フィードバックを用いる情報検索の質問修正モデル

酒井, 浩
九州大学大学院総合理工学研究科情報システム学専攻

原口, 誠
九州大学大学院総合理工学研究科情報システム学専攻

武谷, 峻一
九州大学大学院総合理工学研究科情報システム学専攻

加納, 省吾
北九州大学商学部 | 九州大学大学院総合理工学研究科情報システム学専攻

https://doi.org/10.15017/17612

出版情報：九州大学大学院総合理工学報告. 6 (2), pp.235-242, 1985-01-01. 九州大学大学院総合理工学研究科
バージョン：
権利関係：

フィードバックを用いる情報検索の質問修正モデル

酒井　浩*・原口　誠**
武谷峻一**・加納省吾***

(1984年9月26日　受理)

A Query Adjustment Model of Information Retrieval System with Feedback Process

Hiroshi SAKAI, Makoto HARAGUCHI,
Shun-ichi TAKEYA and Seigo KANŌ

A mathematical model of information retrieval system with feedback process is presented. The model mainly treats the query adjustment problem and employs multi-stage learning process in which the system guesses user's intention of retrieval by means of the document-term matrix and the documents evaluated by user. Precise formulation of the model based on fuzzy theory, and two types of the model, the one reflecting the history of retrieval and the other not, are described.

1. はじめに

今日，電子計算機を用いて情報の蓄積と検索を扱うシステムとしては，情報検索システムを始めとして情報管理システム，データベース管理システム，意思決定支援システム，質問応答システムなどがある．その中で情報検索システムは，一般に文献などの自然言語で記述された文書（一次情報）の検索を目的としており，そのために書誌事項やキーワードなどの二次情報で一次情報を特徴づける機能を有する．この意味で情報検索は狭義には文献検索（document retrieval）と同じと見られるが，一般には広く文献以外の情報の検索にも適用できる．

現在実際に多くの場所で情報検索システムは稼動しているが，その検索方式は大きく2つに分類できる．一つはブール代数を基本とした論理型方式であり，他方は文献をベクトルで表現し，対象となる文献集合に順序を付けて上位から検索結果とする関連度型方式である．商用で稼動している方式は前者の論理型方式がほとんどであるが，関連度型方式は統計的決定理論による Chow のパラメーター推定モデル[1] のように数学的に興味ある研究対象となっている．

本論では，関連度型方式に基づく検索システムで，利用者の評価を基に質問（query）を逐次的に改良していく問題を扱う．これは，システムが利用者の検索意図を表現するように索引語を逐次学習して行く過程のモデル化とも見なすことができる．このような問題は一般に質問修正問題（query adjustment problem[1)2)3)4)]）と呼ばれるが，ここでは履歴を含めた形の多段階検索（過程）の定式化とその表現を考える．そして，決定問題における基本的概念である次の2点[5] を考慮した．

(1) 途中の段階では検索結果は与えられず，最後の段階に与えられる．
(2) 各段階でその都度検索結果が与えられるばかりでなく，それがそれ以後の段階の質問決定の前提となる（動的な扱い）．

実際に人が行う多段階検索は(2)に近いと考えられるにもかかわらず，この過程で扱っている論文は少い．

以後Dは文献集合，Tは索引語集合，dは一つの文献，tは一つの索引語を表すとする．また基本的に次の前提に置く．

(1) 文献は重み付きの索引語集合で表現される．
(2) 利用者は索引語及び索引付けに関する予備的な知識を持たず，システムが提示した文献に [0, 1]

* 情報システム学専攻博士後期課程
** 情報システム学専攻
*** 情報システム学専攻（現在北九州大学商学部）

の値（大きい程関連がある）を与えることで，検索を進めてゆく．

(3) 索引語を媒介として一つの文献から別の文献を見い出すことを類推といい，要因は文献—索引語行列として与えられる．

2. 検索モデル

2.1. ではまず文献—索引語行列の性質を示し，2.2. では検索の対象とすべき文献集合と索引語集合を明確にし，検索過程における履歴を定義する．2.3. でファジィの意味での重み付けを示し，2.4. では多段階検索（過程）Ψ の定式化を行う．2.5. と2.6. ではそれぞれ履歴を持たない場合と持つ場合のモデルを示す．

