自動作成された単語間リンクによる検索質問作成支援

4E－6

津田宏治　仙田修司　美濃導彦　池田克夫
京都大学工学部

1　はじめに

文書データベース（文書ＤＢ）から文書を検索するとき，ユーザは検索質問（キーワードの組）を作成しなくてはならない．しかし，ユーザが適切な単語を探し，検索質問を作るのは容易な作業ではない[1]．我々が提案する文書検索システムは，キーワードの候補を提示することによって，ユーザの検索質問作成を支援する．ユーザが具体的な単語を思いつかないときも，候補中から単語を選ぶことによって検索質問が作成できる．また，キーワードの入力作業を必要としないので，ユーザにかける負担が少ない．

2　検索質問作成の手順

ユーザに文書ＤＢの概要を知らせるために，単語の組を文書ＤＢに蓄積された文書から抽出して図1のように提示する．この単語の組のことを「話題」と呼ぶ．

ユーザが検索質問を作成する過程を図2に示す．はじめに，ユーザはシステムが提示した話題の中から一つを選択し，その話題中から単語を選択する．次に，システムはユーザの選んだ単語に関係の深い単語を10個提示し，ユーザはそこから単語を選ぶ．この作業を何度か繰り返した後，ユーザの選んだ単語の集合が検索質問となる．

（関数 論理 回路 生成 故障 二分決定図 表現 論理回路 ）
（スレッド マルチ レイテンシ シングルプロセッサ ）

図 1: 話題の例

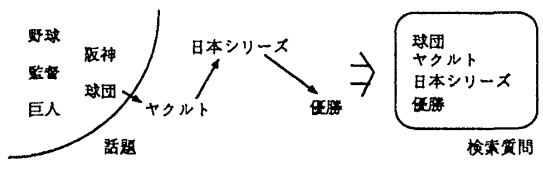

図 2: 検索質問作成の手順

Query formulation support by automatically generated term-to-term links
Koji TSUDA, Shuji SENDA, Michihiko MINOH, Katsuo IKEDA
Faculty of Engineering, Kyoto University

3　文書ＤＢからの話題の抽出法

文書ＤＢに蓄積されたすべての文書から抽出された単語をノードとし，その間に共起頻度（単語が同じ文書に現れた回数）を重みとするリンクを張る．こうして構成されるグラフを**単語グラフ**と呼ぶ．

話題は，単語グラフの中から互いに重みの大きいリンクで連結されている単語を抽出することによって得ることができる．ここでは，数量化IV類[2]を応用したクラスタリング手法を用いて話題を抽出している．本手法は，小規模（大半は10単語以下）で，ほとんどすべて（85%以上）の単語間にリンクが存在する，非排他的な単語集合を生成する．

1 単語 i,j の共起頻度を $(i,j),(j,i)$ 要素とする対称行列（共起行列）を作る．
2 共起行列の固有ベクトルを固有値の大きい順に30個求める．
3 固有ベクトルの要素の内，絶対値がある閾値（0.1程度）以上のものに対応する単語を正負別に集め単語集合を作る．

単語 $i,j(1 \leq i,j \leq n)$ の共起頻度を c_{ij} とすると，単語 i の共起ベクトル $\mathbf{c}_i$ は $\mathbf{c}_i = \{c_{ij}\}(1 \leq j \leq n)$ のように定義される．生成された単語集合中の任意の2単語について，それぞれの共起ベクトルの成す角のコサインを計算し，その平均をとる（単語集合の確信度）．そして確信度がある閾値(0.4)以下の単語集合は複数の話題を示す単語が混ざっている可能性が高いため排除する．その結果，話題数は，文書ＤＢによっても異るが10～20個程度になる．

4　ユーザの興味に合わせた提示単語候補の選択

システムは，ユーザが選んだ単語と単語グラフ上で連結している単語の中から，提示する候補単語を選ぶ．ユーザに提示できる候補の数には限りがあるので，ユーザの興味に関係のない単語の候補を提示するのは極力避ける必要がある．ここでは，ユーザの興味を察知し，それに合わせて提示単語候補を変える手法について述べる．

図3(a)は，ユーザが複数の意味を持つ単語（多義語）を選んだときの候補を示している．候補には複数のカテゴリの単語が混ざり合い，検索質問作成が行き詰まる原因となっている．ＤＢから抽出された話題 $T_k(1 \leq k \leq m)$ すべてについて，各単語の帰属度（単語がその

