WWWから獲得した知識による検索語拡張とレシピ検索タスクにおける評価

野田 雄也 † 高橋 哲朗 † 橋本 力 †† 鳥澤 健太郎 †† *
† ニフティ株式会社
†† 独立行政法人 情報通信研究機構 MASTAR プロジェクト言語基盤グループ
noda.yuya@nifty.co.jp, takahashi.tetsuro@nifty.co.jp
ch@nict.go.jp, torisawa@nict.go.jp

1 はじめに

近年 Web 上の情報量は増加の一途を辿っており適切な情報を探し出すためのツールとして検索エンジンは欠かせないものとなっている．多くの検索エンジンは利用者から与えられた検索語に応じて膨大な量の情報源から関連のあるサイトを探し出すが，利用者が求める情報を得るためには適切な検索語を選定しなければならず，そのためには検索対象に関する十分な知識が必要となる場合が多い．

たとえば今 “冷え性に効く料理の作り方” を知りたいとする．“冷え性” に関する十分な知識を持っていない場合はこの要求をそのまま検索エンジンに与えることになり，この場合記事内に “冷え性” という語を含むサイトが検索結果となる．しかし，“冷え性” に効く食材や冷え性に効く調理法に関する知識を事前に持っていれば，それらの知識を検索語として与えることで，本来の情報要求である” 冷え症に効く料理” についてのページを，たとえそのページが “冷え性” という語を含んでいなくても発見することができる．つまり，“冷え性” に “にんにく” が効くことを知っていれば，“にんにく” を使った料理を探すことができるのである．利用者が情報を求めて検索を行う場面では検索対象に対する知識が不十分なことが多々あり，このような知識の不足は上述の例のように検索エンジン側で補うことが望ましい．

Web 検索利用者に求められる知識量の増加により Web からの検索が困難になっている一方で，その Web が内包する知識量は情報量の増加に伴い一個人の知識量をはるかに上回っている．これらの知識を獲得する手法として，Stijn ら [3] [4] は，Web から特定単語クラスに属する語や，意味的関係知識を獲得する手法を提案している．我々は，これらの Web から獲得した知識を活用し，Web 検索において利用者に求められる知識の補填を行うことで利用者の負担軽減や新たな情報の発見ができるのではないかと考え，検索エンジンの検索語拡張という形で意味的関係知識の適用を試みた．

本稿では，Web から獲得した知識を活用した検索エンジン「みんなのレシピ検索 β」について紹介した後，検索語拡張が検索結果に与える効果や影響について意味的関係知識を用いた検索語拡張を組み合わせた評価実験の結果を報告する．

2 関連研究

検索語拡張の手法として Relevance Feedback を用いた手法が古くから提案されている [5]．本研究では検索語拡張に用いる知識として，検索語に依存せずに静的に求まる検索語拡張の知識の効果を検証する目的があったために直接の比較対象とはしなかったが，実アプリケーション上では Relevance Feedback も併用することにより，より高い再現率が期待できる．

人手もしくは共起関係や構文構造から自動的に獲得したシソーラスを用いて検索語拡張を行う手法も提案されている [1]．また Riezler ら [2] や海野ら [6] は自動的に獲得した言い換え表現を用いて検索語拡張を行っている．これらのクエリ拡張は “リダクション–縮小” “cat–feline” といった同一概念における検索語拡張である．

一方，本研究では食材とその健康効果のような高度な意味的関係知識 (e.g.“にんにく–冷え性”) による検

* NODA Yuya†, TAKAHASHI Tetsuro†, HASHIMOTO Chikara††, TORISAWA Kentaro††.
† Nifty Corporation
†† National Institute of Information and Communications Technology, MASTAR Project Language Infrastructure Group.