2.1. 文献表現と文献—索引語行列の性質

類推の要因である文献—索引語行列の性質などについて調べておく．

〔定義 1〕

$M(d, t)$ を文献 d 上での索引語 t の表現値とし，次の式を満たす．$M(d,t)$ を成分とする行列 Γ を文献—索引語行列という．I は重み付き索引である．

$$I(d)=\{t|M(d,t)>0, t\in T\}, \tag{1}$$

$$\sum_{t\in I(d)} M(d,t)=1. \tag{2}$$

また集合 $D'\subset D$ に対しては，

$$I(D')=\cup_{d\in D'} I(d). \tag{3}$$

Γ 上でもし二つの文献が同じ索引語を持てば，その意味において二つの文献が似ていると考えられる．このことより，次の定義を行う：

〔定義 2〕

$$C^1(d_i)=\cup_{t\in I(d_i)}\{d|M(d,t)>0, d\in D\}. \tag{4}$$

$$C^n(d_i)=\cup_{t\in I(C^{n-1}(d_i))}\{d|M(d,t)>0, d\in D\}. \tag{5}$$

$C^1(d_i)$ は d_i と同じ索引語を共有する文献の集合であり，d_i と最も良く似た文献の集合と見ることができる．$C^n(d_i)$ を n 段階で d_i 類推可能な類と呼ぶ．$C^n(d_i)$ は単調非減少であり，ある $C(d_i)$ に収束する．その集合を d_i 類推可能な類と呼ぶ．

〔命題 1〕

$C(d_i)$ に含まれる任意の d に対しても $C(d)=C(d_i)$.

(証明)

任意の $d\in C(d_i)$ は d_i から類推されるので d_i から d への索引語の路がある．d から改めて類推を始めると，d_i は途中で出て来るから $C(d_i)\subset C(d)$．同様に，$d_i\in C(d)$ より $C(d)\subset C(d_i)$．従って，$C(d)=C(d_i)$：

〔命題 2〕

$d\notin C(d_i)$ なら $I(d)\cap I(C(d_i))=\phi$.

(証明)

$I(d)\cap I(C(d_i))\neq\phi$ とすると，ある索引語 t について $M(d,t)>0$ かつ $M(d',t)>0(^{\exists}d'\in C(d_i))$ であり，d と d' は同じ索引語を持つ．命題 1 より，

$d\in C^1(d')\subset C(d')=C(d_i)$

が成立し，$d\notin C(d_i)$ に反する．

〔命題 3〕

$C(d_i)\cap C(d_j)=\phi$ か $C(d_i)=C(d_j)$ のいずれか一方が成立する．

(証明)

$d_i\in C(d_j)$ のとき命題 1 より $C(d_i)=C(d_j)$ は明らか．$d_i\notin C(d_j)$ のとき $C(d_i)\cap C(d_j)\neq\phi$ とすると，両方の集合に含まれる少くとも一つの d' がある．d' は d_i 類推可能であり，d_j 類推可能である．つまり，d_i と d_j は相互に類推可能であり，$d_i\notin C(d_j)$ に反する：

〔命題 4〕

$I(C(d_i))\cap I(C(d_j))=\phi$ または $I(C(d_i))=I(C(d_j))$ が成立する．

(証明)

$C(d_i)=C(d_j)$ のときは $I(C(d_i))=I(C(d_j))$．命題 3 より残りの場合は $C(d_i)\cap C(d_j)=\phi$ のみである．命題2より任意の$d'\in C(d_i)$に対して，$I(d')\cap I(C(d_j))=\phi$ より，$C(d_i)$ に含まれるすべての d' で和集合を作れば，$\{\cup_{d'\in C(d_i)} I(d')\}\cap I(C(d_j))=\phi$ となりこれは，$I(C(d_i))\cap I(C(d_j))=\phi$ であることを示す．

以上の定義と命題より次の定理を得る．

〔定理 1〕

文献—索引語行列 Γ の類推可能性より，$D=\cup_{i=1}^{n} D_i$，$T=\cup_{i=1}^{n} T_i(T_i=I(D_i))$ なる分割が存在し，Γ は次のように変形できる．

$$\Gamma=\begin{bmatrix} \Gamma_1 & & 0 \\ & \Gamma_2 & \\ 0 & & \Gamma_n \end{bmatrix}. \tag{6}$$

(証明)

分割できることは，命題1から命題4より明らか．Γ の変形はまず $D_1, D_2, \cdots, D_n$ をこの順序で行の成分とし，ついで $T_1, T_2, \cdots, T_n$ をこの順序で列の成分とする．実際に，$d \in D_i$ に対して $I(d) \subset I(D_i)$ より $T \setminus T_i$ に含まれる索引語 t について $M(d,t)=0$ となる．

