LL_005

専門検索エンジンの高速半自動生成法

Rapid Semi-automatic Synthesis of Domain-Specific Web Search Engines

宮川 礼子 † 鈴木 悠生 † 鍋島 英知 † 岩沼 宏治 †
Reiko Miyagawa Yuki Suzuki Hidetomo Nabeshima Koji Iwanuma

1 はじめに

本論文では，汎用検索エンジンのパーソナライゼーションを目的として，非常に少ない手間で専門検索エンジンを構築する手法を提案する．評価実験の結果，実用化に向けて十分な結果が得られたので報告する．

インターネットにおける情報検索手段として検索エンジンが幅広く利用されているが，その検索結果の質はまだ満足できるものではなく，現在も改善のためのたゆまぬ努力が行われている．そうした検索結果の品質向上のための手段の1つが，汎用検索エンジンのパーソナライゼーションまたは専門検索エンジンの提供である．それぞれ，検索範囲を個人の趣味・嗜好に限定することで，または特定の分野に限定することで検索結果の質の向上を図ることを目的としている．この目的の観点からは両者を同一視できるので，本稿ではこれらを総称して専門検索エンジンと呼ぶこととする．

本研究では，専門検索エンジン構築手法の1つである小山らによる検索隠し味を用いた専門検索エンジンの構築手法 [5, 3] に着目する．検索隠し味とは，あるドメインに属する Web ページ群を特定するためのキーワードのブール式である．理想的には，検索隠し味を汎用検索エンジンに入力したとき，検索結果として対象のドメインに関する Web ページのみが漏れなく獲得できることが望ましい．ユーザ質問 q に検索隠し味 s を付加し $(q \wedge s)$，汎用の検索エンジンに与えることで，ドメインに属する Web ページ群からユーザ質問に適合する Web ページ群のみ抽出することが可能となる．

検索隠し味モデルは，ユーザ質問に検索隠し味を付加して汎用の検索エンジンに送るだけで高い適合率と再現率を示す非常に優れた専門検索エンジンの構築手法である．しかし検索隠し味を抽出するためには，人手により2,000件もの Web ページを対象ドメインに属するページかどうか（つまり正例と負例とに）分類する必要があり，訓練集合を作成するために非常に手間と労力を要する．

本研究では，ユーザが所望するドメインに関する専門検索エンジンを少ない手間で，かつ短時間で容易に構築することを目的として，(i) ディレクトリ型検索エンジンを利用した精錬による訓練集合の半自動生成法 [6] と (ii) 類似度に基づく訓練集合の半自動生成法を提案する．

前者は，ディレクトリ型検索エンジンから収集した Web ページを，人手により選別した少数の訓練事例より生成した決定木を用いて精錬することにより高精度の訓練集合を生成する手法である．後者は，ユーザの所望するドメインに関連するディレクトリが存在しない場合に，汎用検索エンジンを用いて人手により収集した極少数の正例ページを基にして機械的に収集した多量のページを単語頻度に基づく尺度によって分類し，高精度な訓練集合を生成する手法である．両手法ともに短時間で十分な精度の訓練集合を生成できることが大きな特徴である．また，両手法により専門検索エンジンを構築し評価実験を行った結果は，人手により訓練集合を生成した場合と同程度の適合率と再現率を示している．これは我々の手法が汎用検索エンジンのパーソナライゼーション手法として非常に有望であることを示している．

本論文の構成を次に示す．まず2章で本研究の基礎となる小山らの検索隠し味抽出方法を紹介する．3章および4章において本稿で提案する訓練集合の半自動生成法を述べる．5章は従来手法との比較評価実験である．最後に関連研究を紹介し，本研究をまとめる．

2 検索隠し味の手動抽出法

小山らの料理レシピドメインにおける検索隠し味の抽出を例として，検索隠し味の抽出方法と評価方法を紹介する．なお本稿では小山らの手法を手動抽出と呼ぶ．

ドメインに属するページを収集するため，将来ユーザが入力すると予想されるキーワードを選ぶ．小山らは料理レシピドメインにおいて，食材である牛肉・鶏肉・ピーマンなど10種のキーワードを選択している．次に各キーワードを個別に汎用検索エンジンに入力し，検索結果上位200件，計2,000件の Web ページを収集する．収集した Web ページから名詞を抽出し，名詞の出現ベクトルを Web ページの属性とする．収集した Web ページを人手により正例と負例に分類し，訓練集合及び検証集合（それぞれ1,000件の Web ページ）を作成する．