話題の文書に使われている確率）を計算すれば，各単語は，話題を座標軸とする m 次元空間（話題空間）上の一点として表される．ユーザが既ににたどった単語と，提示単語候補の話題空間上の位置を比較することにより，ユーザの興味に合わない候補を削除することができる（図3(b)）．単語 i がどの程度ユーザの興味にあっているかを示す評価値を興味適合度（$conformity_i$）と呼ぶ．

話題 $T_k(1 \leq k \leq m)$ に含まれる単語の共起ベクトルの総和を $\mathbf{c}_{T_k}$ とするとき，単語 i の話題 T_k への帰属度 $bel(i,T_k)$ を次のように定義する．

$$bel(i,T_k) = \frac{\mathbf{c}_i \cdot \mathbf{c}_{T_k}}{|\mathbf{c}_i||\mathbf{c}_{T_k}|}$$

単語 i の話題空間上の位置ベクトル $\mathbf{w}_i$ は，単語の各話題への帰属度を要素とするベクトルで表される．

$$\mathbf{w}_i = (bel(i,T_1), bel(i,T_2), \cdots, bel(i,T_m))$$

ユーザの興味の中心を示すベクトル $\mathbf{ctr}$ は，単語グラフで既にたどった単語の位置ベクトルの総和とする．単語 i の興味適合度を，$\mathbf{w}_i$ と $\mathbf{ctr}$ との成す角のコサインで定義する．

$$conformity_i = \frac{\mathbf{w}_i \cdot \mathbf{ctr}}{|\mathbf{w}_i||\mathbf{ctr}|}$$

ここでは，ユーザが選んだ単語と単語グラフ上で連結している単語の内，リンクの重みの大きい順に20個選択し，その内で興味適合度が大きいもの10個を候補として提示している．

/高木:
1:選手 2:監督 3:パス 4:ゴール 5:投手 6:ボール
7:チーム 8:攻撃 9:ラン 10:ラモス
(a) 興味適合度を考慮に入れない場合の候補
(野球とサッカーが混在)

/ベイスターズ/高木:
1:投手 2:ボール 3:ラン 4:実力 5:ドラゴンズ 6:応援
7:解雇 8:先発 9:中日 10:投票

(b) 興味適合度を考慮に入れた場合の候補
(野球に統一)

図 3: 興味適合度の効果

5 実験と結果

本検索システムの評価基準には，次の三つを用いる．

A 候補を選ぶだけで検索可能な文書の割合（検索可能率）
B 文書を検索するのに必要な平均の単語数（平均単語選択回数）
C 提示候補数に対する有効な候補の割合（有効候補提示率）

A,B の評価には，すべての文書の検索を試みる必要があるため，計算機による検索シミュレーションを行った．計算機は文書を与えられると，文書との類似度が最も高い話題を選び，単語を選んで質問を作成する．(文書と話題との類似度には，各々が含む単語の共起ベクトルの総和の成す角のコサインを用いた．文書と単語との類似度も同様)．質問作成中，一つ単語を選ぶ度に検索を行い，与えられた文書が10位以内に検索されていればその文書は検索可能とした．一方，9単語選んだ時点で検索されなかった場合は検索不可能と判断した．ネットニュースを題材に A,B を測定した結果を図4に示す．文書数が比較的少ない場合には，数回の選択で大半の文書が検索可能であることが分かる．

C の評価のため，実際に50回分，500個の候補単語を次の三つのカテゴリに人手で分類した．

1 興味のある事柄に関係あり，検索質問に含まれる可能性のある単語
2 意味が抽象的過ぎて質問になり得ない単語（動詞，一般名詞等）
3 興味のある事柄に全く関係ない単語

結果は，1 が 61%,2 が 32%,3 が 7%であった．単語候補の選択は統計的なデータのみによるにもかかわらず，3 の興味に関係ない単語の提示は低く押えられている．しかし，2 の抽象的な語が高い率で存在することが，質問作成の障害となっている．

6 結論

候補から単語を選択して検索質問を作成する方法は非常に簡便である．文書数が少ない場合には十分検索の役割を果たしているので，個人的な文書データベースでの利用が期待できる．また，大規模なデータベースにおいても，ある程度文書数を絞った後の詳細な検索に利用できると考える．

文書数	100	300	500	1000
A 検索可能率 (%)	96.0	83.8	67.6	40.2
B 平均選択回数	1.5	2.1	2.6	3.2

図 4: 検索可能率と平均選択回数

参考文献

[1] THOMPSON, R. H. and CROFT, W. B. Support for browsing in an intelligent text retrieval system, *Int.J.Man-Machine Studies*, 30 (1989), 639-668.

[2] 田中豊, 垂水共之, 脇本和昌 パソコン統計解析ハンドブック, 共立出版 (1984).