この定理は，Γ を用いる検索過程での類推の限界を示している．つまり，Γ_i からは $\Gamma_j (i \neq j)$ についての何の情報も得られない．もし利用者の持つ初期情報がすべて Γ_i に含まれるなら，以後 Γ でなく Γ_i を文献—索引語行列と考えても良いことになる．

2.2. 利用者評価と対象となる文献，索引語集合

利用者の文献に対する評価と検索過程の履歴などについて定義する．

〔定義3〕

利用者は検索目的に従って，文献 d に対して [0, 1] の値を付ける．この操作を ω とし，

$$\omega: D \to [0,1]. \qquad (7)$$

これは，検索目的の文献集合に対して，ある文献がどの程度その集合に含まれるかを示すファジィ集合のメンバシップ関数に対応する．また ω 値が付けられた文献 d に対する $I(d)$ をシステムが確認した索引語と呼ぶことにする．

〔定義4〕

X^i を利用者が第 $(i+1)$ 段階で評価する文献集合とし，

$$Y^n = \cup_{i=1}^{n} X^i \quad (Y^{n-1} \cap X^n = \phi) \qquad (8)$$

と書く．このように数段階に及び検索することを多段階検索[6]という．

〔定義5〕

索引語 t' が文献集合 K 上で不要であるとは，

$$\omega(d)=0 \quad \text{for } {}^{\forall}d \in \{d \mid d \in K, M(d,t')>0\} \qquad (9)$$

なることである．

定義5より，多段階検索として二つの場合を考えることができる．それは，検索の対象として考えておくべき集合に対する次の二つのアルゴリズムで与えられる．

〔アルゴリズム1〕

begin

n : $=1$;
T^1: $=T$;
D^1: $=D$;
repeat: 利用者の評価 $(\omega(d),\ d \in X^n)$;
T^{n+1}: $=T^1 \setminus \{Y^n$ で不要である索引語$\}$;
D^{n+1}: $=D^1 \setminus \{d \mid I(d) \cap T^{n+1} = \phi\}$;
if $I(Y^n)=T^1$ **then** 終了;
n: $=n+1$;
go to repeat
end

〔アルゴリズム2〕

begin

n : $=1$;
T^1: $=T$;
D^1: $=D$;
repeat: 利用者の評価 $(\omega(d),\ d \in X^n)$;
T^{n+1}: $=T^n \setminus \{X^n$ で不要である索引語$\}$;
D^{n+1}: $=D^n \setminus \{d \mid I(d) \cap T^{n+1} = \phi\}$;
if $I(Y^n)=T^1$ **then** 終了;
n : $=n+1$;
go to repeat
end

〔命題5〕

アルゴリズム2においては次のことが成立する．

$$T^{n+1} \subset T^n,\ D^{n+1} \subset D^n,\ (n \geq 1). \qquad (10)$$

(証明)

定義より明らか．

命題5における性質はアルゴリズム1で成立するとは限らない．つまり，Y^n を一度に全部評価してしまうことと同じであり，多段階検索の意味がない．それに対してアルゴリズム2では，一度不要とされた索引語は検索の対象から外される．このような意味で段階の影響が考えられる．次の定義を行う．

〔定義6〕

多段階検索において履歴を持つとは，各段階で検索の対象とすべき文献と索引語の集合の決定がアルゴリズム2に従って決定されることをいう．

〔命題6〕

多段階検索において，履歴を持つ場合の文献と索引語の集合を D_c^n と T_c^n とし，持たない場合を D_e^n と T_e^n とすると，

$$T_c^n \subset T_e^n,\ D_c^n \subset D_e^n,\ (n \geq 2). \qquad (11)$$

(証明)

$t \in \{Y^n$ で不要である索引語$\}$

$\Leftrightarrow \omega(d)=0$ for $^{\forall}d\in\{d|d\in Y^n, M(d,t)>0\}$

$\Rightarrow {}^{\exists}X^i$ に対して $t\in\{X^i$ で不要である索引語$\}$

$\Leftrightarrow t\in\cup_{i=1}^{n}\{X^i$ で不要である索引語$\}$.