索語拡張を提案する．また検索語拡張を適用した場合，良い効果だけではなく不適切な検索結果が検索されることによる精度の低下も考えられる．実験ではこの不適切な検索結果がどれくらい含まれるかについても検証を行う．

3　みんなのレシピ検索 β

Web から獲得した豊富な知識の活用例として，料理レシピが記載されたブログ記事を検索できる専門検索サイト「みんなのレシピ検索 β」[1]を開発した (図 1)．Web からの知識獲得には [3] [4] の手法を用いた．こ

図 1: みんなのレシピ検索 β

の手法により獲得した「健康効果と食材」に関する知識に対して人手によるクリーニングを行ない，得られた 1373 対の知識を検索語拡張と検索結果への提示に利用している．知識の実例を以下に示す．

- < レーズン > は < 貧血 > に効く
- < ごま > は < 美肌 > に効く
- < にんにく > は < 冷え性 > に効く

検索対象は，2007 年 7 月から 2008 年 12 月までの日本語ブログ記事約 2 億件から，レシピが記載されている約 16 万件を抽出したものである．

検索アルゴリズムを 図 2 に示す．

検索機能は，Word 転置インデックスと Bi-gram 転置インデックスを併用したハイブリッド検索である．Word 転置インデックスはブログ記事を形態素解析し，レシピ名，食材名などのレシピ関連語と，健康効果や栄養素などの Web から獲得した知識に含まれる語を抽出して構築した．Bi-gram 転置インデックスは，文字連続を正しく検索するためにブログ記事の全ての文

[1] `http://labs.nifty.com/beta/recipe/` にて公開中

1. 利用者から与えられた検索クエリ q を係り受け解析し，肯定クエリと否定クエリ[a] に分割する．
2. 各クエリから名詞，形容詞，動詞を抽出し検索語 $\{w_1, ..., w_N\}$ を作成．
3. 抽出された各検索語 w_n について処理．
   (a) 検索語 w_n を知識辞書を用いて拡張し，拡張検索語 $\{x_{n1}, ..., x_{nN}\}$ を作成．
   (b) 作成された各拡張検索語 x_{nm} について以下の 2 つのインデックスを用いて OR 検索を行う．
      i. x_{nm} について Word 転置インデックスを OR 検索．検索結果が得られた場合は，3(b)ii へ進まずに次の $x_{n(m+1)}$ の処理へ．
      ii. x_{nm} について Bi-gram 転置インデックスを OR 検索．
4. 否定クエリから抽出された検索語で得られえた記事を検索結果から除外する．
5. 検索結果として得られた各記事について式 (1)〜(3) でスコアを求める．

[a] “q = 美肌にならないスープ” という検索クエリの肯定クエリは “スープ”，否定クエリは “美肌にならない” である．

図 2: 検索処理の流れ

字列に対して位置情報つき Bi-gram を生成して構築した．

ランキングには式 (1)〜(3) の式を用いた．Score は検索語との合致度を表す MatchScore とブログ記事の固有スコアである SiteScore で構成される．MatchScore は検索語と記事における単語の一致数に重みをかけたものである．レシピ関連語などの重要な語のみがインデキシングされている Word 転置インデックスで発見された語のスコアを高くしている．SiteScore はページそのものの持つスコアを表わしている．第 1 項は，タイトルにレシピ関連語が多く含まれている記事の方がよりレシピを話題の中心に置いていると考えられるために置いている．第 2 項は，複数のレシピを掲載しているページに対するペナルティの役割を果たす．今回対象としたブログ記事の中には，1 記事内に複数のレシピが書かれているページや，レシピのサマリのみを集約したページも存在した．こういったページは式 1 の MatchScore において高いスコアを得て

しまい，適切な検索結果が得られない問題を生じさせるためこれらのページに対するスコアが相対的に低くなるようにこのペナルティを与えた．本来，レシピらしさを判定するためにはレシピ関連語数が多いほど高いスコアを与えるべきであるが，今回検索対象とした記事はあらかじめレシピが記載されている記事に限定したため，レシピらしさに関してはタイトルのみ (第 1 項) について考慮するようにし，本文は上述したペナルティのために用いた．

