自動生成された検索ディレクトリ「鳥式」の現状

鳥澤健太郎　隅田飛鳥　野口大輔　風間淳一
{torisawa, a-sumida, noguchi-d, kazama}@jaist.ac.jp
北陸先端科学技術大学院大学 情報科学研究科

1　はじめに

現代人はありとあらゆる事象に関して適切な問題回避, あるいはイノベーティブなアイディアを切実に求めており, 検索エンジンはそのための格好のツールである, しかしながら, 適切な問題回避, イノベーティブなアイディアを得る為には, そもそも各自にとって想定外であるキーワードを検索エンジンに与える必要が往々にしてある.

本研究の目標は, ユーザーが最初に入力した検索トピックから, そうした想定外のものも含めた価値ある関連語を求め, 検索ディレクトリとして提示することであり, さらに, 検索ディレクトリに提示された関連アイテムが選択されたならば, 検索トピックと関連アイテムの両方に関係の深い文書を提示することである. 我々はこの目標を達成すべく現在検索ディレクトリを大量の Web 文書をもとに自動生成しているが, この検索ディレクトリを「鳥式」と呼んでいる. 「鳥式」では, ユーザーが入力する検索トピックの利用, 対処の文脈における検索にフォーカスする. より具体的には, とりあえず, 利用／対処の文脈にあって価値があることが自明な, トピックに関する（潜在的）トラブル, 方法, ツール／道具等の意味的カテゴリに属する関連語を, カテゴリ毎に分類して網羅的に提示し, ユーザーが想定外の関連語を探し出す支援を行なう.　ここでのポイントは, このように検索の利用／対処の文脈を限定すれば, 価値ある関連語を意味的にある程度限定することが可能であり, また, ユーザーは膨大な関連語に押しつぶされることなく, 比較的に客観的な意味的カテゴリにしたがって, 価値ある想定外のキーワードを発見できるということである. また, 関連語を限定するための別の方法として, 検索履歴等を参考に, 特定ユーザーに特化した関連語を提示することも当然考えられる. しかしながら, このようないわゆるパーソナライゼーションは, ユーザーの主観的観点をシステムに取り入れるという意味において, 我々の行なった利用, 対処という観点での文脈の限定に比べて遥かに難しい課題であると考えている.

現在, 鳥式には通常の HTML で関連語を提示するバージョンと, GUI を用いて関連語を提示するバージョンの二つがある. 図 1 に GUI 版でのブラウジングの様子を