つまり，$\{Y^n$ で不要である索引語$\}\subset\cup_{i=1}^{n}\{X^i$ で不要である索引語$\}$ となり，

$T_c^n=T^1\diagdown\cup_{i=1}^{n-1}\{X^i$ で不要である索引語$\}$

$\subset T^1\diagdown\{Y^n$ で不要である索引語$\}$

$=T_e^n$.

$D_c^n\subset D_e^n$ についても同様に証明される．

各段階で対象とすべき文献集合と索引語集合は，履歴を持つ場合の方が持たない場合より確実に小さくなることがわかる．

2.3. ファジィ重み付け

ここでは，利用者の評価を基にした索引語の重み付けを考える．

〔定義 7〕

$$H^n(t)=Y^{n-1}\cap\{d|M(d,t)>0, d\in D\}. \quad (12)$$

$H^n(t)$ は現在までに評価された文献集合の中で索引語 t をもつ集合である．もちろん，$H^n(t)=\phi$ もある．このとき，ファジィの理論[7]に従って次の定義を行う．

〔定義 8〕

第 n 段階での索引語 t の重みを以下とする．

$$w^n(t)=\begin{cases}\sum_{d\in H^{n-1}(t)}M'(d,t)\times\omega(d) & (H^{n-1}(t)\neq\phi)\\ \text{未定義} & (H^{n-1}(t)=\phi)\end{cases} \quad (13)$$

$$M'(d,t)=M(d,t)/\sum_{d'\in H^{n-1}(t)}M(d',t).$$

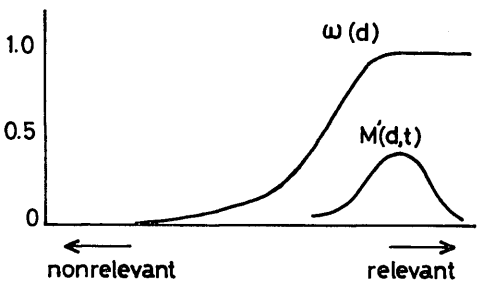

Fig. 1 Evaluation of documents and indexing-terms

$w^n(t)$ は定義式の右辺の値の確率で関連ある文献を表現していることになり，その値を重みとする．1索引語の重みを決定するのは $H^n(t)(\subset I^{-1}(t))$ であり，評価値を索引付けの強さに関して平均化した形である．段階数 n が大きくなる程システムが確認する索引語の個数は増加し，検索目的をよりよく表現するような索引語が見い出され，重みが付けられる．不要な索引語には重み 0 が付く．

2.4. 多段階検索（過程）Ψ の定式化

ここでは，今まで示した性質を利用して多段階検索（過程）Ψ を定式化し，解析する．Ψ は利用者の評価により質問を改良して行くフィードバック系である．

〔定義 9〕

多段階検索（過程）Ψ は次の形で与えられ，二つの局面をもつ．

$$\Psi=((T,D,\Gamma,F),G,\Lambda,\omega), \quad (14)$$

(T,D,Γ,F)：ある関連度型検索システム（Γ：索引語行列，F：検索式 は後に示す），

G：質問を修正するアルゴリズム，

Λ：$\{\cup_{i=1}^{n}X^i|X^i\subset D,\ Y^{n-1}\cap X^n=\phi\}$

評価される文献集合，

ω：利用者の評価．

（第 1 局面）

利用者は索引語の集合Mを入力し，これにより Γ 中の必要な類を決定し，対象とする文献と索引語集合を D^1 と T^1 する．X^1 として $\cup_{t\in M}I^{-1}(t)$ の部分集合を取る．

（第 2 局面）

ω と G により逐次的に質問を修正する（**Fig. 2** 参照）．実際には G の具体的な与え方が問題である．G の与え方によりアルゴリズム 1 と 2 のそれぞれに対応する形で，T^n と D^n が決定されるモデルを考える．以下，具体的に各要因の定義を行う．

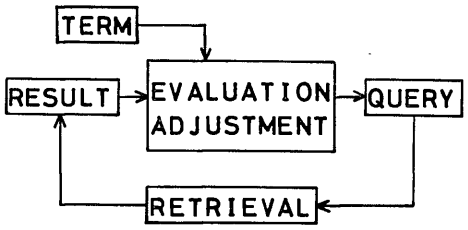

Fig. 2 Multi-stage information retrieval process Ψ

〔定義 10〕（質問形式）

第 n 段階で (13) 式により計算された索引語の重みを基に，$Q^n(t)$ $(n \geq 2)$ を定義し質問とする．

$$Q^n(t) = \begin{cases} w^n(t)/\sum_{t' \in I(Y^{n-1})} w^n(t') & (t \in I(Y^{n-1})) \\ 1/\#|T^1| \ (=Q^1(t)) & (t \notin I(Y^{n-1})) \end{cases} \tag{15}$$

