7J－4

類似事例検索システム
―通信ソフト故障診断問題への適用―

岡本 青史† 佐藤 健† 塙 順吉‡ 松尾 勝美‡
（株）富士通研究所† 富士通株式会社‡

1. はじめに

近年，人工知能の分野における新しい推論の枠組として，事例ベース推論(CBR)や記憶に基づく推論(MBR)が注目を集めている．これらの枠組では，過去の事例を直接知識として用いることで問題解決を行なうため，エキスパートシステムにおける知識獲得のボトルネックを回避したシステムの構築が可能となる．その一方で，既存のCBRシステムやMBRシステムは問題領域に強く依存して提案されているために，汎用性に乏しく，新しいシステムの構築に要する時間的コストが大きくなるといった問題も抱えている．

本論文では，CBRやMBRにおける重要な構成要素の一つである類似事例検索に注目し，事例が自然言語で記述された問題に対して汎用的な類似事例検索システムの提案を行なう．提案システムにおける専門家からの知識獲得は，本質的に事例のカテゴリ分類だけである．さらに，提案システムを通信ソフト故障診断問題に適用し，提案システムの有効性を示す．

2. システムの概要

提案システムは，**キーワード抽出，重み学習，類似事例検索**から成り，その概要を図1に示す．

キーワード抽出：事例ベース中の自然言語で記述された事例から，社内で開発された形態素解析ツールを使用することで，各事例からキーワードを抽出し，キーワードインデックスファイルを生成する．キーワード抽出において，必要であれば，ユーザ定義語辞書およびシソーラスを使用する．

重み学習：一般的に，類似事例検索はk-最小近傍法によって実現される．この場合，各キーワードに対する重み学習と事例間の類似度の定義が，システムの検索性能に直接影響を与える．このため，様々な重み学習の手法が提案されている[1, 2, 3]．

Aha等[1]は，事例検索の成功／不成功によって重みの値をインクリメンタルに学習させる手法を提案した．Creecy等[2]は，属性の出現の有無に関する統計的情報を用いた重み学習法を提案した．Mohri等[3]は，多変量解析における数量化II類に基づく手法を提案した．

本論文では，事例の属するカテゴリ名が記述されたカテゴリ情報ファイルを用いて，Creecy等[2]が提案した*Per Category Feature Importance*(PCFIと略す)を修正した以下の重み学習法を提案する．

総事例数をN，カテゴリjに属する事例数をN_j，キーワードiが出現する事例数をA_i，カテゴリjに属してキーワードiが出現する事例数をA_{ij}とおく．この時，カテゴリjにおけるキーワードiに対する重みW_{ij}を以下のように定義する．

$$W_{ij} = \left(\frac{A_{ij}}{A_i} - \frac{N_j}{N}\right) + \left(\frac{A_{ij}}{N_j} - \frac{A_i}{N}\right)$$

ここで$W_{ij} < 0$の場合は，$W_{ij} = 0$とする．

この重み学習法を用いて，キーワードインデックスファイルとカテゴリ情報ファイルから，重みファイルを生成する．

類似事例検索：類似事例検索では，新事例に対するキーワード抽出を行なった後，重みファイルを用いて事例ベース中の各事例と新事例との類似度を計算し，類似度の大きい順にユーザ指定数の類似事例をユーザに提示する．ここで，新事例Pとカテゴリjに属する事例C_jとの類似度$Sim(P, C_j)$を以下のように定義する．

$$Sim(P, C_j) = \frac{S_{P \wedge C_j}}{S_{C_j}} \sum_i W_{ij}\ f(P, i)\ f(C_j, i)$$

ここで，S_{C_j}は事例C_jに出現するキーワード数を表しており，$S_{P \wedge C_j}$は新事例Pと事例C_jの両方に出現す

Similar Cases Retrieval System
— Toward a Diagnosis Problem in Communication Software —
Seishi Okamoto†, Ken Satoh†, Junkichi Hanawa‡, and Katumi Matsuo‡
Fujitsu Laboratories Ltd.† and Fujitsu Ltd.‡
e-mail: seishi@flab.fujitsu.co.jp

