】

$$AR_j = AR_j - \tilde{M}_{ji} \cdot InfoRich(d_i) \qquad (10)$$

【００４５】
ただし、ＡＲ$_j$はドキュメントｊの類似度の順位を表現し、ｉは選択されたドキュメントである。類似性のペナルティは、類似度の行列に基づくので、ドキュメントが選択したドキュメントに類似しているほど、ドキュメントの類似性のペナルティは大きくなる。

【００４６】
一実施形態では、順位付けのシステムは、全体の順位を生成するために、類似度の順位と、テキストベースの順位（例えば、従来技術の適合度）とを組合せることができる。順位付けを、得点または順位に基づいて組合せることができる。組合せた得点に関して、ドキュメントの最終得点を表現する全体的な得点を得るために、テキストベースの得点を類似度の順位と組合せる。組合せた得点は、テキストベースの得点と類似度の順位との線形結合に基づくことができる。得点は異なるオーダの大きさを有するので、順位付けのシステムは得点を正規化する。組合せた得点は、次の式で表現することができる。

【００４７】
【数１６】

$$\mathrm{Score}(q, d_i) = \alpha \cdot \frac{\mathrm{Sim}(q, d_i)}{\overline{\mathrm{Sim}}_{\Theta}(q)} + \beta \cdot \frac{\log \overline{\mathrm{AR}}_{\Theta}}{\log \mathrm{AR}_i}, \ \forall d_i \in \Theta \qquad (11)$$

【００４８】
ただし、α＋β＝１であり、Θは検索要求ｑに対する検索結果を表現し、Ｓｉｍ（ｑ，ｄ$_i$）は検索要求ｑに対するドキュメントｄ$_i$の類似性を表現し、

【００４９】
【数１７】

$$\overline{\mathrm{Sim}}_{\Theta}(q) = \mathrm{Max}_{\forall d_i \in \Theta} \mathrm{Sim}(q, d_i) \qquad (12)$$

【００５０】
であり、

【００５１】
【数１８】

$$\overline{\mathrm{AR}}_{\Theta} = \mathrm{Max}_{\forall d_i \in \Theta} \mathrm{AR}_i \qquad (13)$$

【００５２】
である。

【００５３】
組合せた順位に関して、ドキュメントの最終順位を得るために、テキストベースの順位を、類似度の順位と組合せる。組合せた順位は、テキストベースの順位と類似度の順位との線形結合に基づくことができる。組合せた順位は、次の式で表現することができる。
【００５４】
【数１９】

$$\mathrm{Score}(q,d_i)=\alpha\cdot\mathrm{Rank}_{\mathrm{Sim}(q,d_i)}+\beta\cdot\mathrm{Rank}_{AR_i},\ \forall d_i\in\Theta \qquad (14)$$

【００５５】
ただし、Ｓｃｏｒｅは、検索要求ｑに対するドキュメントｄ$_i$の最終順位を表現し、
【００５６】
【数２０】

$$\mathrm{Rank}_{\mathrm{Sim}(q,d_i)}$$

【００５７】
は、テキストベースの順位を表現し、
【００５８】
【数２１】

$$\mathrm{Rank}_{AR_i}$$

【００５９】
は、類似度の順位を表現する。結合のアルゴリズムの中のαおよびβの両方は、調整可能なパラメータである。α＝１かつβ＝０の場合、再順位付けは実行されず、検索結果は、テキストベースの検索に基づいて順位付けされる。β＞αの場合、再順位付けを行う際に、類似度の順位付けに、より重みがかけられる。β＝１かつα＝０の場合、類似度の順位付けだけに基づいて、再順位付けが行われる。
【００６０】
図１は、一実施形態における類似度の図式を例示する図である。類似度の図式１００には、ノード１１１～１１５、ノード１２１～１２４、およびノード１３１を含み、各々はドキュメントを表現する。ノードの間の有向のエッジは、別のノードに対するひとつのノードの類似度を示す。例えば、ノード１１１は、ノード１１５に対する類似度を有するが、ノード１１５は、ノード１１１に対する類似度を有していない（または閾値のレベルを下回る類似度を有する）。この例では、ノードグループ１１０は、同一の話題に方向付けられたノード１１１～１１５を含む。理由は、このノードグループのノードの間には多数のエッジが存在するからである。同様に、ノードグループ１２０は、同一の話題に方向付けられたノード１２１～１２４を含む。ノードグループ１３０には、１つのノードしかないのは、このノードが、他のどのノードに対しても類似度を有しておらず、このノードに対する類似度を有するノードもないからである。ノード１１５は、おそらくノードグループ１１０のすべてのノードのうちで最も高い情報の豊富さを有し、ノード１２４は、おそらくノードグループ１２０のすべてのノードのうちで最も高い情報の豊富さを有する。理由はは、各ノードは、ノードに対する類似度を有するノードの数が最も多いからである。
【００６１】
図２は、一実施形態における順位付けのシステムのコンポーネントを例示するブロック図である。順位付けのシステム２００は、データストア２０１～２０４、およびコンポーネント２１１～２１６を含む。ドキュメントストア２０１は、ドキュメントの集まりを収容するが、このストアは、インターネットを介して利用可能なすべてのウェブページを表現する場合もある。類似度の図式を生成するコンポーネント２１１は、ドキュメントストアのドキュメントに基づいて類似度の図式を生成する。類似度の図式を生成するコンポーネントは、類似度の図式ストア２０２内に類似度を格納する。情報の豊富さを計算するコンポーネント２１２は、類似度の図式ストアからの類似度の図式を入力し、各ドキュメン

