SASS: Web文書の発想検索支援システム

4L－8

織田 充　南 俊朗　有馬 淳
（株）富士通研究所 ネットメディア研究センター
{oda,minami,arima}@flab.fujitsu.co.jp

1 はじめに－発想的キーワード探索

膨大な量のWeb文書の中から有益なものを探しだすためのキーワード検索エンジンが多数提供されている．そこでの問題の1つは，適切なキーワードを思いつくことの困難さである．発想支援システム[2]における発想過程と同様，入力キーワードの決定過程はどのような文書を参照するか検索意図を定める発散過程と，実際に参照する文書を絞り込む収束過程の2つからなる．適切なキーワードを思いつくことの困難さは，これらの困難さに起因する．自ら思いつくことが困難であっても，示された候補の中から自分の要求に最も近いと思われるキーワードを選択することは容易である．このことを利用し，キーワード推薦システムが開発されている[1, 3]．しかしそれらの多くは，文書中に出現する単語を推薦キーワードとして用いる．これは提供者側からのキーワード推薦方式である．これに対して検索に用いられたキーワードは，利用者の興味や意図を何らかの意味で反映している．本稿ではこの点に注目し，利用者の検索ログを基に利用者のキーワード発想過程を支援する，発想的キーワード推薦システムSASS (Searching Assistant with Social Selection)を提案し，その基本的な考えおよびシステムの概略を述べる．

2 利用者検索ログに基づく関連性

SASSシステムは，検索者が入力したキーワード列に関連したキーワードを推薦する．検索者の興味に基づいて変化するであろう「関連性」として何が適切かを一般的に定めるのは困難である．SASSでは複数の「関連性」選択肢を検索者に提供し，選択された関連性を元に検索ログからキーワードを推薦している．本稿ではその「関連性」の一つである「検索関連性」に焦点を当て紹介する．

絞り込み検索においては，キーワードを補足や，修正が行なわれるが，その結果出された最終的な検索式を以下では単に「検索」と呼ぼう．

SASS: An Idea-Creating Searching Assistant System for Web Documents

Mitsuru Oda, Toshiro Minami and Jun Arima
Netmedia Laboratory, Fujitsu Laboratories Ltd.
2-2-1 Momochihama, Sawara, Fukuoka 814-8588 Japan

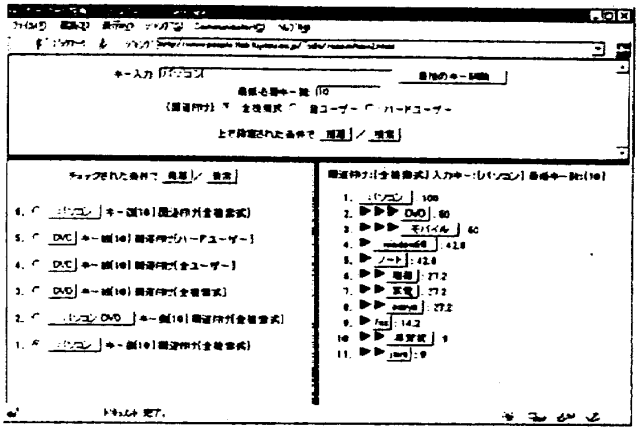

図1: SASSシステムの画面

検索は検索者の興味のコンパクトな表現と見なすことができる．従って，同じキーワードを含む二つの検索があれば，それが表す二つの興味も類似していると推定される．そこで，

仮定 検索者が入力したキーワードkを含む他の検索rがある時，そのrに含まれる他のキーワードk'を推薦することは妥当である．

を前提とすることができる．しかし，ログ解析の結果，同じキーワードを含む検索はそれほど多くなく，興味が多様なそれぞれの検索者に対して推薦するに充分な量のキーワードを見つけられないことが分かった．この問題を解決するために，仮説を推移的に拡張し，それに基づいて下記「検索関連性」を採用している．

推移関連性: キーワードkに対し，i) kを含む検索r上の任意のキーワードk'はkに関連する．また，ii) kとk'が関連している時，k'を含む検索r'上の任意のキーワードはkに関連する．

3 SASSシステム

以下，SASSでのキーワード推薦法を示す．

定義：検索 最終的な検索式に含まれるキーワード集合を「検索」と呼ぶ．ただし時点，検索者による検索の違いが区別できるように，各集合に識別子が付加されているものとする．

定義：関連キーワード $o(r,a,b)$は「検索r中にキーワードa,bが出現する」関係を表す．また，$o'(a,b) = \exists x.o(x,a,b)$とし，$o'$の推移的閉包を$o'^*$

とする．$o'^*(a,k)$ を満たす k を「キーワード a に対する関連キーワード」と呼ぶ．

定義：関連キーワード列

(i) $a_i = a_j \supset i = j$
(ii) $\forall i.\exists x.o(x, a_i, a_{i+1})$ ただし $1 \leq i < n$.
(iii) $\forall i, j, x.o(x, a_i, a_{i+1}) \wedge o(x, a_j, a_{j+1}) \supset i = j$

