L-083

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構築

Semi-automatic Generation of Domain-specific Web Search Engine Based on Refining Training Data

鈴木 悠生 † **鍋島 英知** † **岩沼 宏治** †
Yuki Suzuki Hidetomo Nabeshima Koji Iwanuma

1 はじめに

本論文では，ディレクトリ型検索エンジンから収集した Web ページを精錬することで精度の高い訓練集合を生成し，検索隠し味による専門検索エンジンを半自動構築する手法を提案する．評価実験の結果，手動生成した検索隠し味と同程度の性能が得られたので報告する．

インターネットに存在する無数の Web ページの中からユーザの望む情報を探し出すため，検索エンジンが幅広く利用されている．目的の情報を含む Web ページを効率良く絞り込むためには，適切なキーワードからなる詳細な検索式を検索エンジンに与える必要があるが，そのような詳細な検索式を初心者が作成することは容易ではない．こうした WWW 情報検索における問題の解決法の 1 つとして専門検索エンジンの提供がある [3].

本研究では，専門検索エンジン構築のための手法の 1 つである小久保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, 10] 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．

検索隠し味モデルの利点の 1 つは，汎用の検索エンジンへの入力を修正するだけであるため，隠し味を利用しない場合と比べて大差ない速度で検索できることにある [5]．これに対し汎用検索エンジンの出力をフィルタリングするモデル [9] では，検索結果に含まれる Web ページを分類のために取得するため原理的に速度は遅くなり，使いやすさが損なわれることになる．

検索隠し味は，機械学習の一種である決定木学習アルゴリズムに，対象となるドメインの Web ページ群と非対象ドメインの Web ページ群とを与えることで抽出される．小久保らは，料理レシピを対象ドメインとして検索隠し味を抽出し，適合率 97% 以上，再現率 86% 以上という非常に高い性能を持つことを示した [10]．料理レシピだけでなく，レストランドメイン・中古車ドメインにおいても検索隠し味モデルが優れた性能を持つことを示している [5].

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入力す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（正例）とそうでないページ（負例）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．そこで，ディレクトリ型検索エンジンを利用して，あるカテゴリに登録されている Web ページ群を正例，別のカテゴリの Web ページ群を負例として訓練集合を自動的に生成する試みが行われている [5]．しかし，自動収集した訓練集合にはノイズが多く含まれるため，良い学習結果を得ることは難しい．我々も自動収集による検索隠し味を抽出したが，キーワード単体で検索した場合と比較して，ほとんど性能の向上を得ることはできなかった．

本研究では，精度の高い訓練集合を少ない労力で作成することを目的として，**精錬による訓練集合の半自動生成法**を提案する．我々の手法では，まず人手により 50 件の Web ページを正例と負例とに分類し，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．次にディレクトリ型検索エンジンから，対象ドメインに関するディレクトリとそうではないディレクトリを選び，そこに登録されている Web ページ群を機械的に収集する．それらを精錬用決定木により精錬し，精度の高い正例集合と負例集合を生成する．そしてこの正例と負例の集合から改めて検索隠し味を抽出する．

提案手法を評価するため，病気や怪我の詳細と治療法を対象ドメインとして実験を行った．その結果，我々の手法は，訓練集合を人手により作成した場合と同程度の適合率と再現率を示した．我々の手法で人手を要するのは，50 件の Web のページを正例・負例とに分類する作業と，ディレクトリ型検索エンジンから対象ドメインに関するディレクトリを選択する作業のみである．本手法は，従来手法と比較して非常に少ない手間で同程度の性能を持つ検索隠し味を抽出することが可能である．

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久保らの検索隠し味抽出方法を紹介し，3 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．4 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

従来手法と提案手法との違いは訓練集合の作成方法のみであるので，まず本章では小久保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について紹介する．なお，本稿では小久保らの手法を**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．小久保らは料理レシピドメインにおいて，食材である牛肉・鶏肉・

† 山梨大学, University of Yamanashi

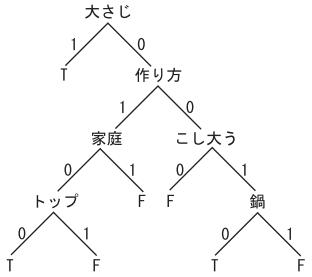

図 1　決定木の例 [10]

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汎用検索エンジンに入力し，その検索結果から Web ページを収集する．料理レシピドメインでは，各キーワードの検索結果から上位 200 件，計 2,000 件の Web ページを収集している．収集した Web ページからすべての名詞を抽出し，これをキーワードの集合とする．Web ページの属性は，各キーワードの出現ベクトルにより与えられる．収集した Web ページを人手により正例と負例とに分類し，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 [7] 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．この時点では決定木の枝狩りは行わない．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” のついた枝を進み，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：

