自然言語処理 108－10
(1995. 7. 20)

ユーザ評価を用いたテキスト照合パターン生成手法

岩本 秀明

NTT情報通信研究所

情報検索や情報抽出などに必要なテキスト照合処理において，その照合結果に対するユーザの適合・不適合評価から，キーワード論理式や正規表現などの照合パターンを生成する手法を提案する．本手法の前提として，まず，テキスト照合によって照合に成功した一連のテキスト情報がユーザに示される．次に，ユーザはこれらに対して適合・不適合判断を行う．本手法では，ユーザが適合と判断した適合情報とユーザが不適合と判断した不適合情報から，照合パターンを生成する．この照合パターンは，より正確な照合結果を得るために，再度テキスト照合を行うために用いる．本手法を新聞記事の標題とその第一文とに繰返し適用して，再現率が向上することを確認した．

A Method for Generating Text Matching Patterns using Relevance Feedback

Hideaki IWAMOTO

NTT Information and Communication Systems Laboratories
1-2356 Take Yokosuka 238-03 Japan
iwamoto@nttnly.isl.ntt.jp

This paper proposes a method for generating text matching patterns using relevance feedback. Text matching patterns are required for the text processing such as information retrieval and information extraction. The way of generating patterns is as follows: (1) A user was displayed a result of text processing. (2) The user classified the results into positive and negative one. Using user's evaluation, our method generates text matching patterns which discriminate positive information from negative one. We applied this method to the newspaper retrieval experiments and confirmed improvement of the recall rate.

1 はじめに

近年，Internet，LAN，パソコン通信，PC，ワープロ，CD-ROMなど様々な媒体で大量の電子化テキストが蓄積されてきている．我々の目標は，これらのテキストから，ユーザの要求にしたがって，情報の収集・整理を行うシステム [1] を構築することである．

このようなシステムにユーザの要求を伝える方法として，通常，自然言語などによるユーザの直接入力を利用する．しかしながら，自然言語による要求表現の曖昧さを解消することは極めて困難である．このため，我々は，ユーザの直接入力に加えて，ユーザの評価を用いる方法 [2] を検討している．ユーザの評価は，照合パターンによってテキストを照合した結果に対して行われる．従来のシステムでは，この照合パターンは，ユーザの直接入力から生成する．

ここで，照合とはキーワード論理式や正規表現などのパターンとテキストとを比較することにより，パターンに合致する情報の位置やその構造を同定する処理である．照合結果は，テキスト全体またはその一部からなるが，これをまとめてテキスト情報と呼ぶことにする．

本稿で提案する手法において，ユーザはシステムから提示されたテキスト情報に対して，それが要求に適合するか否かを判断する．システムでは，ユーザが適合と判断した適合情報とユーザが不適合と判断した不適合情報とから，より照合洩れおよび照合誤りの少ないパターンを生成する．

このような手法は，情報検索の分野でRelevance Feedback [3] として古くから知られている．この手法は，テキストに出現する語のベクトルを前提としている．

しかしながら，語のベクトルは，語句間の関係を直接表現できないため，照合誤りが多くなると考えられる．語句間の関係は，表面的には，語順および格標識や用言の活用形などによって表現される．

本研究の目的は，これらの語句間の関係を表す情報をパターンとして保持することによって照合誤りを減らし，それと同時に，照合パターンの生成にRelevance Feedbackを導入し斬進的に照合洩れを減らすことである．

2 照合パターン生成手法

本稿で提案する手法は，適合情報および不適合情報を入力とし照合パターンを出力する．

適合情報とは，テキスト照合処理によりユーザに提示されたテキスト情報の中で，ユーザが適合と判断したものである．それに対して，不適合情報とは，ユーザが不適合と判断したテキスト情報である．

なるべく誤りがないように，洩れているテキスト情報を照合しようとすると，その照合を行うパターンは，全ての適合情報と照合し，いずれの不適合情報とも照合せず，かつ，適合情報・不適合情報以外のテキスト情報と照合する可能性を持つ必要がある．

テキスト情報として朝日新聞の経済記事における標題およびその本文の第一文を考え，例えば，適合情報を

(2.1)【中古車販売にも陰り　5月の登録数，9ヵ月ぶり前年割れ】
新車販売が低迷する一方で，好調といわれていた中古車販売にもかげりが見え始めた．――('92/6/10)

(2.2)【7月の新車販売，再び前年割れ】
日本自動車販売協会連合会が3日まとめた軽自動車を除く7月の新車販売台数は，前年度に比べ5．6％減の54万1599台となった．――('92/8/4)