トの情報の豊富さの得点を計算する。このコンポーネントは、計算した情報の豊富さの得点を、情報の豊富さストア２０３に格納する。一実施形態では、類似度の図式を生成するコンポーネント、および情報の豊富さを計算するコンポーネントは、検索を実施するのに先立って、類似度の図式および情報の豊富さの得点を生成するためにオフラインで実行することができる。検索を実施するコンポーネント２１３は、ユーザからの検索要求を受信し、ドキュメントストアのドキュメントから検索結果を識別する。検索を実施するコンポーネントは、検索結果を、検索要求に対する検索結果の各ドキュメントの適合度の表示と共に検索結果ストア２０４に格納する。類似性のペナルティを計算するコンポーネント２１４は、類似性ペナルティを計算して、検索結果ストア、類似度の図式ストア、および情報の豊富さストアの情報に基づく類似度の順位に適用する。類似度の順位を計算するコンポーネント２１５は、検索結果の中にある各ドキュメントの類似度の順位を生成する。類似度の順位を計算するコンポーネントは、ドキュメントの情報の豊富さ、類似度の図式ストア、および検索結果を計算に入れる。最終得点を計算するコンポーネント２１６は、類似度の順位と適合度の得点とを組合せて、最終得点を計算する。

【００６２】

順位付けのシステムが実装されるコンピューティングデバイスには、中央処理装置、メモリ、入力装置（例えば、キーボード、ポインティングデバイス）、出力装置（例えば、表示装置）、および記憶装置（例えば、ディスク装置）が含まれる。メモリおよび記憶装置は、順位付けのシステムを実装する命令を含むことができるコンピュータ可読媒体である。さらに、データ構造およびメッセージ構造は、通信リンク上の信号などのデータ伝送媒体を介して格納または伝送することができる。様々な通信リンク、例えば、インターネット、ＬＡＮ、ＷＡＮ、または、ポイントツーポイントのダイヤルアップ接続などを使用することができる。

【００６３】

順位付けのシステムは、様々な動作環境で実装することができる。使用するのに適切となり得る、周知の様々なコンピューティングシステム、環境、および構成には、パーソナルコンピュータ、サーバコンピュータ、ハンドヘルドまたはラップトップデバイス、マルチプロセッサシステム、マイクロプロセッサベースのシステム、プログラム可能家庭用電化製品、ネットワークＰＣ、ミニコンピュータ、メインフレームコンピュータ、および、あらゆる上述のシステムまたは装置を含む分散コンピューティング環境などが含まれる。

【００６４】

１つまたは複数のコンピュータまたは他の装置によって実行される、プログラムモジュールなどコンピュータ実行可能命令の一般的コンテキストにおいて、順位付けのシステムを説明することができる。一般に、プログラムモジュールには、特定のタスクを実行または特定の抽象データ型などを実装する、ルーチン、プログラム、オブジェクト、コンポーネント、およびデータ構造などが含まれる。一般に、様々な実施形態において、プログラムモジュールの諸機能を、所望に応じて、結合または分散することができる。

【００６５】

図３は、一実施形態における順位付けのシステムの全体的な処理を例示する流れ図である。順位付けのシステムには、検索結果を表現することができるドキュメントの集まりが提供される。ブロック３０１では、コンポーネントが、ドキュメントの集まりについての類似度の図式を構成する。コンポーネントは、オフラインでドキュメントの集成の中にあるすべてのドキュメント（例えば、すべてのウェブページ）に及ぶ、またはリアルタイムで集まりのドキュメントのみに及ぶ類似度の図式を構成することができる。ブロック３０２では、コンポーネントは、集まりの各ドキュメントの情報の豊富さを計算する。ブロック３０３では、コンポーネントは、集まりのドキュメントを順位付けし、次いで完了となる。