を満足するキーワード列 $[a_1, a_2, \ldots, a_n]$ をキーワード a_1 から a_n へ至る「関連キーワード列」と呼ぶ．

入力されたキーワード k に対する関連キーワードの集合 S および関連キーワード列の重複を許す集合 P を次の手続きにより求める．

1) 全ての検索を含む集合を R，また関連キーワード列の重複を許す集合を P，関連キーワードの集合を S とする．ただし P, S は共に空とする．
2) 各検索 $r \in R$ について関係 $o(r, k, k')$ が成立するならば，k' を S に，また $[k, k']$ を P に加え，r を R から除く．このような検索 r が存在しなければ終了する．
3) 各検索 $r \in R$ について関係 $o(r, k', k'')$ $(k' \in S, k'' \notin S)$ が成立するならば，k'' を S に，また $[\ldots, k'] \in P$ となる全ての関連キーワード列をキーワード k'' で延長した列 $[\ldots, k', k'']$ を P に加え，r を R から除く．このような検索 r が存在しなければ終了する．
4) ステップ 3 へ戻る．

検索ログの解析から得られた知見からキーワード間の関連性を推移的に拡張したが，逆に複数の検索を介して初めて関連性を持つキーワードと，仮定にあるような直接関連性を持つキーワードとの区別が付かない．そこで関連性の程度を表す評価値として，キーワード間の関連度を導入する．SASS システムは，入力されたキーワードに対する関連キーワード群をその入力キーワードに対する関連度でソートし，その結果を利用者に対して提示する．

定義：キーワード関連度　関連キーワード列集合 P における関連キーワード列 $p(= [a_1, \ldots, a_n])$ の分岐数とは，P に含まれる p を 1 キーワード分延長した関連キーワード列 $[a_1, \ldots, a_n, a_{n+1}]$ の個数である．いまキーワード k から k' へ至る関連キーワード列を $[k, \ldots, k_i'', k'](1 \leq i \leq n)$，またそれぞれの関連キーワード列の分岐数を $N_i(1 \leq i \leq n)$ とするとき，キーワード k に対する関連キーワード k' の関連度 [4] $v_k(k')$ を以下の式で与える．

$$\begin{cases} v_k(k') = \sum_{i=1}^{n} \frac{v_k(k_i'')}{N_i} & if\ k \neq k' \\ v_k(k) = 1 \end{cases}$$

例えば，検索 r_1, r_2, r_3, r_4 において使用されたキーワード集合をそれぞれ $\{A, B\}, \{A, C\}, \{B, C, D\}, \{C, E\}$ とするとき，キーワード A からの D へ至る関連キーワード列は $[A, B, D]$ および $[A, C, D]$ で表現されるものがそれぞれ 1 個ある．またキーワード A に対するキーワード B, C, D, E の関連度はそれぞれ 1/2, 1/2, 3/4, 1/4 となる．

定義：キーワード列に対するキーワード関連度　入力キーワードの列 $k_1, \ldots, k_n$ に対するキーワード k の関連度を $\Sigma_i v_{ki}(k)$ と定める．

SASS システムは，上記認識に基づきユーザの発想的キーワード探索を支援する．図 1 に SASS の画面例を示す．ユーザは，検索エンジンの場合と同様に，まず，自分の目的とする Web 文書を特徴づけるキーワードを指定する (画面上部)．システムは候補毎に，入力されたキーワード列に対する関連度を求め，関連度の高い順に利用者に提示する．(画面右下部分)．過去の履歴も表示され，いつでも元の状態に戻ることができる (画面左下部分)．

4　まとめ

SASS システムは，文書に出現する単語から推薦キーワードを選択する提供者側の意図に基づくキーワード推薦ではなく，利用者の検索情報を基にし，利用者の興味を反映した発想的キーワード推薦を行うことが大きな特徴である．

この方向での文書選択の支援を更に進めるために次の課題の探求が重要である．(1) 検索履歴データより利用者にとり有効なキーワード群を抽出する技術が，本方式の有効性のキーとなる．効果の高い関連度抽出技術を更に追求する必要がある．(2) よりきめ細かな推薦機能をユーザに提示するためには，検索エンジンとキーワード推薦機能の効果的な融合法の研究が必要である．

参考文献

[1] 河野 浩之：問答：検索支援システム構築技術としてのデータマイニング，第 9 回メディア統合技術研究会，画像電子学会，1997.

[2] 國藤 進：発想支援システムの研究開発動向とその課題，人工知能学会誌 Vol.8 No.5, pp.552-559, 1993.

[3] NTT: TITAN, URL: http://sting.navi.ntt.co.jp/titan/titan-clx.html

[4] 南 俊朗，織田 充：関連度を用いた Web 文書のナビゲーション，マルチメディア通信と分散処理研究会，情報処理学会，2 月 1998.