4ZE-07　WWWハイパーリンク構造を利用した文書検索方式

鈴木 賢太郎†，野口 進祐††，木下 哲男††，白鳥 則郎††
†東北大学工学部
††東北大学電気通信研究所/情報科学研究科

1. はじめに

本稿では，WWW 文書検索における検索精度を改善するためにハイパーリンク構造の情報を利用して文書ベクトルを拡張する方式を提案する．

2. WWWハイパーリンク構造を用いた文書検索

WWW 上の文書の代表的な検索手法としては，tf･idf ベクトルを用いる方法がある．tf･idf ベクトルは，文書中の単語の出現頻度及び各単語のもつ情報量に基づき文書をベクトル表現する．この方法では，文書中の単語の意味は考慮されないので，文書中にない単語を検索要求として入力した際には無効である．この欠点を補う方法として，検索要求語から，類義語，共起語の情報を辞書から得て拡張する手法もあるが，もともと情報量の少ない検索要求語を拡張するのには限界がある．これに対して，検索要求語ではなく文書側の情報により文書ベクトルを拡張する方法が提案されている．研究[1]では学術論文に付与されたキーワードを，研究[2]では学術論文の引用関係をそれぞれ利用して，関連する文書の集合を形成し，その解析から文書ベクトルを拡張し文書検索の精度改善に効果があることが報告されている．本稿では，ハイパーリンク構造から得られる情報を利用することにより文書部分集合を形成し，その解析により tf･idf ベクトルを拡張する手法を提案する．

3. ハイパーリンクを利用した文書ベクトルの拡張

3.1. 文書固有の特徴量抽出

最初に，tf･idf 法により文書固有の特徴量である文書ベクトルを計算する．文書 D_i を次のベクトルで表現する．

$$d_i = (d_{i1}, d_{i2}, ..., d_{in})$$

ここで，d_{ij} は文書 D_i に対する単語 W_j の重みで，次の式で計算される．

$$d_{ij} = tf_{ij} \cdot \log(N/df_j)$$

ここで，tf_{ij} は文書 D_i 中の単語 W_j の出現頻度，N は全文書数，df_j は単語 W_j の出現文書数である．

3.2. 文書部分集合の形成

ハイパーリンクの構造情報をもちいて文書部分集合を形成する手法を述べる．ある文書 D_i に着目したとき，D_i 自身と，D_i が直接リンクする文書群で文書部分集合を形成する．一般に，WWW のハイパーリンクは多様な意味で使われているため，すべてのハイパーリンクを用いたのでは意味のある文書集合にはなりにくい．そこでハイパーリンクを選別する必要がある．ここではハイパーリンクの分類[3]を利用して，情報を持たないページへのリンク，ページ内での移動などに用いられる便宜的なリンクを除外して文書集合を形成する．

3.3. 文書部分集合を利用した文書ベクトルの拡張

形成された文書部分集合を利用して，各文書ベクトルを拡張する．具体的には下記の式に基づき，文書 D_i における文書部分集合の文書ベクトルの平均を計算し，d_i と各要素について比較し，その値が最大となる要素を d_i の要素として選択する．

$$d'_{ij} = \max\left(d_{ij}, \frac{1}{|N_{D_i}|} \sum_{x \in N_{D_i}} d_{xj} \right)$$

ここで N_{D_i} は文書 D_i のリンクによって形成される文書部分集合である．

このようにして，文書ベクトルを拡張することにより，文書側において関連する語群を強調する効果が得られる．また，WWW 情報空間上の文書は同じテーマに関する文書でも文書表現が曖昧であることが多い．文書を個別に扱う場合にはこの文書表現の曖昧さによる影響が検索精度の悪化につながるが，関連文書から単語を補うことによって，検索精度が検索語に依存しにくくなる．これにより，検索者の曖昧な検索に対応でき，平均的な検索精度が向上する．

4. おわりに

本稿では WWW ハイパーリンク構造を用いて文書ベクトルを拡張する方法を提案した．現在，試作システムを用いて，提案手法と既存手法の比較実験を行っている．

参考文献

[1] 金沢 輝一，高須 淳宏，安達 淳：文書関連性を考慮した検索方式，情報処理学会 DBS 研究会，DB ワークショップ'98，(48)

[2] 野口 進祐，木下 哲男，白鳥 則郎：学術論文の引用関係に基づく特徴量の抽出手法，情処研報 DPS，94-13，1999

[3] 小野田 浩平，土肥 浩，石塚 満：WWW ハイパーリンクの意味による分類とノードリンクの提示，情処全大(56)，Vol.3，pp.155-156，1998

A Retrieval Method Based on WWW Hyperlink Structure
Kentaro Suzuki†, Shinsuke Noguchi††, Tetsuo Kinoshita†
† and Norio Shiratori††
†School of Engineering, Tohoku University
††Research Institute of Electrical Communication / Graduate School of Information Science, Tohoku University