【００６６】

図４は、一実施形態における類似度の図式を構成するコンポーネントの処理を例示する流れ図である。コンポーネントは、ドキュメントの集まりが伝わり、これらのドキュメン

トについての類似度の図式を構成する。ブロック４０１～４０３では、コンポーネントは、ドキュメント集まりの中にある各ドキュメントについてのドキュメントベクトルを生成するループを実行する。ブロック４０１では、コンポーネントは、集まり内の次のドキュメントを選択する。決定ブロック４０２では、集まり内のすべてのドキュメントを既に選択した場合、ブロック４０４に進み、そうでなければ、ブロック４０３に進む。ブロック４０３では、コンポーネントは、選択されたドキュメントに対するドキュメントベクトルを生成し、次いで、集まり内の次のドキュメントを選択するためにブロック４０１へのループを実行する。ブロック４０４～４０８では、コンポーネントは、集まり内のドキュメントの各組についての類似度を計算する。ブロック４０４では、コンポーネントは、第１のドキュメントから開始し、集まりの中にある次のドキュメントを選択する。決定ブロック４０５では、すべてのドキュメントが既に選択された場合、コンポーネントは、類似度の図式を返し、そうでない場合、ブロック４０６に進む。ブロック４０６～４０８では、コンポーネントは、集まりの各ドキュメントを選ぶループを実行する。ブロック４０６では、コンポーネントは、第１のドキュメントから開始し、集まりの中にある次のドキュメントを選択する。決定ブロック４０７では、集まり内のすべてのドキュメントが既に選ばれた場合、コンポーネントは、集まり内の次のドキュメントを選択するためにブロック４０４へのループを実行し、そうでない場合、ブロック４０８に進む。ブロック４０８では、コンポーネントは、式１に従って（ブロック４０６で）選んだドキュメントに対する（ブロック４０４で）選択されたドキュメントの類似度を計算し、次いで、集まり内の次のドキュメントを選ぶブロック４０６へのループを実行する。

【００６７】

図５は、一実施形態におけるドキュメントを順位付けるコンポーネントの処理を例示する流れ図である。コンポーネントには、類似度の図式が生成され、各ドキュメントの情報の豊富さが計算されたドキュメントの集まりを伝える。ブロック５０１～５０３では、コンポーネントは、集まり内の各ドキュメントの類似度の順位を、その情報の豊富さに初期設定するループを実行する。ブロック５０１では、コンポーネントは、集まり内の次のドキュメントを選択する。決定ブロック５０２では、すべてのドキュメントを既に選択した場合、ブロック５０４に進み、そうでない場合、ブロック５０３に進む。ブロック５０３では、コンポーネントは、選択したドキュメントの類似度の順位を、選択したドキュメントの情報の豊富さに設定し、集まり内の次のドキュメントを選択するブロック５０１へのループを実行する。ブロック５０４～５０８では、コンポーネントは、ドキュメントの組を識別し、類似性のペナルティによって類似度の順位を調整するループを実行する。ブロック５０４では、コンポーネントは、次に最も高い類似度の順位を有するドキュメントを選択する。決定ブロック５０５では、終了条件に到達した場合、コンポーネントは、順位付けしたドキュメントを返し、そうでない場合、ブロック５０６に進む。ブロック５０６～５０８では、コンポーネントは、ドキュメントを選び、類似性のペナルティにより類似度を調整するループを実行する。ブロック５０６では、コンポーネントは、選ばれたドキュメントから選択されたドキュメントへの類似度についての類似度の図式において、零以外の値で指示されるように、選択されたドキュメントに対する類似度を有する、次のドキュメントを選ぶ。決定ブロック５０７では、このようなドキュメントがすべて既に選ばれた場合、コンポーネントは、次に最も高い類似度の順位を有するドキュメントを選択するブロック５０４へのループを実行する。ブロック５０８では、コンポーネントは、式１０に従って、類似性のペナルティによって選ばれたドキュメントに対する類似度の順位を調整する。次いで、コンポーネントは、選択されたドキュメントに対する類似度を有する次のドキュメントを選ぶブロック５０６へのループを実行する。

【００６８】

本明細書では、例示の目的のために、順位付けのシステムの特定の実施形態について説明したが、本発明の精神および範囲から逸脱することなく様々な変形形態を構成することができることは当事業者には理解されよう。一実施形態では、順位付けのシステムは、ドキュメントごとではなくブロックごとを基礎にして類似度および情報の豊富さを計算する