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

最後に Rule post-pruning [8] に基づくルールの単純化を行う．単純化の指標として，検証集合に対する連言肢とルールの適合率と再現率の調和平均を用いる．ルールの単純化は以下の 2 つのステップからなる．

1. **リテラルの除去**：ルールに含まれる各リテラル L に対し，もし L を削除することで，連言肢の調和平均の値が悪化しない（小さくならない）ならば，L を削除する．
2. **連言肢の除去**：ルールに含まれる各連言肢 C に対し，もし C を削除することで，ルールの調和平均の値が悪化しないならば，C を削除する．

小久保らは，訓練集合から生成した決定木をルールに変換し，単純化アルゴリズムを適用した結果，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．

次に検索隠し味の性能を評価するため，小久保らは豚肉・ほうれん草・エビの 3 つのキーワードを検索した結果を評価している．これらは訓練集合を生成する際に使用した語とは異なる新しい検索語である．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．次に再現率を評価するため，インターネット上に存在するレシピページの総数を次式で推定する．

$U_{index} \cong$ (キーワード単体でのヒット数)
$\times$(キーワード単体で検索したときの 1,000 件の適合率)

同様にして検索隠し味を付加した場合に検索されるレシピページの総数を次式で推定する．

$U_{spice} \cong$ (検索隠し味を付加したときのヒット数)
$\times$(検索隠し味を付加したときの 1,000 件の適合率)

このとき，検索隠し味を付加した場合の**推定再現率** R は次式で与えられる．

$$R \cong U_{spice} / U_{index}$$

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．次章では，精錬により精度の良い訓練集合を半自動で生成する新しい手法を提案する．

3　精錬による訓練集合の半自動生成

我々の手法では，まずディレクトリ型検索エンジンから対象ドメインに関するディレクトリ 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り，サイトのトップページはコンテンツの紹介を主に含み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[*1]．この作業によって，ノイズとなる Web ページ群をさらに収集する可能性があるが，後述する精錬手続きにより，そのようなユーザの意図とは異なる Web ページを除去することを試みる．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．ここで他のディレクトリではなく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

[*1] ただし，その Web サイト外へのリンクは無視する．

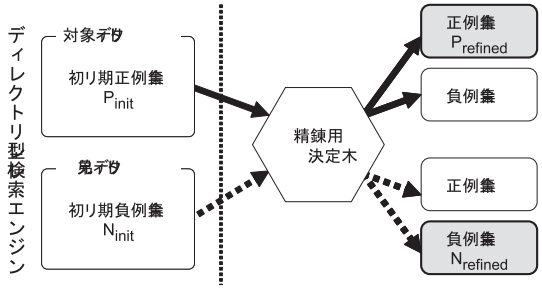

図 2　訓練例の精錬

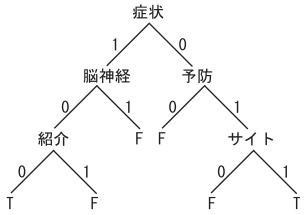

図 3　精錬用決定木

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．

そして精錬によって得られた正例・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．

4　３つの訓練集合作成法による評価実験

提案手法を評価するため，病気や怪我の詳細と治療法を対象ドメインとして小久保らの**手動生成法**と我々の**半自動生成法**との比較実験を行った．さらに精錬アルゴリズムの有用性を示すために，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し，評価した．この手法は人手を要さないので**自動生成法**と呼ぶ．

まず半自動生成の評価実験について述べる．ディレクトリ型検索エンジン 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” のサブディレクトリ “病気・症状” を選択した．このディレクトリに属する Web ページを 3,000 件収集し，初期正例集合とした．また兄弟カテゴリである “看護” や “職場” などから 7,000 件の Web ページを収集し初期負例集合とした．

次に初期正例集合 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．これを図 3 に示す．

そして P_{init} と N_{init} を精錬用決定木により精錬する．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 $N_{refined}$ とする．

表 1　抽出された検索隠し味

	検索隠し味
半自動生成	(症状 ∧ ¬紹介 ∧ ¬脳神経) ∨ 看病
手動生成	(症状 ∧ ¬注文 ∧ ¬チーム∧ ¬評価 ∧ ¬サイズ ∧ ¬デザイン) ∨炎症 ∨ 頭痛
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）