ただし，$\sum_{t \in T^1} Q^n(t) = 1$ を満たす．

すなわち，質問を索引語集合での確率分布として定義する．各段階では，必要な索引語が不要な索引語の占めていた重みを取り上げるような形で質問が生成されていくことになる．

〔定義 11〕(検索式 F)

第 n 段階における質問に対する文献 d の検索値 $F^n(d)$ を次で与える．

$$\begin{aligned} F^n(d) &= \sum_{t \in I(d)} M(d,t) \times Q^n(t) \\ &= \sum_{t \in I(d,1)} M(d,t) \times Q^n(t) \\ &\quad + \sum_{t \in I(d,2)} M(d,t) \times Q^1(t) , \end{aligned} \tag{16}$$

ただし，$I(d,1) = \{t \mid t \in I(d) \cap I(Y^{n-1})\}$，

$I(d,2) = I(d) \setminus I(d,1)$.

$F^n(d)$ を各 d について計算し，$F^n(d) > 1/\#|T^1|$ を満足する文献を上位から検索結果とする．

〔命題 7〕

$$\min_{t \in I(d)} Q^n(t) \leqq F^n(d) \leqq \max_{t \in I(d)} Q^n(t) . \tag{17}$$

特に，$I(d,1) = \phi$ のときは，

$$F^n(d) = 1/\#|T^1| . \tag{18}$$

(証明)

$$\begin{aligned} F^n(d) &= \sum_{t \in I(d)} M(d,t) \times Q^n(t) \\ &\leqq \max_{t \in I(d)} Q^n(t) \times \sum_{t \in I(d)} M(d,t) \\ &= \max_{t \in I(d)} Q^n(t) . \end{aligned}$$

min についても同様．特に $I(d,1) = \phi$ の時は，

$$\begin{aligned} F^n(d) &= \sum_{t \in I(d)} M(d,t) \times Q^1(t) \\ &= Q^1(t) = 1/\#|T^1| . \end{aligned}$$

検索式で $F^n(d) > 1/\#|T^1|$ は，(18) からわかるように初期段階より確実に検索値が増加していることを示す．今まで示した Ψ の各要因の関係を **Fig. 3** に示す．

具体的アルゴリズム G は，次の二つのモデルで示す．

2.5. Ψ の履歴のないモデル

2.4. で多段階検索 Ψ の定義を行った．ここでは，アルゴリズム 1 の検索方式で対象となる文献と索引語集合が決定されるモデル ($Q^n(t) = 0, {}^\forall t \notin T^n$ を満足)

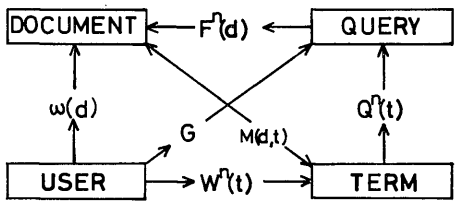

Fig. 3 Relations among factors in Ψ

を示す．

モデル I (履歴なし)

質問を作るアルゴリズム G を定義 10 で与える：この時，全体のアルゴリズムは次の形となる．

〔アルゴリズム 3〕

begin

comment D^1, T^1 は決定しているとする；

$Q^1(t_i) := 1/\#|T^1|$,

$(1 \leq i \leq \#|T^1|)$;

$Y^0 := \phi$;

$n := 1$;

start: 利用者評価 $(\omega(d), d \in X^n)$;

$Y^n := Y^{n-1} \cup X^n$;

weight: calculate $w^{n+1}(t)$ $(t \in I(Y^n))$;

query: $Q^{n+1}(t)$

$$= \begin{cases} w^{n+1}(t)/\sum_{t' \in I(Y^n)} w^{n+1}(t') & (t \in I(Y^n)) \\ Q^1(t) & (t \notin I(Y^n)) \\ (\sum_{t \in T^1} Q^{n+1}(t) = 1) \end{cases}$$

retrieve: $F^{n+1}(d) > 1/\#|T^1|$ なる集合を検索；

$n := n+1$;