ことができる。ブロックは、単一の話題に一般的に関係するウェブページの情報を表現する。ウェブページの順位付けは、そのウェブページに対するブロックの重要度に部分的に基づくことができる。ブロックの重要度に関しては文献に記載されている（例えば、特許文献１参照。この文献を参照により本明細書に援用する。）。したがって、本発明は添付の特許請求の範囲を除いて限定されるものではない。

【図面の簡単な説明】

【００６９】

【図１】一実施形態における類似度の図式を例示する図である。

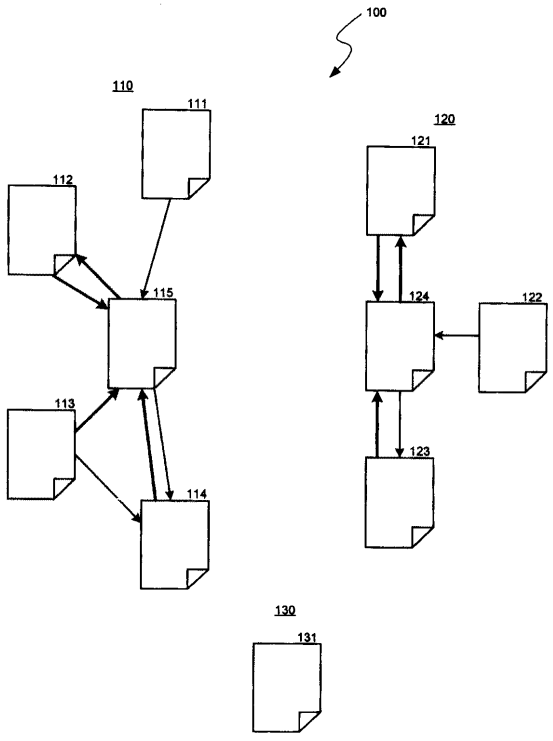
【図２】一実施形態における順位付けのシステムのコンポーネントを例示するブロック図である。

【図３】一実施形態における順位付けのシステムの全体的な処理を例示する流れ図である。

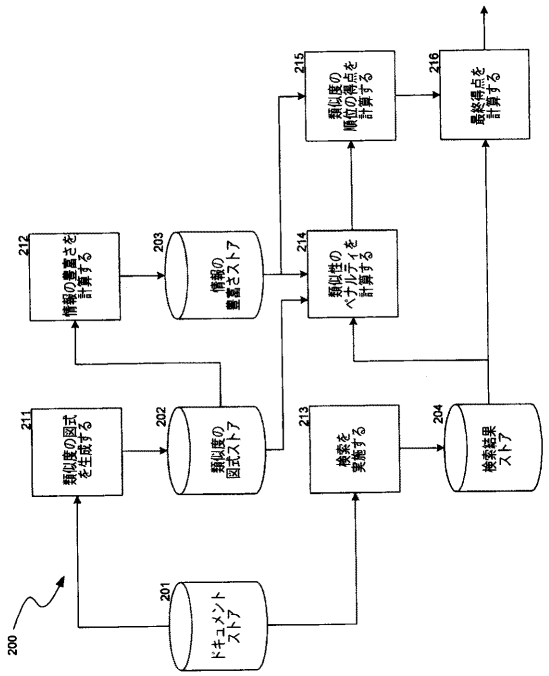
【図４】一実施形態における類似度の図式を構成するコンポーネントの処理を例示する流れ図である。

