E-039

ユーザの要望に対する Web 検索型推薦システム

Recommendation System to Fulfill a Request of User Using Web Retrieval

稲井 聡† 吉村 枝里子† 土屋 誠司† 渡部 広一†

Satoshi INAI Eriko YOSHIMURA Seiji TSUCHIYA Hirokazu WATABE

1. はじめに

現在，インターネットの普及により PC や携帯電話などに Web 機能が標準装備され，以前と比べ気軽に自分の行きたい場所や欲しい商品などの情報を入手する事が可能になった．しかし，Web の情報量は日々増加しており，その雑音の多い Web ページから自分にとって有益な情報を検索エンジンで見つけ出す労力が増えて来ている．そこで，雑音の多い Web からの情報収集に慣れていないユーザでも，自分の行きたい場所や欲しい商品名などを自動で得られる検索・推薦システムの構築が必要であると考えられる．

本稿では，場所や商品名などを求めるユーザの要望を表し，かつ文末が「～したい」の文（以下，要望文）を基に Web で情報収集を行い，その要望に適合した回答を出力するシステムを提案する．なお，本システムは一般人がよく Web 検索で調べていると想定される，行きたい店などの場所や欲しい商品名などの固有名詞の出力をする．

2. Web 検索型推薦システムの概要

本システムでは，入力された要望文をもとに Web 検索を行い，その要望に適合した複数の場所や商品名などを出力する．図 1 にシステムのイメージを示す．

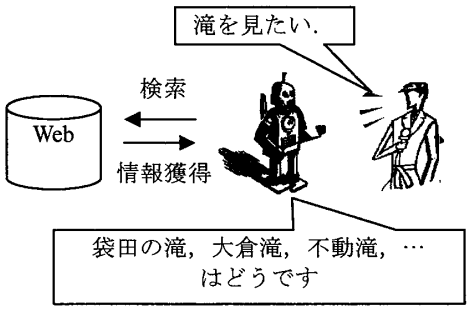

図1 Web 検索型推薦システムの使用例

3. 使用技術

3.1 概念ベースと関連度

概念ベース[1]は，複数の国語辞書や新聞などから機械的に構築した語（概念）に対し，その意味・特徴を表す語（属性）とその重要性を表す数値（重み）の組を付与した知識ベースである．概念ベースには，約 12 万語の概念が収録されている．なお，本稿では概念ベースに登録されていない概念を未定義語と定義する．

関連度[2]とは，概念ベースに定義された語と語の関連の強さを，同義性・類似性のみに関わらず定量化した値である．なお，関連度は概念ベースに存在する属性・重みを用いて算出するため，未定義語は関連度を計算することが出来ない．

3.2 シソーラス

シソーラス[3]とは，単語の意味や概念を分類，整理して用語を階層的に体系化したものである．各節点に相当する語をノード，ノードに含まれる語をリーフと呼ぶ．

シソーラスは，一般名詞の意味的用法を表す約 2700 語の意味属性(ノード)の上位下位関係・全体部分関係が木構造で示されたものであり，約 13 万語(リーフ)が登録されている．

3.3 未定義語の属性・重み獲得手法

未定義語の属性・重み獲得手法[4]とは，未定義語の意味的特徴を表す属性とその重要性を表す重みの組を Google[5]を用いて獲得する手法である．なお，本稿では未定義語の属性・重み獲得手法をオートフィードバック(AF)と呼ぶ．

3.4 Web から自立語を獲得する手法

Web から自立語を獲得する手法[6]は，入力した語句をもとに Google で検索を行い，上位 100 件の検索結果ページ内から入力した語句と共に出現する自立語(概念ベースに含まれない語句も含む)を，自動的に獲得する手法である．なお，本稿では Web から自立語を獲得する手法を拡張オートフィードバック(拡張 AF)と呼ぶ．

3.5 シソーラスマッピング

シソーラスマッピング[7]は，AF と関連度を用いることで，シソーラスに定義されていない語句が大局的にどのノードに所属するかを判断する手法である．

4. システムの流れ

本システムの流れは，まず入力された要望文から検索にかける語句を作成する．次にその語句で Web 検索を行い，要望文の回答の候補となる自立語を取得する．最後に形態素解析やシソーラス・関連度などを用いて，取得した自立語の処理・順位付けを行い出力する．各々の処理の流れについて，次の 4.1 節～4.4 節で説明する．

4.1 検索にかける語句の作成

要望文を形態素解析し，動詞・名詞・形容詞を抽出する．この時，要望文「TV を買いたい．」のように出力すべき内容が「TV の機種」か「TV を販売する店」の 2 種類想定される場合，ユーザに「あなたの欲しい回答は場