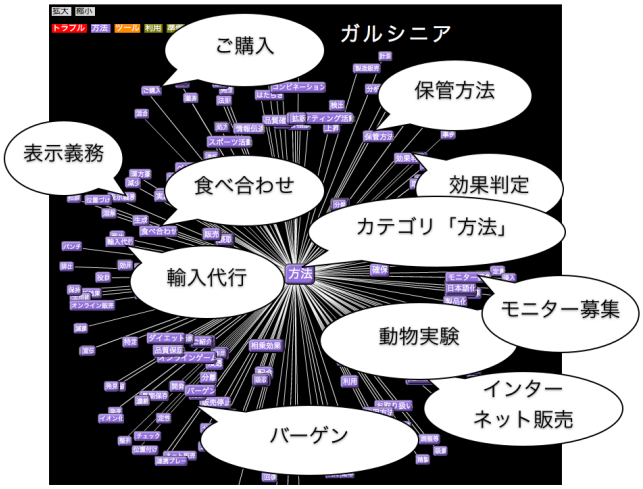

図 1　鳥式（GUI 版）でのブラウジング

示す. これは, ダイエットのサプリメントとして販売されている「ガルシニア」をトピック語としてユーザーが入力した後の状況である. 前述したように, 関連語は, トラブル, 方法等のいくつかのカテゴリによって分類されて提示されるが, この図では, 方法というカテゴリ中の関連語をブラウズしている. 中心にカテゴリ名の「方法」が表示され, その周囲に関連語が表示されている. 例えば, ガルシニアを利用するために必要な入手方法である,「輸入代行」「インターネット販売」などの関連語があり, 利用方法である「食べ合わせ」, さらに, 効果を確認する手段である「動物実験」「表示義務」なども提示されている. おそらく, ガルシニアを利用したいと思うユーザーであっても, 動物実験の結果をチェックしたり, 輸入代行によって入手を行なうといったことは稀であり, また, これらの検索キーワードを実際に検索エンジンに入力する可能性も低いであろう. 鳥式の目的はこのような「想定外」のキーワードをユーザーに気づかせ, 有用な情報を入手する支援を行なうことである. ユーザーは, 何もない所から関連語を思いついて検索エンジンに入力をおこなうのではなく, 単に提示されたキーワードをクリックするだけで, トピックと関連語の両方を含む文書を得ることができる.

なお, この GUI においては, 各関連語の中心からの距離は, 関連語のマイナーさの度合いを示し, 関連語間の角度は関連語の意味的類似性を反映している. （現在の実装では意味的類似度の計算法に問題があり, かならずし

も, このようには動作していない. 今後改善の予定である.) このようなデザインにした意図は, まず, メジャーな想定内の関連語の出現している位置によって, 欲しい想定外のキーワードの「方向」のあたりをつけ, そこからよりマイナーな関連語を中心から離れる方向に探すことで, 「欲しい」想定外の関連語が効率的に見つけられるようにするためである. これにより, 提示された関連語をはじから逐一チェックするのではなく, より効率的に価値ある想定外の関連語が見つけられるようになると期待している.

2 想定外？

さて, これまで「想定外」とひとことで述べてきたが, 一般に「想定」の意味する所は非常に曖昧であるので, ここで少し整理をしてみたい. 人によっては例えば, 「トラブル」という言葉を意識に昇らせた瞬間, 「可能な (タイプとして既知の) トラブル」(e.g., 肝炎) はすべて「想定内」と言うかもしれない. しかしながら, 我々は, こうした立ち場はとらない. 一言で言えば, 検索エンジンを前にして, ユーザーが検索キーワードとして実際に入力はしないが, 実は重要で価値がある関連語を, 想定外であると見なす. 特に, 「トラブル」のような一般的なキーワードを想起することは簡単であるが, あるトピック (e.g., アガリクス) に関するトラブルとして「肝炎」のような詳細度の高い具体的なキーワードを網羅的に想起し, 検索キーワードとして入力することは容易ではない. ここで述べている想定外の関連語としては, そうした詳細度の高い具体キーワードがまずは候補として挙げられる. これらをある程度網羅的に提示して, ユーザーに価値ある情報にアクセスしてもらうのが鳥式の狙いである.

より具体的に, 我々が想定外であると考えている関連語としては, 以下のようなものがある.

1. そもそも未知の関連語
2. 存在は知っていたが, トピックワードと重要な関係があるとは思っていなかった関連語
3. トピックワードと重要な関係があるとは薄々思っていたが, わざわざ検索をしてみようとは思い至らない関連語

ここで, 以上のものの具体的な例をあげてみたいが, そもそもある関連キーワードが想定外であるか否かは, ユーザーの知識, 経験に依存する. したがって, 以下では本論文の第一著者の視点で例をあげるしかないが, 例えば, 1 のタイプの例としては, 「ディズニーランド」の関連語でトラブルと分類されて提示される「身長制限」がある. 第一著者は遊園地のアトラクションを利用する際に, 身長に制限があることは薄々知ってはいたが, 「身長制限」なる用語が存在することは知らなかった. また, 2 のタイプとしては, 「ダイエット」のツールとしての「砂糖」がある. もちろん, 第一著者はこれらの語の意味する所は知っていたが, 砂糖はダイエットをする上で障害となるものであり, それをツールとして利用するダイエット法があることは知らなかった. また, 3 のタイプとしては, 「skype」の「セキュリティー問題」がある. 多くの skype ユーザーは skype にセキュリティー問題があるかもしれないということは薄々思っているかもしれない. しかしながら, その内で実際にネット検索を行ない, skype のセキュリティーについてチェックをしたことのあるユーザーはまれであろう. ここでは, このような関連語も一種の想定外であると見なす. 仮に skype というキーワードを入力すると, 「セキュリティー問題」が提示され, それをクリックするだけで有用な情報が得られるのであれば, それだけ skype のセキュリティー問題の具体的な情報に触れるユーザーが増え, そこに有用性を見いだせるであろう.