不適合情報を

(2.3)【トヨタ「GM車販売は困難」　米国側に波紋も　GMがヤナセ気遣う】
トヨタ自動車は5日，今年1月のブッシュ米大統領訪日時に表明したゼネラルモーターズ（GM）車を年間5000台販売する計画について「（輸入代理権を持つ）ヤナセとGMの合意が得られない．――('92/10/6)

とする．ただし，【】内が標題であり，本文の第二文以降は省略している．これらを入力として，

(2.4) ～新車販売～∨～中古車販売にも～

という照合パターンを生成する．ただし，"～"は，任意の1文字の繰り返しを表し，"∨"は，論理和を表すとする．

照合パターン(2.4)は，適合情報(2.1)および(2.2)とは，照合し，不適合情報(2.3)とは照合しない．

このようなパターンを生成するために，本稿で提案する手法は，以下の処理からなる．

(1) 適合情報から部分情報$_{(a)}$を取り出す．
(1') (a)の集合から不適合情報に含まれる要素を除く．
(2) (a)を組み合わせて適合パターン$_{(b)}$を構成する．
(2') (b)の集合から不適合情報と照合する要素を除く．
(3) (b)の集合から全ての適合情報と洩れなく照合する最少の組み合わせを選ぶ．

最終的な照合パターンは，処理(3)で選択した適合パターンを論理和の形式で結合して生成する．

このために，一つまたは複数の適合情報と照合する適合パターンを構成し，その中から処理(3)により，全ての適合情報と洩れなく照合し不適合情報とは照合しない適合パターンの組合せを選ぶ．

処理(2)で構成する適合パターンの集合(P)は，その要素($p_i, p_j \in P$)の間に包含関係を持つ．

すなわち，適合情報の集合を T とし，その中で適合パターン p_i および p_j と照合する適合情報の全体集合における部分集合をそれぞれ T_i および T_j とする．

ここで，パターンpでテキスト情報tを照合することを，

$$t \in L(p)$$

のように表すと，適合情報の部分集合 T_i および T_j は，それぞれ，

$$T_i \subseteq L(p_i) \text{ および } T_j \subseteq L(p_j)$$

という条件を満たす．このとき，適合情報の部分集合 T_i および T_j が次の条件

$$T_i \supseteq T_j$$

を満たすとき，パターン p_i はパターン p_j を包含すると呼び，これを

$$L(p_i) \supseteq L(p_j)$$

のように表す．また，パターンの包含関係において，包含する方のパターンは，包含されるパターンに対して上位のパターンであると定義し，包含されるパターンは，包含するパターンに対して下位のパターンであると定義する．さらに，相対的に上位にあるパターンは，共通的な，共通性が高い，または，個別性が低いパターンであるといい，相対的に下位にあるパターンは，個別的な，個別性が高い，または，共通性が低いパターンであるということにする．

(3)における処理は，適合パターンの集合から，その要素間の包含関係における最上位の適合パターンを選び出すことに相当する．ここで，不適合情報と照合するパターンは適合パターンの集合から取り除かれているものとすると，処理(3)を経て最終的に生成される照合パターンは，全ての適合情報と照合し，いずれの不適合情報とも照合しない．加えて，処理(3)では最も共通性の高いパターンを選ぶため，最終的に生成される照合パターンは，適合情報・不適合情報以外のテキスト情報と照合する可能性を持つ．

処理(1)，(2)および(3)について，それぞれ第3.1節，第3.2節および第3.3節で述べ，処理(1')および(2')については，第3.4節でまとめて議論する．

2.1 部分情報の抽出

適合情報から部分情報を抽出するために，文字のn-gramを用いる．それぞれの適合情報に対し，文字単位に1-gramからn-gramまで数え上げる．ただし，nは，ある適合情報の文字列の長さである．

例えば，適合情報(2.1)および(2.2)に対し，部分情報を抽出した結果の一部を表2.1に示す．

文字列	長さ×頻度	長さ	頻度
車販売	18	3	6
中古車販売にも	14	7	2
販売	12	2	6
車販	12	2	6
新車販売	12	4	3
中古車販売に	12	6	2
古車販売にも	12	6	2
月の新車販売	12	6	2