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

次にこれらを訓練集合と検証集合に分割し，従来手法と同様の手順で検索隠し味を抽出する．その結果得られた検索隠し味を表 1 に示す．

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．各キーワードを汎用検索エンジン 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．次にこれらを訓練集合と検証集合に分割し，表 1 に示す検索隠し味を抽出した．

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) を訓練集合と検証集合に分割し，これらをそのまま利用して検索隠し味を抽出した．自動生成による検索隠し味を表 1 に示す．

半自動生成，手動生成，自動生成により抽出された検索隠し味を評価するため，汎用検索エンジン Google にユーザ質問としてキーワード q だけを入力した場合と，検索隠し味 s を付加した場合（$q \wedge s$）とでの適合率と再現率を求め評価を行う．ユーザ質問として，足首・頭・肺の 3 種類のキーワードを使用した．

表 2 に検索結果上位 100 件における各手法の適合率を示す．検索語としてキーワードのみのを使用した場合の適合率の値は低く，最大でも 22% であった．それに対し，手動生成により抽出された検索隠し味をキーワードに付加して検索した場合は 69% 以上という適合率を得ている．半自動生成の場合では 60% 以上となり，手動生成に近い適合率を示した．訓練集合の精製をしなかった自動生成の場合は最大でも 26% であり，キーワードのみの場合と比べてもあまり向上していない．これは訓練集合の精錬手法の有効性を示している．

表 3 は推定再現率による各手法の評価結果である．適合率の場合と同様に，半自動生成と手動生成は比較的近い値を示している．自動生成の場合も 72% 以上という高い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ている．

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例として 50 件の Web ページを 3 セッ

表 3　各手法における推定再現率

	足首	頭	肺
半自動生成	0.622	0.703	0.680
手動生成	0.783	0.794	0.736
自動生成	0.746	0.786	0.726

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

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．

5　関連研究

検索隠し味モデルと同様に，ユーザ質問に検索式を付加することにより検索の質を向上させる試みが Pahlevi らによって行われている [6]．彼らの手法では，ディレクトリ型検索エンジンの検索結果にディレクトリ情報が付随することに着目し，ユーザにカテゴリを指定させることで検索結果を自動分類し，訓練集合を生成する．

精度の高い訓練集合を機械的に精製するためには，文書の自動分類技術が深く関わってくる．Glover らは，リンク元 Web ページのアンカー付近のテキストを利用して SVM で Web ページを分類する手法を提案している [1]．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 [4]．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 [2]．我々の手法の特徴は，初期正例・負例集合を精錬することによって精度の良い訓練集合を生成することにあり，これらの手法とは異なる．

6　おわりに

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．本論文では，この訓練集合生成の手間を大幅に軽減するため，ディレクトリ型検索エンジンから収集した訓練集合を精錬する手法を提案した．評価実験の結果，我々の手法は，わずか 50 件の Web ページを人手で分類することで，従来手法と同程度の能力をもつ検索隠し味を抽出することが可能であることを示した．

安定性と性能の向上のために，人手で分類する訓練例集合はドメインにおける特徴をまんべんなく集めたものであることが望ましい．適切な訓練集合を用意することは今後の課題の 1 つである．特に，ユーザの希望するカテゴリがディレクトリ型検索エンジンに存在しない場合に訓練集合をどのようにして生成するかは重要な課題である．また今回は，訓練集合の精錬のために名詞の出現ベクトルを属性値とした決定木学習アルゴリズムを利用したが，これとは異なる属性値を用いた SVM などの他の学習指針に基づくアルゴリズムを利用することで，より優れた精錬を行える可能性がある．本手法を他ドメインにも適用して有用性をさらに実証することや，手順の完全自動化なども重要な課題である．

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] Eric J. Glover, Kostas Tsioutsiouliklis, Steve Lawrence, David M. Pennock, and Gary W. Flake. Using Web structure for classifying and describing Web pages. In *Proceedings of WWW-02*, 2002.

[2] 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.

[3] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. A machine learning approach to building domain-specific search engines. In *Proceedings of IJCAI-99*, pages 662–667, 1999.

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

[5] 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.

[6] S. M. Pahlevi and H. Kitagawa. Taxonomy-based adaptive web search methoda. In *Proceedings of 3rd IEEE International Conference on Information Technology: Coding and Computing*, pages 320–325, 2002.

[7] 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.

[8] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[9] J. Shakes, M. Langheinrich, and O. Etzioni. Dynamic reference sifting: A case study in the homepage domain. In *Proceedings of WWW-97*, 1997.

[10] 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.