ここで, 1 , 2 のタイプの想定外は, 万人に想定外として受け入れられるであろうが, 一方で 3 のタイプを「想定外」と呼ぶことに抵抗がある人は多いものと思われる. しかしながら, 今の文脈において重要なのは, 検索という行為, そしてその瞬間におけるキーワードの想起である. つまり, 検索という行為をしている瞬間に, 検索キーワードとして想起されていないが実は価値あるキーワードというのが重要なのである. このような立ち場に立てば, 3 のタイプもやはり想定外と捉えるべきであり, このような想定外も提示することに意義があることになろう. また, このような意味での想定外のキーワードは, 実は広範囲に渡り, これらをカバーするためには, 鳥式で行なっているようにある程度網羅的に関連語を表示することに意義があるということになる.

さらに一点重要なポイントを挙げると, これまでは暗に一般ユーザーの情報収集の支援という観点で説明を行なって来た. しかしながら, 類似したニーズは企業サイドにもあるはずである. 例えば, 企業にとって, ネット上で議論されている自社製品の問題点をチェックするのは, もはや必須であろう. 当然, 想定外の問題点もチェックを行う必要があり, 鳥式はそのような状況においても有用であると考えている. 現在のシステムでも例えば, 「XBOX」に対して「傷」といったものが提示されている. (ネット上で話題になっている XBOX の傷は, XBOX 本体の傷ではなくて, XBOX に挿入された DVD 等の傷である.) このような関連語を特に外部から示唆されることなく, 企業の担当者が思いつくのは, 往々にして困難であり, やはり, 鳥式のようなシステムである程度網羅

的に列挙する必要があると考えている.

3 鳥式の構造

鳥式では, 想定外を含む多数の関連語をユーザーに提示する以上, 人手で検索ディレクトリを作成するのは現実的ではない. 我々は, 大量の Web 文書から日本語表現の意味的分類を自動的に行ない, また, 分類された表現間の意味的関係を自動的に認識することで, 検索ディレクトリを自動生成している. 検索ディレクトリは二階層からなっている. 上位層は, **対象特定レベル**と呼ばれ, 検索のトピックを階層的に意味分類したものであり, 検索トピックの名称 (例:「アガリクス」) を特定できないユーザーがブラウジングによって, それを特定するのに利用される. 下位層は**様相特定レベル**と呼ばれるが, 上の対象特定レベルでのブラウジングによって検索トピックが充分に特定できた場合か, そもそもユーザーが検索トピックを直接入力した場合に, そのトピックの利用, 対処に関して関連語 (例: 「肝炎」) をブラウズしつつ情報を収集するレベルである. 現状ではクリックされた関連語とその検索トピックは既存の検索エンジンにクエリーとして与えられ, AND 検索を行なうことで関連のある文書が提示されるようになっている. また, 関連語の提示方法としては, 前述の GUI で例示したように, 連続的な意味的類似性や, 関係付けの一種の強度といった情報も用いて, 最終的な目標である想定外の発見の支援を行なう手法を開発中である.

以下では, 対象特定レベルと様相特定レベルのそれぞれに対して, その自動生成手法の概要を述べる.

■対象特定レベル 対象特定レベルは, 検索トピックを階層的に分類したものである. 一般に, 「北陸先端大学院大学」と「大学」のように, ある語とその語の属する上位概念のラベルとなる語の間の関係を上位下位関係と呼び, 上位概念のラベルを上位語, その上位概念に属する語のことを下位語と呼ぶ. (本来, 上位下位関係はあくまで概念間の関係であり, 概念とそれに属するインスタンスの間の関係は含まないが, ここでは議論を単純にするため, 概念/インスタンス間の関係も上位下位関係に含めることにする.) 対象特定レベルは一言でいえば, 大量の上位下位関係の集合であり, それを階層的に提示するものである.