表2.1 部分情報の抽出結果

2.2 適合パターンの構成

適合情報と文字のn-gramによって抽出した部分情報とによって，適合パターンの集合を構成する．

まず，部分情報をその長さおよび適合情報中の頻度によって順位付けする．次に，この順序にしたがって，部分情報の組み合わせを増やしつつ，それらが適合情報に出現する順序を適合パターンとして蓄積していく．部分情報の組み合わせが増えるのにしたがって，構成されるパターンは各適合情報に個別化される．一般に，頻度を優先すると，助詞や助動詞などの付属語が強調され，長さを優先すると，複合語や文節が強調される．

適合情報(2.1)および(2.2)と表2.1に示した部分情報とを用いて，適合パターンを構成する例を示す．ただし，部分情報の長さと頻度との積が等しい場合，頻度の高いものを優先してパターンを構成する．

部分情報の組み合わせとして，まず，”車販売”を選ぶ．この”車販売”は，適合パターン上で，

(2.5)【～車販売～】　#(2.1)の標題
～車販売～車販売～　#(2.1)の第一文

【～車販売～】　#(2.2)の標題
～車販売～車販売～　#(2.2)の第一文

のように現れる．ここで二種類の適合パターン”～車販売～”と”～車販売～車販売～”とを得る．

次の部分情報の組み合わせとして，部分情報の順序にしたがって，”車販売”と”中古車販売にも”とを選び，それらは適合情報上で，

(2.6)【中古車販売にも～】　#(2.1)の標題
～車販売～中古車販売にも～　#(2.1)の第一文

【～車販売～】　　#(2.2)の標題
～車販売～車販売～　　#(2.2)の第一文

のように現れる．最初の組み合わせで得た先の二種類のパターンに加え，新たに適合パターン”中古車販売にも～”と”～車販売～中古車販売にも～”とを得る．

適合情報でのパターンが各適合情報と一対一に照合するまでこれを繰り返すと，適合情報上で，

(2.7)【中古車販売にも～月の～前年割れ】　　#(2.1)の標題
～車販売～中古車販売にも～　　#(2.1)の第一文

【～月の～車販売～前年割れ】　　#(2.2)の標題
～車販売～車～月の～車販売～前年～　#(2.1)の第一文

のようなパターンが構成され，この場合，次のような適合パターンの集合を得る．

(2.8)～車販売～
～車販売～車～月の～車販売～前年～
～月の～車販売～前年割れ
～車販売～中古車販売にも～
～車販売～前年割れ
中古車販売にも～月の～前年割れ
～車販売～車販売～
中古車販売にも～
～車販売～車～車販売～

2.3 適合パターンの組み合わせ最少化

順序付られた部分情報にしたがって構成した適合パターンの集合は，その要素の間に包含関係を持つ．適合パターンの集合(2.8)における包含関係を図2.1に示す．

全ての適合情報と洩れなく照合するパターンは，この包含関係における最上位パターンからなる．したがって，適合パターンの組み合わせを最少化することは，適合パターンの集合から，その包含関係における最上位のパターンを特定することに相当する．図2.1のパターン包含関係においては，”～車販売～”がその最上位パターンである．

適合パターンの包含関係における最上位のパターンは，複数有り得る．例えば，図2.1の包含関係において，”～車販売～”を適合パターンの集合から削除すると，残った適合パターンの包含関係は，3種類の最上位パターン”中古車販売にも～”，’’～車販売～前年割れ”および’’～車販売～車販売～”を持つ．

適合パターンの包含関係における最上位パターンは，適合パターン中，最も共通性の高いパターンである．このために，最上位パターンによる照合結果には，照合誤りが多くなる可能性がある．

このような場合，構成する個々の適合パターン自体を個別なものにするか，あるいは，適合パターンの集合において共通性の高いパターンを削除しておくことによって，最終的に生成するパターンを個別的なものにする必要がある．しかし，生成したパターンによる照合結果の誤りはユーザにしか判断できない．したがって，適合情報だけでなく，不適合情報を利用することによって，少なくともユーザにとって必要でないと明らかになったテキスト情報と照合しないように照合パターンを個別化することができる．