$X^n \subset \{d \mid F^n(d) > 1/\#|T^1|, d \notin Y^{n-1}\}$ なる X^n があれば取る；

if $X^n \neq \phi$ **then goto** start

end

〔命題 8〕

$Q^n(t)$ はアルゴリズム 1 で定義した T^n に対して，

$$\begin{aligned} &\text{i)} \quad Q^n(t) > 0 \quad (t \in T^n), \\ &\text{ii)} \quad Q^n(t) = 0 \quad (t \notin T^n) . \end{aligned} \tag{19}$$

(証明)

任意の $t \in T^1 \setminus T^n$ に対して定義より，

$w^n(t) = \sum_{d \in H^{n-1}(t)} M'(d,t) \times \omega(d)$

$= 0$ $(\because \omega(d) = 0$ for ${}^\forall d \in H^{n-1}(t))$.

$t \in T^n$ なる t では，$w^n(t) > 0$ より $Q^n(t) > 0$ か，$Q^n(t) = Q^1(t) > 0$.

検索値は，質問と表現値との期待値を取る形だから，$T^1 \setminus T^n$ なる集合は検索に影響を与えない．

実際にモデルⅠのアルゴリズムで生成される質問は，

$$Q^n(t)=\begin{cases}\dfrac{\#|I(Y^{n-1})|}{\#|T^1|}\times\dfrac{w^n(t)}{\sum_{t'\in I(Y^{n-1})}w^n(t')} & (t\in I(Y^{n-1}))\\ 1/\#|T^1|\ (=Q^1(t)) & (t\notin I(Y^{n-1}))\end{cases} \tag{20}$$

となる．先にも触れたように $w^n(t)$ は $H^{n-1}(t)$ $(\subset I^{-1}(t))$ の評価を $M(d,t)$ の大きさで平均化しており，$H^{n-1}(t)$ の評価が高ければ $Q^n(t)$ は大きくなる．ところが上の式において "$Q^n(t)=0 \Rightarrow Q^{n+1}(t)=0$" なる性質は成立するとは限らない．つまり，履歴を持たないことになる．この点を改良したのが次に示すモデルⅡである．

2.6. Ψの履歴を持つモデル

2.5. では履歴のないモデルⅠを示したが，次に "$Q^n(t)=0 \Rightarrow Q^{n+1}(t)=0$" が保証されるモデルⅡを示す．

モデルⅡ（履歴あり）

質問を作るアルゴリズムGを次のようにする．

begin

calculate　$w^n(t)$ $(t\in I(Y^{n-1}))$;

$\tilde{w}^n(t) := w^n(t)\times Q^{n-1}(t)$ $(t\in I(Y^{n-1}))$;

$$Q^n(t) := \begin{cases}\tilde{w}^n(t)/\sum_{t'\in I(Y^{n-1})}\tilde{w}^n(t') & (t\in I(Y^{n-1}))\\ 1/\#|T^1|\ (=Q^1(t)) & (t\notin I(Y^{n-1}))\end{cases}$$

end

つまり，第 n 段階での重みを第 $(n-1)$ 段階での質問との期待値で定義する．モデルⅡの全体のアルゴリズムは，モデルⅠの場合の query の所を上に示した G でおき直した形である．次の命題を得る．

〔命題 9〕

$Q^n(t)$ はアルゴリズム 2 で示した T^n に対して，

i)　$Q^n(t)>0$　$(t\in T^n)$，

ii)　$Q^n(t)=0$　$(t\notin T^n)$，　(21)

iii)　$Q^n(t)=0$ なら $Q^{n+1}(t)=0$　$(n\geqq 2)$．

が成立する．

(証明)

$T^n=T^{n-1}\setminus\{X^{n-1}$ で不要である索引語$\}$ より，$t\notin T^n$ なる索引語はある $X^m(1\leqq m\leqq n-1)$ において，$\omega(d)=0$ for ${}^\forall d\in H^m(t)$ であり，$w^{m+1}(t)=0$ より $Q^{m+1}(t)=0$．従って，$\tilde{w}^{m+2}(t)=w^{m+2}(t)\times Q^{m+1}(t)=0$ より $Q^{m+2}(t)=0$．逐次的に ii)，iii) が示せる．