そして訓練集合に対して，ID3 [4] で使用されている情報量に基づく決定木学習アルゴリズムを適用して決定木を生成する．図1に料理レシピドメインにおける単純な決定木の例を示す．節はキーワードであり，そのキーワードが Web ページに含まれるならばラベル "1" の枝を進み，含まれない場合は "0" の枝を進む．各葉はクラスを表す．Web ページがドメインに属する場合はクラス T であり，そうでない場合は F である．

次に決定木を検索エンジンに入力できるブール式に変換する．決定木の根からクラス T の葉へのパスを連言肢とし，各連言肢の選言を取りブール式に変換する．図1の決定木では以下の選言標準形のルールが生成される：

$$\text{大さじ} \vee (\neg \text{大さじ} \wedge \text{作り方} \wedge \neg \text{仮定} \wedge \neg \text{トップ}) \vee (\neg \text{大さじ} \wedge \neg \text{作り方} \wedge \text{こしょう} \wedge \neg \text{鍋})$$

一般的にこのブール式は複雑となり，汎用検索エンジンに入力することができない．そこで次に Rule post-pruning に基づく単純化を行う．単純化では，検証集合に対するルールの適合率と再現率の調和平均が改善する限り，リテラルまたは連言肢の除去を行う．小山らは，訓練集合から生成した決定木をルールに変換し，単純化を行った結果，検索隠し味 "(材料 $\wedge \neg$ 専門 $\wedge \neg$ 商品) $\vee$ 大さじ" を抽出している．

検索隠し味の評価では，訓練集合を生成する際に使用

† 山梨大学, University of Yamanashi

した語とは異なる新しい検索語として，豚肉・ほうれん草・エビの3つのキーワードを選ぶ．各キーワードに検索隠し味を付加して汎用検索エンジン goo に入力し，検索結果の上位 1,000 件に含まれるレシピページの割合（適合率）を調べた結果，97% 以上という高い値を示した．次に再現率の評価では，インターネット上に存在するレシピページの総数を知ることは困難であるため，検索エンジンが返すヒット数と適合率に基づいた推定再現率 [3] を利用する．先の各キーワードについて推定再現率を調べた結果，86% 以上という高い値を示した．

検索隠し味による専門検索エンジンの構築は，高い適合率と再現率を示す優れた手法であるが，訓練集合を人手で作成するため非常に手間と労力を要する．これを改善するために，続く2つの章において十分な精度の訓練集合を半自動的に生成する我々の提案手法を示す．

3 精錬による訓練集合の半自動生成

精錬による訓練集合の半自動生成手法 [6] では，まずディレクトリ型検索エンジンからユーザの所望するドメインに関係するディレクトリ P_{dir} を選択し，P_{dir} に登録されている Web ページを正例の候補として収集する．ただしユーザの意図するドメインと P_{dir} とが完全に一致することは，ディレクトリ管理者の編集方針を把握することが困難であることからも殆ど無いといえる．従って P_{dir} はユーザの意図とは異なる Web ページ群をノイズとして多く含むことになる．また，ディレクトリ型検索エンジンに登録されている URL のほとんどは Web サイトのトップページであるため，コンテンツの紹介文が多く，具体的で有用な情報に欠ける場合が多い．そこで我々は P_{dir} に登録されている URL からさらにリンクを1つ辿ったページまでを収集する（ただし，その Web サイト外へのリンクは無視する）．P_{dir} から得られた Web ページの集合を初期正例集合 P_{init} とする．同様にして，P_{dir} の兄弟ディレクトリに含まれる Web ページを初期負例集合 N_{init} として収集する．兄弟ディレクトリに注目する理由は，正例と負例とが類似している方が，ドメインに属するか否かをより厳密に判定可能な分類能力の高い決定木が得られると考えたためである．