† 同志社大学大学院工学研究科
Graduate School of Engineering , Doshisha University

所ですか？」と聞き，それらの結果を踏まえ検索にかける語句（検索語）を作成する．表 1 に検索語の作成例を示す．

表 1　検索語の作成例

要望文	検索語	質問
TV を買いたい	TV　買う　店　場所	Yes
紅葉を見たい	紅葉　見る	ー

4.2　回答候補語の取得

拡張 AF を用いて，回答の候補となる自立語(以下，回答候補語)30 語を Web から取得する．この時，回答候補語の中に検索語と同じ語句・動詞・形容詞・記号も含まれる．本システムは場所などの固有名詞の出力を目指している為，それらを削除する．図 2 に検索語「紅葉　見る」を基に拡張 AF で取得した回答候補語の処理の例を示す．

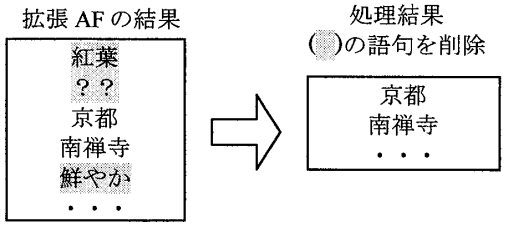

図 2　回答候補語の処理

4.3　シソーラスによる精錬

シソーラスを用いて，検索語の名詞のノードと回答候補語のノードを取得する．なお，シソーラスに存在しない語句はシソーラスマッピングを用いてノードを取得する．次に回答候補語のノードを検索語の名詞のノードと比較し，共通するノードが 3 個未満の回答候補語を削除する．図 3 に，検索語の一つ「紅葉」と回答候補語「奥津渓」と「海の精」の精錬の例を示す．

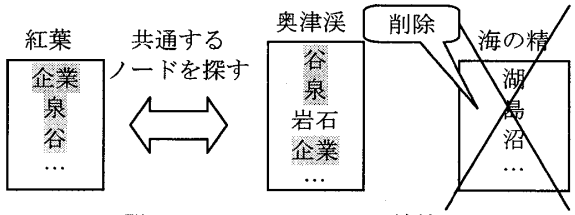

図 3　シソーラスによる精錬

4.4　関連度による精錬

まず，回答候補語の属性を AF で取得する．次に要望文と解答候補語の関連を調べる為に全ての検索語を一つの概念とみなし AF で属性を取得する．そして検索語と各々の回答候補語の関連度を調べ，関連度の高い順に回答候補語を並び替える．表 2 に関連度による精錬の例を示す．

表 2　関連度計算による精錬の一例

検索語	解答候補語	検索語との関連度
紅葉　見る	奥津渓	0.685
	高尾山	0.522
	長谷寺	0.467
	・・・	・・・

5. 評価・考察

アンケートで収集した要望文 50 文と出力すべき回答の種類を基に，システムの評価を行った．被験者 3 人に要望文 50 文の出力結果を提示し，要望文に対する解答として適切か否かチェックしてもらった．要望文 50 文のうち，上位 1 位に 2 人以上の被験者が正解とした回答候補語が出力された文は 32 文，上位 2 位までででは 37 文，上位 3 位までででは 38 文存在した．

不正解とした出力の例として，要望文「RPG のソフトを買いたい」に対し出力された「プレステ」「任天堂」「在庫」がある．「プレステ」「任天堂」は，シソーラスによる精錬においてユーザが欲しいノード「ソフト」ではなく上位の「映画・映像」に分類された為，削除する事が出来なかった．また，「在庫」は検索語「RPG　ソフト　買う」と共に通販サイトなどの Web ページ上に頻出される為，多くの共通属性を獲得してしまった．この事により関連度が高くなり，上位に出力されてしまったと考えられる．

6. おわりに

本稿では，「商品名」や「場所」などを求めるユーザの要望に対する Web 検索型推薦システムを提案した．その結果，上位 3 位で 50 文中 38 文正しい答えを少なくとも一つ出力できた．

この提案手法により，ユーザが得たい「本のタイトル」や「店の名前」などの検索をある程度コンピュータに任せる事が可能になると考えられる．

7. 謝辞

本研究の一部は，科学研究費補助金(若手研究(B)21700241)の補助を受けて行った．

参考文献

[1] 奥村紀之，土屋誠司，渡部広一，河岡司：“概念間の関連度計算のための大規模概念ベースの構築”，自然言語処理，Vol.14，No.5，pp.41-64，2007.
[2] 渡部広一，奥村紀之，河岡司：“概念の意味属性と共起情報を用いた関連度計算方式”，自然言語処理，Vol.13，No.1，pp.53-74，2006.
[3] NTT コミュニケーション科学研究所：“日本語語彙体系”，岩波書店，1997.
[4] 梅田司：“Web を用いた未定義概念の属性獲得手法”，同志社大学工学部知識工学科卒業論文，2005.
[5] “Google”, http://www.google.co.jp/
[6] 辻泰希，渡部広一，河岡司：“www を用いた概念ベースにない新概念およびその属性獲得手法”，第 18 回人工知能学会全国大会論文集，2D1-01，2003.
[7] 後藤和人，土屋誠司，渡部広一，河岡司：“Web を用いた未知語検索キーワードのシソーラスノードへの割付け手法”，自然言語処理，Vol.15，No.3，pp.91-113，2008.