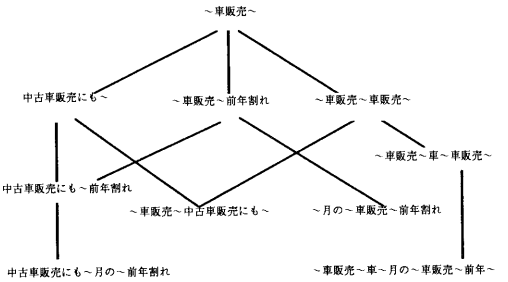

図2.1 適合パターンの包含関係

2.4 不適合情報による適合パターンの個別化

不適合情報を用いて適合パターンを個別化するのに，次の二つの方法が考えられる．

(1')部分情報の集合から不適合情報にも含まれる要素を除く．
(2')適合パターンの集合から不適合情報と照合する要素を除く．

処理(1')は適合パターンを構成する以前に行い，処理(2')はそれ以後に行う．

処理(1')を経て適合パターンを構成する場合，これまでに得た適合パターンによって適合情報を洩れなく照合すれば，その時点でさらにパターンを個別化するのを終了する．

処理(1')を経ずに構成した適合パターンの集合において，どの程度共通性の高いパターンが不適合情報と照合するのかはあらかじめ明らかでない．したがって，処理(2')を適用する際には，適合パターンは各適合情報と一対一に照合するまで個別化する．

テキスト情報(2.1)および(2.2)を適合情報として，(2.3)を不適合情報として，(1)部分情報の抽出，(2)適合パターンの構成および処理(2')を行うと，処理(2)で得た適合パターン集合(2.8)から，

(2.9) ～車販売～

が除かれ，残った適合パターンの集合から，適合情報と洩れなく照合するパターンの組合せを選び，これを論理和の形式で結合すると，

(2.10)中古車販売にも～∨～車販売～前年割れ∨
～車販売～車販売～

という照合パターンを最終的に得る．

また，同様の入力に対して，処理(1)，処理(1')および処理(2)を行うと，

(2.11) ～新車販売～
～中古車販売にも～
中古車販売にも～
新車販売～中古車販売にも～

という適合パターンを構成し，この適合パターンの集合から，適合情報と洩れなく照合するパターンの組合せを選び，これを論理和の形式で結合すると，

(2.12) ～新車販売～∨～中古車販売にも～

という照合パターンを最終的に得る．

3 実験評価

3.1 実験手法

実験手法として，前節で述べた処理(1)(1')(2)(3)を採用する．つまり，適合パターンを構成する以前に，不適合情報にも含まれる部分情報を除いておく．また，適合パターンを構成する際，部分情報の長さと頻度との積が等しい場合，頻度の高いものを優先して適合パターンを構成する．

3.2 評価データ

照合対象は，新聞記事の標題およびその第一文それぞれ1038件を用いる．正解データを作成するため，「電気通信の新しい技術や手段の開発」や「リサイクルおよび再生利用に関する情報が欲しい」などの8つの質問に対して，照合対象から解答となるものを被験者4名に選ばせた．ただし，標題と第一文とはそれぞれ別に被験者に提示した．

ここで，4名全員が正解とみなした標題または第一文の集合を正解データとする．

正解データの中で，同一記事において標題および第一文のどちらも4名全員が正解としたものを適合情報とする．また，同一記事において標題および第一文のどちらも3名が正解とみなしたものを不適合情報とする．

これにより，各質問に対応して，正解データ，適合情報および不適合情報からなる評価データを8組作成した．評価データを"ele_com"や"recycle"などと名付けた．

3.3 実験方法

第3.2節で述べた評価データ8組に対し，次のように第3.1節で述べた実験手法を繰り返し適用する．

(1)適合情報および不適合情報から実験手法により照合パターンを生成する．
(2)生成した照合パターンを用い，照合対象に対して照合を行う．
(3)照合結果と正解データを比較し，適合および不適合情報を作成する．

照合結果が以前の結果と等しくなった場合に，この繰り返しを停止する．また，適合および不適合情報は，照合結果に対する正解データから作成し，以前の適合および不適合情報は，蓄積しないものとする．

3.4 評価結果

評価結果を表3.1に示す．ただし，

$$再現率=\frac{照合結果数-照合誤り数}{正解データ数}$$

$$適合率=\frac{照合結果数-照合誤り数}{照合結果数}$$

である．評価データ"recession"と"accident"は，照合パターンを生成しなかった時点で終了した．評価データ"home_ele"は，2回目の繰り返しにおける照合結果と3回目の繰り返しにおける照合結果とが等しくなり，この時点で終了した．それ以外は，前々回の照合結果と等しくなった時点で終了した．

適合率が10%以下となる回が1/3を占めるが，適合率をそれほど下げずに再現率が向上する回も同程度存在し，再現率は全体的に向上する．

3.5 考察

適合率が10%以下になった回は，30回行った生成・照合の繰り返しの中で1/3を占める．例えば，評価データ"recycle"の2回目から3回目の繰り返しにおいて，次の照合パターンが生成されている．