そして初期正例集合 P_{init} から 50 件をランダムに選び，それらを人手により正例と負例とに分類し，初期訓練集合とする．これを決定木学習アルゴリズムに与え，精錬用決定木を作成する．訓練例に含まれる属性数は少なく，ここで生成される決定木のサイズは小さいため，枝狩りによる精錬用決定木の単純化は行わない．

次に初期正例集合 P_{init} と初期負例集合 N_{init} を精錬用決定木により分類する（図 2）．精錬用決定木は，初期正例集合の部分集合（初期訓練集合）から生成されており，ディレクト管理者の意図とユーザの意図との違いを識別するモデルであるといえる．従って，P_{init} 中の Web ページが，精錬用決定木により再び正例として分類されるならば，その Web ページは対象ドメインに関係する可能性が高い．このようにして精錬された精錬集合を $P_{refined}(\subseteq P_{init})$ とする．同様にして初期負例集合 N_{init} から精錬用決定木により負例として分類されたページを抽出し，これを負例集合 $N_{refined}(\subseteq N_{init})$ とする．負例集合を初期負例集合から精錬用決定木によって生成することで，ユーザの意図と適合しない Web ページ群のみを収集することが可能となる．そして精錬により得られた正例・負例の集合から改めて検索隠し味を抽出する．これ以降の手順は従来手法と同様である．

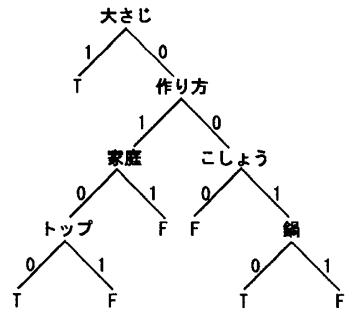

図 1 決定木の例 [5]

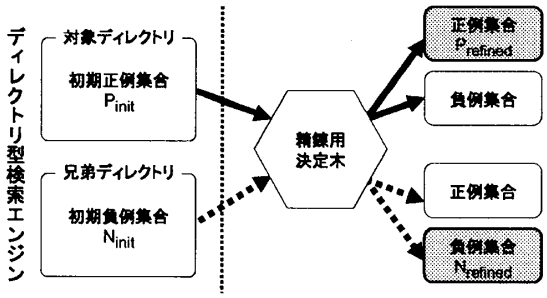

図 2 訓練例の精錬

4 類似度による訓練集合の半自動生成

本章では，ディレクトリ型検索エンジンに依存しない，訓練集合の半自動生成手法を提案する．

手動抽出の場合と同様に，所望のドメインに関するキーワードを 10 個選ぶ．次に各キーワードを汎用検索エンジンに入力し，その検索結果上位 200 件，計 2,000 件の Web ページを機械的に収集する（これを W とする）．次にユーザは，各キーワードの検索結果から所望のドメインに適合する文書を 1 件選ぶ．これを代表正例と呼ぶ．ユーザは全ての検索結果を閲覧する必要はなく，検索結果の上位から見ていきドメインに適合する文書が見つかればそれを選択すればよい．各キーワード毎に代表正例を選び，計 10 件抽出する（これを R とする）．

W は，所望のドメインに適合しない文書を数多く含んでいる．そこで，代表正例と類似した文書を W から抽出することにより精度の高い訓練集合を生成する．2つの Web ページ p,q 間の類似度を，TF 法による単語頻度を属性とする文書ベクトル $\vec{p},\vec{q}$ 間のコサイン尺度として定義する．名詞の種類数を m とすると，文書ベクトル $\vec{p},\vec{q}$ は次式で表される：

$$\vec{p} = [p_1 p_2 \cdots p_m], \quad \vec{q} = [q_1 q_2 \cdots q_m]$$

ここで p_i, q_i は，それぞれ文書 p または q における単語 i の出現頻度である．文書 p,q 間の類似度を以下のコサイン尺度により与える：

$$\cos(p,q) = \frac{\sum_{i=1}^{m} p_i q_i}{\sqrt{\sum_{i=1}^{m} p_i^2}\sqrt{\sum_{i=1}^{m} q_i^2}}$$