各スコアの算出式を以下に示す．

$$\begin{aligned} MatchScore &= (\text{記事中の一致 Word 数} * 5) + \\ &\quad (\text{記事中の一致 Bi-gram 数} * 1) \quad (1) \\ SiteScore &= (\frac{\text{タイトルに含まれるレシピ関連語数}}{\text{タイトルの総語数}}) + \\ &\quad (\frac{1}{\text{本文に含まれるレシピ関連語数}}) \quad (2) \\ Score &= MatchScore * SiteScore \quad (3) \end{aligned}$$

4 評価実験

4.1 実験条件

実験で用いた検索クエリは，「みんなのレシピ検索 β」公開後に一般利用者が入力した検索クエリと，3 名のアノテータが作成した自然文クエリから，計 275 件をランダムに選択した．アノテータは一般家庭の主婦であり，主婦の目線から，日常の料理で役立つ情報が得られそうな検索クエリを作成してもらった．作成した検索クエリの実例を以下に示す．

- 「メタボ撃退メニュー」
- 「鴨と合う食材は？」

検索エンジンは「みんなのレシピ検索 β 」を使用し，検索語拡張機能の有効／無効を切り替えて実験を行った．検索語拡張には，3 で述べた知識のみを使用した．

実験は以下のパターンで行った．

(A1) 検索語拡張なし

(A2) 意味的関係知識を用いた拡張（健康効果-食材）

評価は検索クエリの作成に携わった 3 名が行い，検索結果の上位 10 件について検索クエリの要求を満たしているかどうかを TRUE/FALSE で判定した．検索結果の上位 10 件を評価対象としたのは，実サービスにおいて利用者は検索結果の最初のページを重要視することが経験的に分かっており，検索結果の 1 ページ目に相当する上位 10 件を評価対象とすることで実際の現場に近い視点で評価することができると考えたためである．判定には，検索クエリと「みんなのレシピ検索 β 」の検索結果詳細ページ，実際のブログ記事の内容を使用した．

4.2 実験結果

検索結果の上位 5 件の総計は 1375 件，上位 10 件の総計は 2750 件である．両手法の検索結果上位 5 件と，上位 10 件を対象として検索クエリと適合していると評価された件数の累計を求めた結果を表 1 に示す．両手法を比較すると，(A1) より (A2) の方が累計適合件数が増加しているのが分かる．

表 1: 検索結果の累計適合件数

	手法 (A1)	手法 (A2)
上位 5 件	735	746
上位 10 件	1392	1422

検索結果の上位 10 件を対象として検索クエリ単位で適合件数を集計して比較した結果を表 2 に示す．対象とした検索クエリは 275 件中，手法 (A2) により検索語拡張が行われた 22 件である．

表 2: 検索クエリ単位の適合件数の比較

クエリ単位の適合件数比較	
(A1) の方が多い	(A2) の方が多い
3	13

さらに詳細に検索結果について分析を行った．まず，手法 (A2) により検索語拡張が行われた検索クエリを調べると，275 件中 22 件であることが分かった．先の表 2 より，検索結果が改善された検索クエリの件数は 13 件，悪化した検索クエリの件数は 3 件であり，意味的関係知識を用いた検索語拡張は，検索結果の改善に効果があるといえる．残りの 6 件については検索結果の適合数が変わらなかった．

検索結果が大きく改善された検索クエリの例として「花粉症に効果のある食材」を挙げる．この検索クエリにおいて検索語拡張の対象となった語は “花粉症” であり，{“にがり”,“ヨーグルト”,...} のように拡張される．“花粉症” という語は，ブログ記事中のレシピに直接関係ある場面より，レシピ前後に書かれる日記的内容の箇所に現れることが多い．このため検索語拡

張せずに “花粉症” のみで検索をすると，日記的内容の箇所に “花粉症” を記載したような，検索クエリとは不適合な記事が発見されてしまう．しかし，検索語拡張により効果のある食材名を検索語として使用することで，検索クエリに適合する記事を新たに発見することができた．