【図５】一実施形態におけるドキュメントを順位付けるコンポーネントの処理を例示する流れ図である。

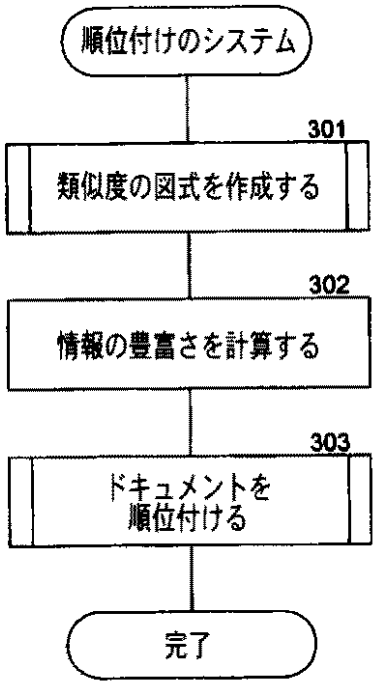
【図１】



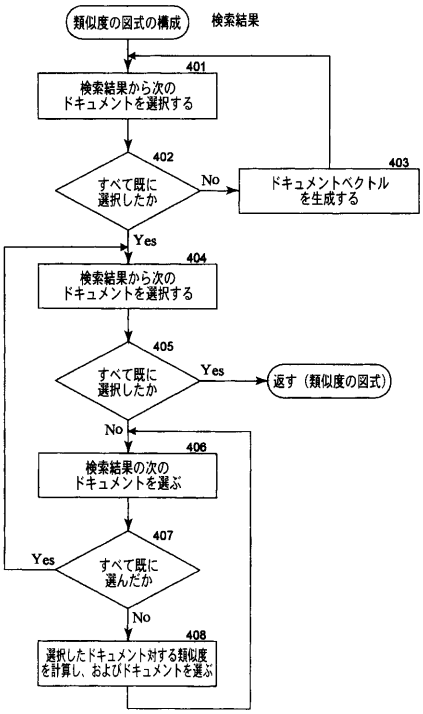
【図２】



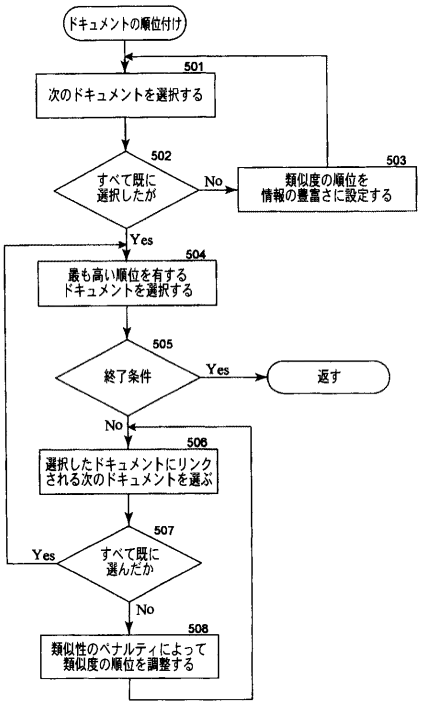
【図３】


順位付けのシステム
301
類似度の図式を作成する
302
情報の豊富さを計算する
303
ドキュメントを
順位付ける
完了

【図４】


類似度の図式の構成
検索結果
401
検索結果から次の
ドキュメントを選択する
402
すべて既に
選択したか
No
403
ドキュメントベクトル
を生成する
Yes
404
検索結果から次の
ドキュメントを選択する
405
すべて既に
選択したか
Yes
返す（類似度の図式）
No
406
検索結果の次の
ドキュメントを選ぶ
407
Yes
すべて既に
選んだか
No
408
選択したドキュメント対する類似度
を計算し、およびドキュメントを選ぶ

【図５】


ドキュメントの順位付け
501
次のドキュメントを選択する
502
すべて既に
選択したが
No
503
類似度の順位を
情報の豊富さに設定する
Yes
504
最も高い順位を有する
ドキュメントを選択する
505
終了条件
Yes
返す
No
506
選択したドキュメントにリンク
される次のドキュメントを選ぶ
507
Yes
すべて既に
選んだか
No
508
類似性のペナルティによって
類似度の順位を調整する

フロントページの続き

(72)発明者 ホアージュン チュン
アメリカ合衆国 ９８０５２ ワシントン州 レッドモンド ワン マイクロソフト ウェイ マイクロソフト コーポレーション内
(72)発明者 ウェイーイン マ
アメリカ合衆国 ９８０５２ ワシントン州 レッドモンド ワン マイクロソフト ウェイ マイクロソフト コーポレーション内
(72)発明者 チョン チェン
アメリカ合衆国 ９８０５２ ワシントン州 レッドモンド ワン マイクロソフト ウェイ マイクロソフト コーポレーション内

審査官 打出 義尚

(56)参考文献 特開２００４－３１８５２８（ＪＰ，Ａ）
Carbonell J, The use of MMR, diversity-based reranking for reordering documents and producing summaries, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 米国, ＡＣＭ, １９９８年, ｐｐ. ３３５－３３６, ＵＲＬ, http://doi.acm.org/10.1145/290941.291025
Goldstein J, Multi-document summarization by sentence extraction, NAACL-ANLP 2000 Workshop on Automatic summarization - Volume 4, 米国, Association for Computational Linguistics, ２０００年, ｐｐ. ４０－４８, ＵＲＬ, http://dx.doi.org/10.3115/1117575.1117580
Inderjeet Mani, 自動要約, 日本, 共立出版株式会社, ２００３年 ６月２０日, 初版, ｐｐ. １８１－１８３

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｆ １７／３０