各 Web ページ $w \in W$ に対し，各代表正例 $r \in R$ との類似度 $\cos(w,r)$ を算出し，もし閾値以上の値ならば，そのページを正例として分類する．すべての代表正例との類似度が閾値未満ならば負例として分類する．

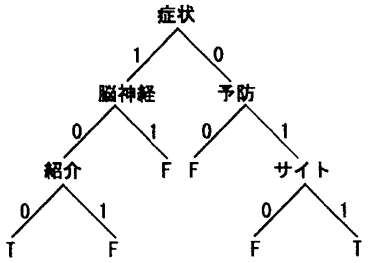

図3　精錬用決定木

以上の手続きより得られた正例と負例の集合から検索隠し味を抽出する．以降の手順は従来手法と同様である．

5　評価実験

まず精錬による半自動生成法の評価実験結果を示す．対象ドメインは“病気や怪我の詳細と治療法”とし，Yahoo! からドメインに関するディレクトリとして“健康と医学 > 病気・症状”を選択した．このディレクトリから Web ページを 3,000 件収集し初期正例集合 P_{init} とした．また兄弟カテゴリである“看護”や“職場”等から 7,000 件の Web ページを収集し初期負例集合 N_{init} とした．次に P_{init} から 50 件をランダムに選び，それらを人手により正例と負例とに分類して決定木学習アルゴリズムに与え，精錬用決定木を生成した（図 3）．次に P_{init} と N_{init} を精錬用決定木により精錬し，P_{init} のうち正例として分類された 824 件の Web ページを正例集合 $P_{refined}$ とし，N_{init} のうち負例として分類された 6,487 件の Web ページを負例集合 $N_{refined}$ とした．

手動生成では，まず将来ユーザが入力すると予想されるキーワードとして，“歯，鼻，膝，肩，腰，胸部，目，心臓，血，頭部”の 10 個を選択した．各キーワードを Google に入力し，検索結果から上位 200 件，計 2,000 件の Web ページを収集し，これを人手により正例集合（306 件）と負例（1694 件）に分類した．

精錬アルゴリズムの有用性を示すため，精錬を行わずに初期正例集合と初期負例集合から検索隠し味を抽出する手法についても評価を行った．この手法は人手を要さないので自動生成法と呼ぶ．自動生成では，半自動生成において収集した初期正例集合 (3,000 件) と初期負例集合 (7,000 件) から検索隠し味を抽出した．

各手法より生成された検索隠し味を表 1 に示す．表中の“精錬半自”は精錬による半自動生成法を示す．次に各検索隠し味の評価を行った．評価用の検索キーワードとして“足首・頭・肺”を選び，Google にキーワードだけを入力した場合と，検索隠し味を付加した場合の適合率と推定再現率を算出した（表 2 および表 3）．

表 2 より，キーワードのみの場合の適合率は非常に低いが，精錬による半自動生成は手動生成と同程度の適合率を示した．自動生成はキーワードのみの場合と大差なく，これは精錬手法の有効性を示している．表 3 より，推定再現率においても精錬による半自動生成は手動生成と同程度の良い値を示している．自動生成も高い推定再現率を示しているが，これは他の手法と比べて検索隠し味による Web ページの絞り込みが甘く，ヒット数が何倍もあるため，結果として高い再現率となった．

次に提案手法の安定性を調べるため，人手により分類した精製用の訓練集合として 50 件の Web ページを 3 セット (A,B,C)，100 件の Web ページを 2 セット (D,E) 用意して適合率を調べた．それぞれの検索隠し味と適合率を表 4, 5 に示す．A,B,C においてはばらつきのある結果となっているが，訓練例を 100 件とした場合は比較的安定した性能が得られている．