現在の対象特定レベルは, 128 万語のトピック語をカバーし, それらの間の関係が 247 万個の上位下位関係で結ばれている. これらの上位下位関係は大量の Web 文書ならびに Wikipedia から自動で抽出されたものであり, 精度は約 90% である. 技術的には, Wikipedia から階層構造, 定義文, カテゴリーページに機械学習を適用することで上位下位関係を自動獲得している (4) 他, 一般の Web 文書に語彙統語パターンを適用し上位下位関係を自動獲得している (3).

■様相特定レベル 様相特定レベルにおいては, 対象特定レベルにおいて特定されたトピック語, もしくは, ユーザーが直接キーボードから入力したトピック語に関係のある関連語を, それらの意味的分類にしたがって提示する. この意味的分類としては, 現在, トピック語の利用／対処を想定した時に有用である, 以下のような基本的なカテゴリを用意している.

トラブル トピック語が指し示す対象, すなわちトピックを利用する, あるいはトピックに対処する上で障害となる (潜在的) トラブルのカテゴリである. 例えば, トピックを「ディズニーランド」とすると, それを利用する上で障害となる「身長制限」, 「渋滞」等は, このカテゴリに属する.

方法 トピックを利用／対処する上で有用／必要な具体的方法を含むカテゴリーである. 例えば, ダイエットサプリメントである「ガルシニア」を利用するにあたってはそれを購入する必要があるが, そのための一方法である「輸入代行」などがこれに属する.

ツール/道具 トピックを利用する／対処する上で, 道具は重要である. 例えば, 「ダイエット」に対処する, つまり, ダイエット行なう際にサプリメントは道具として重要であるが, 「ガルシニア」のような具体的なサプリメントの名称は「ダイエット」というトピックの関連語としては, ツール／道具というカテゴリのもと提示される.

場所 トピックを利用する／対処する上で, 場所が重要なことは多々ある. 例えば, 「魚」の「鮎」を食べるにあたっては, どの川の「鮎」であるかは重要な要素であろう. この観点から, トピックと関連の深い場所も様相特定レベルでは提示する.

実際に, これらのカテゴリで提示される関連語は, まず, トピックとは無関係に教師あり学習によって大量のコーパスから自動的に獲得される (1). より具体的には, 検索エンジン TSUBAKI(2) が提供している大量の Web 文書から係り受け関係を抽出し, 各名詞に対して, それと係り受け関係にある動詞, 助詞の対を素性として学習を行なう. より具体的には, 名詞の中から抽出されたサンプルに対して, 各カテゴリに属するか否かの判断を人手で行なった学習データを作成し, それをもとに SVM で学習を行ない, その結果得られた二値分類器で, 全ての名詞が各カテゴリに含まれるか含まれないかの判定を行なう. 現在, 各カテゴリに対して, 各 3 万語程度の学習データを用意して, これをもとに, トラブル, 方法, ツー

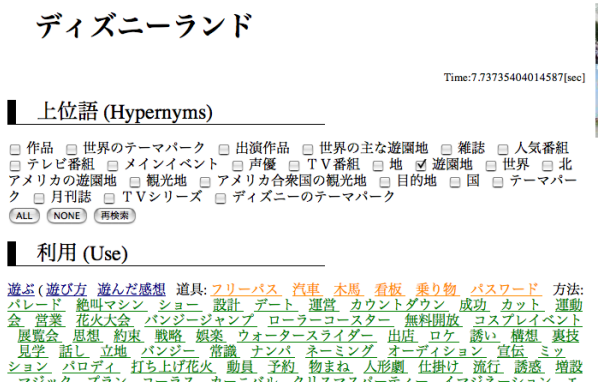

図 2　様相特定レベルでのブラウジング：ディズニーランドの利用を表す「遊ぶ」と関連語が提示されており，また継承用に上位語がチェックボックスと共に表示されている．)

ル／道具，場所に分類される合計 30 万語以上の語を得ている．（精度はカテゴリによって異なるが，70% から 80% 程度である．)

次のステップとしては，各トピックと，上記の方法で得られた各カテゴリ中のキーワードとの関連を，「＜トピック＞の＜キーワード＞」のような語彙統語パターン，もしくは，動詞との共起関係などをもとに計算する．例えば，コーパス中に「ディズニーランドの身長制限」という表現が出現したとすると，トピック語であるディズニーランドに，身長制限というトラブルに属する語が関連していると判断される．

