診断型 WWW 情報検索システム (3) 診断機能*

1AC－6　二瓶克己　富沢伸行　柴田晃宏　島津秀雄†

NEC C&C メディア研究所‡

1　はじめに

ユーザサポート部門向けの情報検索システムである WWW 版 Help Desk Builder を開発した [4][5]. 本稿では,WWW 版 Help Desk Builder の診断機能について示す. Java アプレットで実装した診断機能は, 階層構造インデックスとデータの生起確率を利用して, 検索結果の絞り込みに効率の良い条件を質問とその回答選択肢という形で順番にユーザに提示する.

2　診断機能の設計方針

診断機能の設計においてとられた主な方針は以下の 3 点であった.

1. 作成, 拡張が容易であること
2. 適用範囲が広いこと
3. 既存インデックスを活用すること

従来の診断型エキスパートシステムは診断能力は高いがルールの作成, 維持は困難であるため, ユーザサポート部門のようなデータが次々と追加されていく場合には問題となる. またその対象とする領域は限定されたものである. 本診断機能の場合, 作成が容易かつデータの追加にも対応できること, 診断能力は領域限定した場合よりは落ちても広範囲に適用できることを重視した. Help Desk Builder では検索対象を多観点から階層的に分類するインデックスを持っており, その構造情報を診断機能に利用している [3].

ここで診断機能の基本的な考え方を説明する. 例えば, 農家からの雑草に関する問い合わせに対応するシステムの場合, 図 1のような雑草を葉の形や生息地から階層的に分類した雑草データへの階層構造インデックスを持つ. 検索結果としてアメリカセンダングサ, シロバナセンダングサ, コセンダングサ, シオザキソウ, センダングサの 5 件の雑草データが得られていたとする. 階層構造

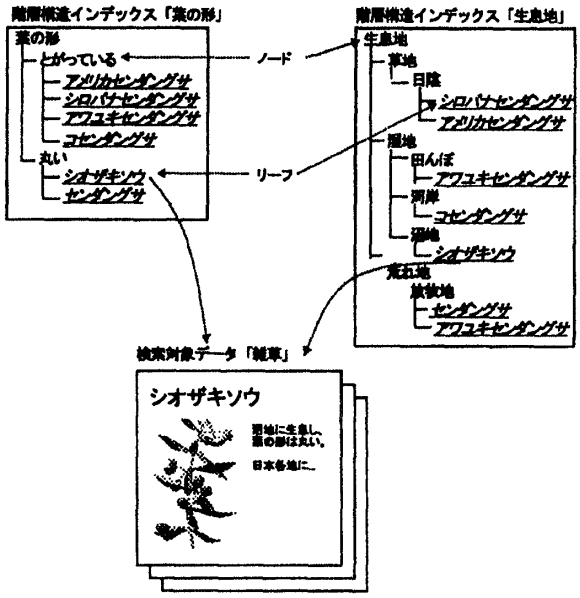

図 1: 階層構造インデックスの例

インデックス「葉の形」では葉の形がとがっているならば 3 件, 丸いならば 2 件に絞り込めることが分かる. 同様に, 階層構造インデックス「生息地」でも, 生息地が草地なら 2 件, 湿地なら 2 件, 荒れ地なら 1 件に絞り込まれる. ユーザに質問「葉の形」, 回答「とがっている」, 「丸い」と提示して回答を選択させることで絞り込みが実行できる. このとき「葉の形」と「生息地」どちらを先に質問するかで絞り込みの効率が変わってくる. そこで, 絞り込みに効率的な質問の順位付けを行なう.

さらに,Help Desk Builder に実装した診断機能 [3] を使用して得られた知見から, 本 WWW 版 Help Desk Builder の診断機能は以下の 3 機能を追加した.

(1) 検索対象データの生起確率の使用　よく検索されるデータと滅多に検索されないデータを同じように扱うのは効率的ではない. よく検索されるデータを含む検索結果集合へと早く絞り込めるような質問順序を計算で求めるために, 検索対象データに生起確率を設定する.

(2) 選択されなかった質問の順位制御　提示した質問の内, 選択された質問の上位にある質問はユーザにとって回答の対象にはなりえない場合も見られた. そこで, 選択されなかった質問が再び質問として上位に現われた場

*A Diagnostic-based Information Retrieval System on the WWW (3), Diagnostics.

†Katsumi NIHEI, Nobuyuki TOMIZAWA, Akihiro SHIBATA, and Hideo SHIMAZU

‡C&C Media Research Laboratrics, NEC Corp.