(3.1) ～ている～∨スチール缶リサイクル～∨缶プレスカーでリサイクル～∨～再生利用～∨～き～

適合パターン”～ている～”や”～き～”が照合対象中に多く現れることは明らかである．実際，”～ている～”で189件，”～き～”で239件のテキスト情報と照合する．

適合率が10%以下となった回は，通常このような共通性の高い適合パターンを含む．このような共通性の高い適合パターンによる照合パターンの生成を避けるためには，部分情報の順位付けを頻度優先でなく長さ優先にする方法が考えられる．また，ユーザとのやりとりを考慮すれば，照合結果があまりに多い場合には，ユーザに対して絞り込みを行うかどうかを問い合わせることができる．

評価データ		繰り返し						
		0	1	2	3	4	5	6
medical	再現率	8/14	12/14	11/14	12/14	—	—	—
	適合率	8/14	12/48	11/113	12/48			
recession	再現率	12/42	12/42	20/42	19/42	29/42	42/42	—
	適合率	12/24	12/23	20/174	19/41	29/225	42/2076	
ele_com	再現率	6/13	8/13	6/13	8/13	—	—	—
	適合率	6/10	8/100	6/18	8/100			
asia	再現率	6/22	8/22	9/22	9/22	9/22	—	—
	適合率	6/10	8/22	9/35	9/25	9/35		
nhk	再現率	6/7	7/7	7/7	7/7	—	—	—
	適合率	6/8	7/114	7/13	7/114			
accident	再現率	12/28	19/28	17/28	28/28	—	—	—
	適合率	12/14	19/836	17/21	28/2076			
home_ele	再現率	4/10	4/10	6/10	6/10	—	—	—
	適合率	4/12	4/9	6/16	6/16			
recycle	再現率	12/17	13/17	15/17	15/17	17/17	17/17	17/17
	適合率	12/18	13/95	15/18	15/403	17/27	17/722	17/27

表3.1 評価結果

前回よりも再現率が下がった回は，評価データ”medical”，”ele_com”および”accident”の1回目から2回目の繰り返しにそれぞれ存在する．この原因は，適合情報から抽出した部分情報の多くが不適合情報に含まれていたために，適合情報と洩れなく照合する適合パターンを構成するのに十分な数の部分情報が得られなかったためである．

例えば，評価データ”ele_com”の1回目から2回目の繰り返しにおいて，次の2件の適合情報を洩らしている．

(3.2)【米・モトローラ社とデジタル通信で提携　NEC】
(3.3)【仕組み　通信のデジタル化：上（なんでもQ&A）】

このとき，評価データ”ele_com”の1回目および2回目の繰り返しで生成した照合パターンは，それぞれ，

(3.4)～電話～∨～デジタル～∨～ダイヤル～
(3.5)～に変わってきている～に変わってきている～∨シャープ～∨～通信網～通信網～∨移動電話用～三井物産と～∨三井物産と～移動電話用～

となっており，部分情報”デジタル”が，不適合情報に含まれ，これを用いて適合パターンを構成できなかったために，適合情報(6.3)および(6.4)の照合洩れが起きる．

これを解決するためには，不適合情報を利用する際に，部分情報ではなく，不適合情報と照合する適合パターンを除く方法が考えられる．

4 今後の課題

前節の考察で述べたように，まず，部分情報の順位付けや不適合情報の利用方法に関して，それが精度に与える影響を明らかにする必要がある．

さらに，今回，適合情報および不適合情報のみから照合パターンを生成する手法を提案したが，照合の繰り返しを想定すると，前回の照合に用いたパターンや適合情報および不適合情報の履歴など利用できる情報は多くあり，これらを利用すれば精度を向上させることが可能であろう．

5 おわりに

本稿では，適合情報と不適合情報から照合パターンを生成する手法を提案した．また，この手法を用いた実験評価により，この手法が再現率を改善することを確認した．さらに，手法の検討および実験評価の過程で，部分情報の順位付けや不適合情報の利用方法が精度に影響を及ぼすことがわかり，今後の指針を得ることができた．

参考文献

[1] 日本電子工業振興協会: 自然言語処理技術に関する調査報告書, pp.110-140, (1995).

[2] Maes, P.: “Agents that Reduce Work and Information Overload“, Communications of the ACM, vol. 37, No7, pp.31-42, 145-146, (1994).

[3] Salton, G. and Buckley C., “Improving Retrieval Performance by Relevance Feedback“, Journal of the American Society for Information Science, 24, pp.288-297, (1990).