命題 8 と比べると iii) の条件が加わった．これで，一度不要とされた索引語は以後の検索には関係しない．

次に，このモデルで逐次的に生成される質問を考える．今度はモデルⅠの質問より詳しくなり，どの段階で索引語がシステムに確認されたのかまで問題となる．特に $n=2$ では，

$$Q^2(t)=\begin{cases}\dfrac{\#|I(Y^1)|}{\#|T^1|}\times\dfrac{w^2(t)}{\sum_{t'\in I(Y^1)}w^2(t')} & (t\in I(Y^1))\\ 1/\#|T^1|\ (=Q^1(t)) & (t\notin I(Y^1))\end{cases} \tag{22}$$

$n=3$ では，

$$\tilde{w}^3(t)=\begin{cases}w^3(t)\times Q^2(t) & (t\in L(2))\\ w^3(t)\times Q^1(t) & (t\in L(3))\\ \text{未定義} & (t\notin I(Y^2))\end{cases} \tag{23}$$

$L(i)$: 第 i 段階で初めてシステムに確認された索引語の集合．

$\tilde{w}^3(t)$ により $Q^3(t)$ を計算すると，

$$Q^3(t)=\begin{cases}\dfrac{\#|I(Y^2)|}{\#|T^1|}\times\dfrac{\Pi_{i=2}^3 w^i(t)}{\sum_{k=2}^3\{\sum_{t'\in L(k)}\Pi_{i=k}^3 w^i(t')\cdot E[Y^{k-1}]\}} & (t\in L(2))\\ \dfrac{\#|I(Y^2)|}{\#|T^1|}\times\dfrac{w^3(t)\cdot E[Y^2]}{\sum_{k=2}^3\{\sum_{t'\in L(k)}\Pi_{i=k}^3 w^i(t')\cdot E[Y^{k-1}]\}} & (t\in L(3))\\ 1/\#|T^1|\ (=Q^1(t)) & (t\notin I(Y^2))\end{cases} \tag{24}$$

ただし，$E[Y^2]=(1/\#|I(Y^2)|)\times\sum_{t\in I(Y^2)}w^2(t)$，

$E[Y^1]=1$．

一般には次の形になる．

〔定理 2〕

第 n 段階での質問は次の式で与えられる索引語集合上での確率分布となる．

$$Q^n(t)=\begin{cases}\dfrac{\#|I(Y^{n-1})|}{\#|T^1|}\times\dfrac{\Pi_{l=2}^{n}w^l(t)}{\sum_{k=2}^{n}\{\sum_{t'\in L(k)}\Pi_{l=k}^{n}w^l(t')\cdot E[Y^{k-1}]\}} & (t\in L(2))\\ \quad\vdots \qquad\qquad \vdots & \quad\vdots\\ \dfrac{\#|I(Y^{n-1})|}{\#|T^1|}\times\dfrac{\Pi_{l=m}^{n}w^l(t)\cdot E[Y^{m-1}]}{\sum_{k=2}^{n}\{\sum_{t'\in L(k)}\Pi_{l=k}^{n}w^l(t')\cdot E[Y^{k-1}]\}} & (t\in L(m))\\ \quad\vdots \qquad\qquad \vdots & \quad\vdots\\ \dfrac{\#|I(Y^{n-1})|}{\#|T^1|}\times\dfrac{w^n(t)\cdot E[Y^{n-1}]}{\sum_{k=2}^{n}\{\sum_{t\in L(k)}\Pi_{l=k}^{n}w^l(t')\cdot E[Y^{k-1}]\}} & (t\in L(n))\\ 1/\#|T^1|\ (=Q^1(t)) & (t\notin I(Y^{n-1}))\end{cases}\tag{25}$$

ただし，$E[Y^n]=\sum_{k=2}^{n}\{\sum_{t\in L(k)}\Pi_{l=k}^{n}w^l(t)\cdot E[Y^{k-1}]\}/\#|I(Y^n)|$,

$E[Y^1]=1$.