次に意味的関係知識を用いた検索語拡張をしたことで悪化した検索クエリとして「目の疲れに効く料理」があった．この検索クエリにおいて検索語拡張の対象となった語は “疲れ” であり，{“黒米”,“ゴーヤ”,...} のように拡張される．しかし，今回使用した意味的関係知識辞書における “疲れ” は “体全体の疲れ” の意味であり，本来は “目の疲れ” に効果のある食材へ展開されるべきだったが，“体全体の疲れ” に効果のある食材へ展開されてしまったため，検索結果が悪化したと考えられる．

このような他の語と合わせて使うことで意味が変わるような語については，意味的関係知識による検索語の拡張を行うべきかどうか検討すべきである．

今回，意味的関連知識による検索語展開が行われなかった検索クエリに注目すると，本来であれば検索語展開されることが期待される以下のような検索クエリが存在した．

- 目が悪い
- 胃にやさしいおかず
- 高脂血症で困っている

これらは今回使用した知識に該当する項目が存在しなかったことや，単語対単語の拡張を想定した知識であり健康効果や症状を文章で記述した検索クエリに対応できなかったことから，今回の実験では検索語展開をすることができなかった．

5 まとめ

本稿では，Web より獲得した意味的関係知識を用いた検索語拡張の効果の検証を行なった．実験の結果，健康効果と食材に関する意味的関係知識を用いることで検索語拡張前には得られなかった検索結果を新たに得られることをを確認した．また同時に，検索精度の改善に対し悪化した事例は少なく，リスクの少ない検索語拡張ができていることも確認できた．

実験対象とした 275 件の検索クエリのうち，クエリ拡張が行なわれた件数は 22 件であったが，そもそも今回用いた “健康効果と食材” に関する意味的関係知識が使われるべきであった検索クエリは 25 件であった．ここから次の 2 つのことが言える．まず “健康効果と食材” に関する拡張が必要であった 25 件のうち 22 件は今回獲得した知識で拡張できているので，この知識のカバレッジは 88% と低くないものであることが確認できた．次に，今回用いた知識では，検索クエリ全体のうち 1 割弱の検索クエリについてのみ適用が可能であった．今後は “健康効果と食材” 以外の知識についても同様に Web からの獲得を行ない，検索に用いていく必要がある．

また，本研究のような意味的関係知識による検索語拡張では，拡張に使用した知識をどのように提示するかということも重要な課題である．検索結果がどのような知識に基づいて得られたものであるかを適切に提示しなければ，利用者はその結果が得られた根拠を知らないまま検索クエリに適合していないという判断を下さなければならない．「みんなのレシピ検索 β」では，食材と健康効果の知識を含むいくつかの意味的関係知識について検索結果とともに提示しているが，これらの提示手法についても検討する必要がある．

ランキングのスコア計算に用いた式 (1)〜(3) やこれらの中で用いているパラメータについては，評価データを用いたパラメータチューニングを行なうなどまだまだ改善の余地が残されている．意味的関係知識を検索語拡張に用いた場合における適切なランキング手法を探ることも，今後の課題の一つである．

参考文献

[1] Rila Mandala, Rila M, Takenobu Tokunaga, and Hozumi Tanaka. Combining multiple evidence from different types of thesaurus for query expansion. In *SIGIR '99*, 1999.

[2] Stefan Riezler, Er Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. Statistical machine translation for query expansion in answer retrieval. In *ACL '07*, 2007.

[3] Stijn De Saeger, Jun'ichi Kazama, Kentaro Torisawa, Masaki Murata, Ichiro Yamada, Kow Kuroda, and Chikara Hashimoto. A web service for automatic word class acquisition. In *the 3d International Universal Communication Symposium (IUCS'09)*, 2009.

[4] Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, and Masaki Murata. Large scale relation acquisition using class dependent patterns. In *the IEEE International Conference on Data Mining (ICDM'09)*, 2009.

[5] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, Vol. 41, pp. 288–297, 1990.

[6] 海野裕也, 宮尾祐介, 辻井潤一. 自動獲得された言い換え表現を使った情報検索. 言語処理学会第 14 回年次大会論文集, 2008.