以上で，各トピック語に対して，関連語の意味的分類が得られるが，他にも (5) にある手法によって獲得された，トピック語の利用を表す動詞，もしくは利用をするための準備を表す動詞も同時に表示される．(図 2 参照) さらに，上述したトラブル等のカテゴリに含まれる関連語も，利用／準備を表す動詞と関連づけられた形で表示される．例えば，「アオブダイ」に関するトラブルである「毒」は，アオブダイの利用を表す動詞「食べる」に関連したトラブルとして提示され，また，「ディズニーランド」に関するトラブルである「渋滞」は，ディズニーランドを利用する準備である「行く」という動詞に関連したトラブルとしてユーザーに提示される．これらの動詞も有用な想定外の関連語を探すための手がかりとして利用できる．

また，様相特定レベルの重要な機能として継承がある．これは，あるトピック語が入力された時に，その上位概念の関連語を入力された検索トピックの関連語として表示するものである．現在，検索トピックと関連語間の関係は主として語彙統語パターンによって獲得されているが，そのカバレージは低頻度の検索トピックに関しては充分とは言えない．このような場合に，より高頻度で出現する上位概念の関連語で補完を行なう訳である．図 2 にある上位語にはチェックボックスが共に表示されているが，これは継承を行なう上位概念をユーザーが指定する為に使われている．

4　おわりに

以上，簡単に検索ディレクトリ「鳥式」の概要について述べて来た．今後の課題としては，精度及びカバレージの向上のほか，より直接的に「価値ある想定外」を見つける支援を行なうことがある．一つの可能性としては，トピックとその上位概念，それぞれの関連語を比較することが挙げられる．例えば，トピック「アオブダイ」とその上位概念「魚」，さらにそれぞれと関連したキーワード「毒」というものを考えると，実際に毒を持つ「アオブダイ」と「毒」の共起頻度は，「魚」と「毒」の共起頻度より相対的に大きくなるものと考えられる．このような上位概念とトピック語の比較を行ない，特異な振る舞いをする関連語を抽出できれば，それらは「価値ある想定外」になる可能性が高いのではないかと考えている．

また，より広い視点で今後を展望すると，「想定外」の情報を入手する機能は今後ますます重要になるものと考えている．例えば，今後ロボット等が普及したとすると，ユーザーにとって想定外であるようなトラブルを警告したり，ユーザーが行なっている行動をより適切にするためのアドバイスを行なう機能がそのロボット等にも期待されるであろう．現在，鳥式で提供すべく自動獲得している一群の言語資源，知識は，そのような高度なサービスを行なう際の基礎として利用できるものと考えており，今後，さらに自動獲得手法を改善すると同時に，人手でのクリーニングを行ない，適宜一般公開して行く予定である．

参考文献

[1] Stijn De Saeger, 鳥澤健太郎. トラブルを見つける. 言語処理学会第 1 4 回年次大会予稿集, 2008.

[2] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. Tsubaki: An open search engine infrastructure for developing new information access. In *Proc. of IJCNLP*, pp. 189–196, 2008.

[3] Asuka Sumida, Kentaro Torisawa, and Keiji Shinzato. Concept-instance relation extraction from simple nouns sequences using a search engine on a web repository. In *Proc. of the Workshop on the Web Content Mining with Human Language Techonologies*, 2006.

[4] 隅田飛鳥, 吉永直樹, 鳥澤健太郎, 萬成賢太郎. Wikipedia からの大規模な上位下位関係の獲得. 言語処理学会第 1 4 回年次大会予稿集, 2008.

[5] 鳥澤健太郎. 対象の用途と準備を表す表現の自動獲得. 自然言語処理, Vol. 13, No. 2, pp. 125–144, 2006.

[6] 風間淳一, 鳥澤健太郎. Web 上の資源から構築した複数の固有表現辞書を用いた日本語固有表現認識. 言語処理学会第 1 4 回年次大会予稿集, 2008.