次に，類似度による半自動生成法の評価実験結果を示す．対象ドメインは“星の伝説，伝承”とした．ユーザが将来入力するであろうキーワードとして“アンタレス・

表1　抽出された検索隠し味

	検索隠し味
手動生成	(症状 ∧ ¬注文 ∧ ¬チーム ∧ ¬評価 ∧ ¬サイズ ∧ ¬デザイン) ∨ 炎症 ∨ 頭痛
精錬半自	(症状 ∧ ¬紹介 ∧ ¬脳神経) ∨ 看病
自動生成	¬パン ∧ ¬運営 ∧ ¬血清 ∧ ¬医薬品 ∧ ¬その他

表2　各手法における適合率（検索結果上位 100 件）

	キーワードのみ	手動生成	精錬半自	自動生成
足首	0.08	0.69	0.60	0.10
頭	0.04	0.75	0.66	0.04
肺	0.22	0.73	0.70	0.26

表3　各手法における推定再現率（検索結果上位 100 件）

	手動生成	精錬半自	自動生成
足首	0.78	0.62	0.75
頭	0.79	0.70	0.79
肺	0.74	0.68	0.73

表4　異なる訓練例における検索隠し味

	検索隠し味
A (50 件)	原因 ∨ (状態 ∧ ¬闘病)
B (50 件)	症状 ∧ ¬脳外科 ∧ ¬進歩 ∧ 児
C (50 件)	(原因 ∧ ¬予定) ∨ 食事
D (100 件)	症状 ∧ ¬検索 ∧ ¬等
E (100 件)	(症状 ∧ ¬ホームページ) ∨ 発症

表5　異なる訓練例における適合率（検索結果上位 100 件）

	足首	頭	肺
A (50 件)	0.47	0.45	0.68
B (50 件)	0.63	0.64	0.60
C (50 件)	0.48	0.51	0.59
D (100 件)	0.63	0.71	0.69
E (100 件)	0.61	0.63	0.63

表6　類似度に基づく分類より得られた訓練集合の適合率

		適合率
分類前		0.13
代表正例 5 件	正例	0.69
	負例	0.90
代表正例 10 件	正例	0.71
	負例	0.96
代表正例 20 件	正例	0.60
	負例	0.94

表7　抽出された検索隠し味

	検索隠し味
手動生成	神話 ∨ 退治 ∨ 大神
類似半自 5	アルゴル ∨ アルゴ
類似半自 10	エチオピア ∨ 物語 ∨ アルゴ
類似半自 20	(南中 ∧ ¬ヒドラ) ∨ (ペガサス ∧ ¬アルゴル)
精錬半自	ゼウス ∨ (神話 ∧ ¬データ ∧ ¬神前 ∧ ¬国際)

表8　各手法での適合率 (検索結果上位 100 件)

	キーワードのみ	手動生成	類似半自 5	類似半自 10	類似半自 20	精錬半自
オリオン	0.05	0.90	0.38	0.60	0.32	0.69
カシオペア	0.03	0.72	0.34	0.59	0.15	0.58
射手座	0.04	0.53	0.19	0.30	0.12	0.36

表9　各手法での推定再現率 (検索結果上位 100 件)

	手動生成	類似半自 5	類似半自 10	類似半自 20	精錬半自
オリオン	0.94	0.03	0.97	0.30	0.16
カシオペア	0.97	0.01	0.76	0.00	0.02
射手座	0.48	0.00	0.68	0.00	0.09

アンドロメダ・エリダヌス・ケンタウルス・ペルセウス・牡牛座・大熊座・乙女座・琴座・白鳥座" の 10 個を選び，それぞれ Google に入力し 1 キーワードあたり 200 件，計 2,000 件の Web ページを機械的に収集した．次に，キーワード毎に代表正例を選び，2,000 件の Web ページを類似度に基づき分類した（類似度の閾値は 0.65 とした）．代表正例数がそれぞれ 5,10,20 件の場合の分類後の適合率を表 6 に示す．表中の "分類前" は人手により 2,000 件を分類した場合の適合率を示す．表 6 より，類似度に基づく分類によって適合率が大きく向上していることがわかる．代表正例 20 件の場合に適合率が少し低下しているが，これは代表正例数の増加に伴い正例として間違って分類される傾向が強くなるためである．