(証明)

帰納法によるが，$Q^2(t)$, $Q^3(t)$ は前に示した通り，$t^*\in L(S)$ としておく．定義より，

$$Q^{n+1}(t^*)=\frac{\#|I(Y^n)|}{\#|T^1|}\times\frac{w^{n+1}(t^*)\times Q^n(t^*)}{\sum_{t\in I(Y^n)}w^{n+1}(t)\times Q^n(t)}$$

$t'\in L(m)$ $(1\leqq m\leqq n+1)$ とすると，

$$Q^n(t')=\frac{\#|I(Y^{n-1})|}{\#|T^1|}\times\frac{\Pi_{l=m}^{n}w^l(t')\cdot E[Y^{m-1}]}{\sum_{k=2}^{n}\{\sum_{t\in L(k)}\Pi_{l=k}^{n}w^l(t)\cdot E[Y^{k-1}]\}}$$

とかけるから，$Q^n(t)$ の所に代入すると，

$$\begin{aligned}Q^{n+1}(t^*)&=\frac{\#|I(Y^n)|}{\#|T^1|}\times\frac{w^{n+1}(t^*)\cdot\{\#|I(Y^{n-1})|\cdot\Pi_{l=s}^{n}w^l(t^*)\cdot E[Y^{s-1}]\}}{\sum_{t\in I(Y^n)}\cdot w^{n+1}(t)\cdot\{\#|I(Y^{n-1})|\cdot\Pi_{l=k}^{n}w^l(t)\cdot E[Y^{k-1}]\}}\\&=\frac{\#|I(Y^n)|}{\#|T^1|}\times\frac{\Pi_{l=s}^{n+1}w^l(t^*)\cdot E[Y^{s-1}]}{\sum_{k=2}^{n+1}\{\sum_{t\in L(k)}\Pi_{l=k}^{n+1}w^l(t)\cdot E[Y^{k-1}]\}}\end{aligned}$$

となり証明できた．

上の式では，1つの索引語がシステムに確認されてから以後は，各段階で計算できる重みを積の形で付加している．この式では，索引語がシステムに確認された段階数まで関係してくる．

3. おわりに

本論文では，多段階検索（過程）における各段階の意味を明確にすること，つまり履歴を考えることによって，動的に変化して行く質問の形を考えた．似たモデルとして Salton のrelevance feedback モデルが考えられている[2]．実際に，モデルⅠは利用者の評価から個々の索引語の重みを計算するだけで，基本的には Salton のモデルとほとんど変らない．しかし，モデルⅡは対象とする索引語がシステムに確認される段階まで問題となり，かつ $Q^n(t)=0$ なら $Q^{n+1}(t)=0$ が保証される．このような意味で，検索過程を詳しく扱ったモデルになっていると思われる[8][9][10]．

また，前半で示した文献—索引語行列 Γ の性質は，転置ファイルを用いるいかなる検索方式でも基本となることである．質問の重み付けは，ファジィの理論に基づいており必要とする索引語にのみ重みが付く．このことで検索に関係する索引語と関係しない索引語との判別はできる．しかし，質問の最適性，収束性についてはかなり難しく，現在の所はうまく示せていない．関連度方式での最適性をうまく扱っているのは，Chow のモデル[1] くらいと考えられる．

論理型検索方式における質問修正については，シソーラスを用いる場合に関して，その定式化を考慮中である[11]．

参考文献

1) Chow, D. and Yu, C. T.: On the construction of Feedback Query, J. ACM, vol. 29, No. 1, pp. 127–151 (1982).
2) Salton, G. and McGill, M. J.: Introduction to Modern Information Retrieval, p. 448, McGraw-Hill, New York (1983).
3) Vernimb, C.: Automatic Query Adjustment in Document Retrieval, Inform. Process. Manage., vol. 13, No. 6, pp. 339–353 (1977).
4) Driyanskii, V. M.: Retrieval Models in Online Documentary Information Sys-

tems: An Analytic Review, Cybernetics., vol. 17, No. 2, pp. 269–287 (1981).
5) 日本数学会編集: 数学辞典, p. 1140, 岩波書店, 東京 (1968).
6) Arikawa, S. and Kitagawa, T.: Multistage Information Retrieval System Based upon Researcher Files, Res. Rep. Fund. Inform. Sci., Kyushu Univ., No. 51 (1975).
7) 浅井ほか: あいまいシステム理論入門, p.269, オーム社, 東京 (1978).
8) 酒井ほか: フィードバックを用いる情報検索法の1つの定式化, 第26回情処全大 (1983).
9) 酒井ほか: フィードバックを用いる情報検索法の1つの定式化II, 第27回情処全大 (1983).
10) Sakai, H. et al: Mathematical Model of Information Retrieval with Feedback Process, Bull. Inform. Cyber., Res. Assoc. Stat. Sci., vol. 21, No. 1–2, (1984).
11) 酒井, 原口, 武谷: シソーラスを用いる情報検索の質問修正モデルについて, 第29回情処全大 (1984).