情報検索におけるユーザモデルとその利用*

4Q－9

宮崎 哲夫† 太田 浩人‡ 古城 則道§
学習情報通信システム研究所¶

1 はじめに

情報検索では，ユーザは検索要求をキーワードの論理式の形で入力し要求することが多い．このとき，ユーザは自分の要求をシステムの持つ索引語で的確に表現できるとは限らない．これは，用語の持つ意味が人によって異なる場合や同義／多義などの言葉の多様性による場合もある．さらには，ユーザが欲しい情報そのものが明確でない場合や的確な用語が思い浮かばない場合もある．このようにユーザの入力したキーワードは曖昧さを含むことが多く，本報告では，このような状況の対応として，情報検索の場面において，個々のユーザの持つ用語や概念の連想特性 [1] および興味関心領域 [2] をユーザモデルとして構築し，それを利用したユーザ支援機構の枠組についての提案を行う．

2 ユーザモデル

ここでは，情報検索においてユーザの入力キーワードを拡張するための連想特性のユーザモデル，大量の検索結果をフィルタリングするための興味関心領域のユーザモデルに関して述べる．

2.1 連想特性のユーザモデリング

キーワード検索では，ユーザの入力したキーワードをもとに関連文書が提示され，その関連文書の中から1つを選択する．この情報をユーザの概念体系における入力キーワードと選んだ文書との関連としてとらえる．キーワードと文書の想起関係は，文書につけられている索引語を用いると，用語空間における想起関係にマッピングできる．用語間の関連性は，文書中の共起頻度を利用して測り，また，連想の強さは実際にユーザが用語を結びつけた回数に関係していると考え，ユーザの用語 k_i から用語 k_j の連想度 $R_{k_i k_j}$ は共起度 $C_{k_i k_j}$ および想起度 $A_{k_i k_j}$ を用いて

$$R_{k_i k_j} = C_{k_i k_j} \times A_{k_i k_j} \tag{1}$$

のように定義する．

用語 k_i と k_j の共起度は，それまでにユーザが読んだ文書中での用語 k_i と k_j の出現状況をもとに

$$C_{k_i k_j} = \frac{N_{k_i k_j}}{N_{k_i} + N_{k_j} - N_{k_i k_j}} \tag{2}$$

により定義される．ここで，N_{k_i} と N_{k_j} はそれぞれ用語 k_i および k_j を索引語とする文書数，$N_{k_i k_j}$ は k_i および k_j の両方を索引語に持つ文書数である．

想起度は，用語 k_i から 用語 k_j を想起する度合であり，入力キーワードに用語 k_i，選択文書の索引語に用語 k_j が現れた回数をカウントした想起回数 $M_{k_i k_j}$ を用いて，想起度を

$$A_{k_i k_j} = \frac{M_{k_i k_j}}{\sum_{k_x} M_{k_i k_x}} \tag{3}$$

で定義する．

このように定義した連想リンクは，ユーザの連想特性であり，用語の連想ネットワークとなる．これを連想特性ユーザモデルと定義する．

2.2 興味関心領域のユーザモデリング

ユーザのある時点における興味関心を，その時点でもっともよく見ている文書カテゴリで定義し興味関心領域と呼ぶことにする [2]．以下に，ユーザの検索した文書を逐次的にカテゴリに分類することにより，文書検索履歴をカテゴリ検索履歴に翻訳し，ユーザの各時点での興味関心領域を推定する手法を述べる．

2.2.1 検索文書の逐次的分類

ユーザが読んだ文書をカテゴリに分類する．ここでは，逐次的にカテゴリを生成する分類手法を用いる．

カテゴリが複数個あり，たとえばカテゴリ C^i の特徴ベクトルを $\vec{C}^i$ とする．ユーザが文書 $\vec{D}$ を読んだ場合，まず文書との類似度が最大となるカテゴリ $C^{\max}$ を選定する．次にこの類似度がしきい値 ϵ 以上ならば，文書はそのカテゴリに属するとし，カテゴリベクトルを

$$\vec{C}^{\max} = \frac{n_{c(\max)}\vec{C}^{\max} + \vec{D}}{n_{c(\max)} + 1} \tag{4}$$

$$n_{c(\max)} = n_{c(\max)} + 1 \tag{5}$$

---

*User Model and its Applications on Information Retrieval
†Tetsuo Miyazaki
‡Hiroto Ohta
§Norimichi Kojo
¶Software Research Laboratory