また比較のため，このドメインに対し精錬による半自動生成法も適用した．ただし Yahoo! には一致するディレクトリが存在しなかったため，関連の深い "生活と文化 > 神話・民話と民俗学" を選択した．

表 7 にそれぞれの手法において抽出された検索隠し味を示す．表中の "類似半自 5,10,20" はそれぞれ代表正例 5,10,20 件の場合の結果である．検索隠し味の評価用のキーワードとして" オリオン・カシオペア・射手座" を選び，Google にキーワードだけを入力した場合と検索隠し味を付加した場合の適合率および推定再現率を算出した．その結果をそれぞれ表 8，表 9 に示す．

類似度による半自動生成法では，手動生成には及ばないが，キーワードのみ場合と比べて高い適合率を示している．訓練集合の精度が最も高かった代表生成 10 件の場合においては，適合率・推定再現率ともに高く，検索結果の質を大きく改善していることが分かる．代表正例の数によって推定再現率にばらつきがあるが，そもそも検索エンジンがインターネット上の全文書を網羅することは困難であるので正確な再現率を求めることは無理がある．キーワードのみでは 100 件中数件であった適合文書の数が，検索隠し味を付加することによって 50 件以上の適合文書が得られることは，ユーザにとって大変有用であり，高速に計算可能な本手法は実用化へ向けての大きな可能性を持っているといえる．

6　関連研究

訓練集合の効率的な生成は，専門検索エンジンを手軽に構築する目的において非常に重要である．Nigam らは，訓練集合作成の手間を削減するため，少数の分類済み文書と多数の未分類文書を使って，分類済みの文書を増やす手法を提案している [2]．Liu らは，正例集合と未分類の文書を使って，未分類文書中の正例を特定する手法を提案している [1]．これらの手法では，我々の手法と比較して比較的多数の分類済み文書を仮定している点と，高精度な分類を目的としているため計算時間が非常にかかる点が大きく異なる．我々の訓練集合生成法は TF ベクトルの類似性に基づく比較的単純な手法ではあるが，検索結果の質を向上するために十分な精度を持っている．本研究ではユーザが所望のドメインに関する専門検索エンジンを手軽に構築できることを目的としているので分類済み文書の数は少ない方が望ましいが，分類精度の安定性を向上させるためにも，これらの手法の併用を検討することは今後の課題の 1 つである．

7　おわりに

本研究では，汎用検索エンジンのパーソナライゼーションを目的として，非常に少ない手間で専門検索エンジンを構築するため，(i) ディレクトリ型検索エンジンを利用した精錬による訓練集合の半自動生成法と (ii) 類似度に基づく訓練集合の半自動生成法を提案した．これらの手法により生成された検索隠し味を利用することで，ユーザの所望するドメインに関する検索エンジンを高速に構築することが可能である．評価実験の結果は検索結果の質を大きく改善できることを示しており，本手法は実用化へ向けた大きな可能性を持っているといえる．

謝辞：本研究は一部，文科省科学研究費補助金 (No.16500078) ならびに中部電力基礎技術研究所研究助成の援助を受けている．

参考文献

[1] B. Liu, W. S. Lee, P. S. Yu, and X. Li. Partially supervised classification of text documents. In *Proceedings of ICML-2002*, pages 387–394, 2000.

[2] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39:103–134, 2000.

[3] S. Oyama, T. Kokubo, and T. Ishida. Domain specific search with keyword spices. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):17–27, 2004.

[4] J. R. Quinlan. Induction of decision trees. In Jude W. Shavlik and Thomas G. Dietterich, editors, *Readings in Machine Learning*. Morgan Kaufmann, 1990. Originally published in *Machine Learning* 1:81–106, 1986.

[5] 小久保 卓, 小山 聡, 山田 晃弘, 北村 泰彦, and 石田 亨. 検索隠し味を用いた専門検索エンジンの構築. 情報処理学会論文誌, 43(6):1804–1813, 2002.

[6] 鈴木悠生, 鍋島英知, and 岩沼宏治. 精錬手法に基づく検索隠し味型専門検索エンジンの半自動生成. In 第 4 回情報科学技術フォーラム, *L-83*, pages 190–202, 2005.