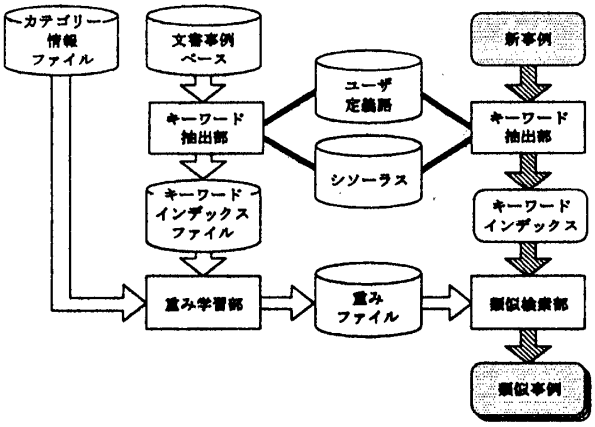

図 1: システムの概要

るキーワード数を表している．また関数 f は，事例 C とキーワード i に対して，以下で定義される．

$$f(C,i) = \begin{cases} 1 & \text{事例} C \text{に} i \text{が出現する場合} \\ 0 & \text{事例} C \text{に} i \text{が出現しない場合} \end{cases}$$

本提案システムにおける専門家からの知識獲得は，本質的に事例のカテゴリ分類のみであり，知識獲得のボトルネックはある程度解消されている．さらに，本システムの構築にかかる時間的コストは，カテゴリ分類に要するコストとほぼ等しくなるため，小さいと言える．

3. 通信ソフト故障診断問題への適用

本論文で扱う通信ソフト故障診断問題において，各事例は自然言語で記述された交換機ソフトの障害状況とその対処法で表現されている．この問題では，現在の障害状況を自然言語でシステムに入力し，過去の類似した障害状況を持つ事例を検索することで，ユーザの現在の問題解決を支援することを目的としている．

事例ベースには 1668 の事例が格納されており，それらの事例を障害状況に応じて 212 のカテゴリに分類した．この事例ベース中の事例を，1566 の訓練事例と 102 のテスト事例に分割し，テスト事例に対する正答率を計算することで，提案システムの検索性能を評価する．ここで，テスト例と同じカテゴリに属する事例が検索された複数の事例中に存在する場合に正解として，正答率はシステムが正解した確率で表される．また，抽出されたキーワード数は 2749 であり，1 事例に出現するキーワードの平均数は 11.5 である．

検索事例数	PCFI	提案手法
5	81.4%	88.2%
10	86.3%	92.2%

表 1: 正答率

Creecy 等 [2] が提案した重み学習手法 PCFI と，我々の提案手法の正答率を表 1に示す．表 1から分かるように，いづれの検索事例数の場合も，我々の提案手法は PCFI よりも高い正答率を得ることが出来ている．

さらに，我々の手法では，自然言語で障害状況が記述されたテスト事例を入力して，類似事例を検索するまでに要する時間は平均 1.2 秒であり，1566 の訓練事例から重みファイルを生成するのに要する時間は 7.8 秒である（共に，S-4/IX を使用）．このように，検索時間，重み計算時間とも実際の運用上問題がなく，インクリメンタルな類似事例検索システムを可能にしている．また，本提案システムは，自然言語で記述された事例がカテゴリ分類されている問題に対して汎用的であり，特許文書の類似検索や，ヘルプデスクシステム等に容易に適用できると考えられる．

4. 謝辞

通信ソフトウェア品質保証部の高橋部長，ネットメディア研究センターの浅川センター長並びに丸山主任研究員には，本研究を行なう機会を与えて頂きました．また，ソフトウェア研究部の颯々野研究員には，形態素解析ツールを提供して頂きました．深く感謝申し上げます．

参考文献

[1] Aha, D.W., Kibler, D. and Albert, M.K. Instance-Based Learning Algorithms. *Machine Learning, 6*, pp.37–66, 1991.

[2] Creecy, O.H., Masand, B.M., Smith, S.J. and Waltz, D. Trading Mips and Memory for Knowledge Engineering. *CACM, 35*, pp.48–63, 1992.

[3] Mohri, T. and Tanaka, H. An Optimal Weighting Criterion of Case Indexing for Both Numeric and Symbolic Attributes. *Proceedings of AAAI'94 Workshop on CBR*, pp.123–127, 1994.