のように変更する．ここで，$n_{c(\max)}$はカテゴリ $C^{\max}$ に属する文書数である．他方，最大類似度がしきい値 ϵ 未満ならば，新しいカテゴリ C^{new}

$$\vec{C}^{\mathrm{new}} = \vec{D} \qquad n_{c(\mathrm{new})} = 1 \tag{6}$$

を生成する．以上の手続きにより，ユーザの読んだ文書は逐次分類される．

2.2.2 興味関心領域の推定

ユーザのカテゴリ検索履歴から現時点の興味関心領域を求める．このとき，現在みている情報の重みを 1 とし，過去の検索文書の寄与に対して一定の忘却率 [3] に従って減少する重みを設定する．

忘却率をγ $(0 \leq \gamma \leq 1)$，学習の開始時を 0，現時点を T，ユーザが t 時点に読んだ文書を $D(t)$，その文書が属するカテゴリを $G(D(t))$ とすると，ユーザがカテゴリ C に興味をもっている指数は

$$I(C) = \sum_{t=0}^{T} \gamma^{T-t} \delta(G(D(t)), C) \tag{7}$$

で測られる．ただし，$\delta(G, C)$ は，カテゴリ G がカテゴリ C と同じ場合には 1，違う場合には 0 の値となる関数である．このとき，現時点の興味関心領域は，指数 $I(C)$ の大きい順に並べたカテゴリのリストで得られる．

以上のようにカテゴリを興味のある順序で並べられたリストを興味関心領域ユーザモデルと定義する．

3 ユーザモデルに基づく検索支援

ここで提案したユーザモデルを用いる検索支援の概要は図 1に示す．検索支援に用いるリソースは，抽象性の高い単語を用語とした用語集合，および，文書の意味内容を用語によって索引付けした文書索引テーブルの 2 つである．

3.1 連想によるキーワード拡張

ユーザの入力したキーワードにより，連想ネットワーク内が活性化される．活性は，連想ネットワークのリンクを通じて伝播する．連想ネットワークを行列 R，しきい値以上の活性度を持つ用語を残す処理を射影関数 $P(x)$ で表現すると，1 伝搬後の活性用語は

$$\vec{a}(t+1) = \vec{a}(t) + P(\vec{a}(t) \cdot R) \tag{8}$$

により変化する．ここで，$\vec{a}(t)$ は t 時点における活性用語集合を表すベクトルである．このような活性伝搬に従い作られる連想鎖 [4] の平衡状態をユーザの入力キーワードに対する関連キーワードとして定義する．

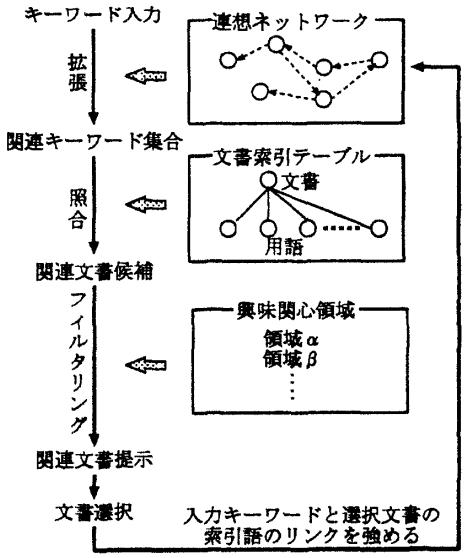

図 1: 検索支援の流れ

3.2 関連文書の選定

入力キーワードに対する関連キーワードが決まると，文書索引テーブルを参照して，関連キーワードを含む関連文書が選定される．

3.3 興味関心領域によるフィルタリング

選定された関連文書は，興味関心領域推定で作られた各カテゴリとの類似度を計算し，最大類似度をもつカテゴリに割りふる．分類された文書は，興味関心領域が与えるカテゴリ順に順序付けされ，各カテゴリ内についてはカテゴリとの類似度に従い順序付けされる．

4 おわりに

本報告では，ユーザの連想特性および興味関心領域を用いた検索支援の枠組についての提案を行った．

今後の課題として，ここで提案したユーザモデルの有効性の検証，逐次分類アルゴリズムの検討などがある．

参考文献

[1] 芽島 路小，岡本敏雄，連想にもとづく文献検索システム，信学技報，ET95-9，pp.63-70，(1995)．

[2] 郭 斑，古城 則道，宮崎 哲夫，古家 弘幸，自由探索型学習システムの学習者モデル構築とその利用，電子情報通信学会誌，DII，Vol.J79-D-II，No.9，pp.1619-1623，(1996)．

[3] 野本 豊裕，松田 憲幸，平嶋 宗，豊田 順一，文脈情報に基づくブラウジング支援-Context Sensitive Filtering-，第 11 回人工知能学会全国大会論文集，pp.460-463，1997.

[4] 小淵 洋一，斉藤 隆，布目 英修，意味素によるファジー・シソーラスの自動生成 (2)，第 7 回人工知能学会全国大会論文集，pp.321-324，1993.