合，順位を下げる制御を行なう．

(3) 質問に使用しない階層構造インデックスの指定　質問として提示するのに適切でない階層構造インデックスは，質問に使用しないよう指定可能にする．

3　質問順の計算

質問順の計算方法について説明する．質問ノードとは，検索結果として得られたデータへのインデックスを持つ階層構造インデックスの各リーフから，階層を上げていって最初にたどりつく共通ノードとする．回答ノードとは，検索結果として得られたデータへのインデックスを持つ各リーフから，階層を上げていって得られた質問ノードの一つ下の階層のノードとする．

ステップ１　質問に使用可能であると設定された階層構造インデックスから質問ノードとその回答ノードを得る．

ステップ２　質問順を決定する．質問順の決定はID3[1]等で使用されている期待獲得情報量最大化原理にもとづく．期待獲得情報量の大きい質問ノードから質問することで質問回数を最小にすることが期待される．

ステップ2-1　生起確率と検索結果集合の情報量を計算する．C を検索結果の集合，k を検索結果数，検索結果データ $r_j(1 \leq j \leq k)$ の検索回数を h_j とした場合，検索結果データ r_j の生起確率 p^j と，検索結果集合 C の情報量 $M(C)$ は以下の式で表される．

$$p^j = \frac{h_j}{\sum_{i=1}^{k} h_i}$$

$$M(C) = -\sum_{j=1}^{k} p^j \log_2 p^j$$

ステップ2-2　検索結果集合 C を質問ノード a の回答ノード $a_1, a_2, ..., a_n$ によって部分集合 $C_1, C_2, ..., C_n$ に分割した場合の期待情報量 $B(C,a)$ を計算する．

$$B(C,a) = \sum_{i=1}^{n} \frac{|C_i|}{|C|} M(C_i)$$

ステップ2-3　質問ノード a の獲得情報量の期待値 $gain(C,a)$ を計算する．

$$gain(C,a) = M(C) - B(C,a)$$

ステップ2-4　検索結果集合にまだ取り出してない質問ノードがあるならステップ2-2へ戻る．そうでないならステップ３へ進む．

ステップ３　質問順の調整を行なう．前回選択した質問よりも上位にあった質問がステップ２で得られた質問の中に存在した場合，質問順位を下げる．

ステップ４　質問順にしたがって質問ノードと回答ノードをユーザに提示し，ユーザからの入力を受け付ける．回答ノードを選択すると，絞り込みを実行するため，選択された回答ノードによって得られるリーフに設定されたデータを新たな検索結果集合とし，ステップ１へ戻る．

4　診断実行例

農林水産省草地試験場と共同開発した農家からの雑草に関する問い合わせに対応する知識情報検索システム[2]を拡張したものを題材として診断実行例を示す．収録されている雑草は外来雑草194件である．階層構造インデックスは和名，葉，花，生息地など12個でうち質問に使用するのは９個，総ノード数は2903である．194件の雑草に対し，診断実行により生成された質問は46であった．図2に生成された質問と回答選択肢の例を示す．

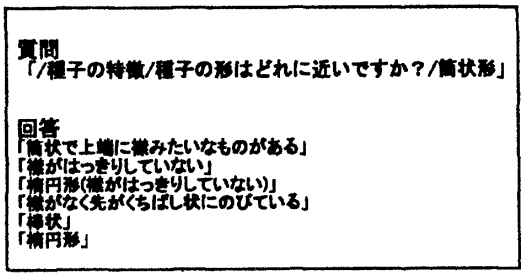

図 2: 生成された質問と回答選択肢の例

5　おわりに

ユーザサポート部門向けの情報検索システムであるWWW版 Help Desk Builder の診断機能について示した．Javaアプレットで実装した診断機能は，階層構造インデックスとデータの生起確率を利用して，検索結果の絞り込みに効率の良い条件を質問とその回答選択肢という形で順番にユーザに提示する．

診断実行例に用いた雑草データを作成した農林水産省草地試験場 黒川俊二氏に感謝する．

参考文献

[1] J.R.Quinlan, "Induction of Decision Trees" , Machine Learning, Vol.1, pp.81-pp.106, 1986.
[2] "農林水産業の高度情報システム - 農林水産業における高度情報システム開発に関する調査委員会報告書 -", 農林水産技術情報協会, 1996.
[3] 二瓶ほか,"ヘルプデスク構築支援システム「Help Desk Builder™」の開発 知識情報の検索 -Help Desk Builder/BT-", 情処54全大, 1997.
[4] 柴田ほか,"診断型WWW情報検索システム (1) 開発方針", 情処55全大, 1997.
[5] 富沢ほか,"診断型WWW情報検索システム (2) 構成と実装", 情処55全大, 1997.