関係データベースを使った事例ベース検索 (2) − 実用システム*

6H−6

柴田晃宏‡　島津秀雄‡　北野宏明§　佐藤亜津美¶　梶原寿一郎¶†

日本電気（株）C＆C情報研究所‡
日本電気（株）ソフトウェア生産技術開発研究所§
日本電気（株）ソフトウェア企画室¶

1　はじめに

本稿では、関係データベースを使った事例ベース検索 [5] の実用システムの開発例について述べる。開発中の実用システムは、社内ソフトウェアの品質向上・生産性向上事例を対象として組織的に進められてきた活動事例の類似事例検索システム [3][4] である。同システムは、従来事例ベースと類似検索処理の部分を独自に作成していたが、[5] に基づいて事例ベース検索シェル [1] の再構築を行なったことに従って、事例ベースを商用関係データベースシステム上に移植し、[5] に基づく類似事例検索機能を実現した。商用の関係データベースシステムは、事例件数の増大に耐える検索能力、ネットワーク対応機能を備えているため、現実の運用に耐えるシステムを構築する上で大きなメリットがある。本稿では、[5] で提案した事例検索手法に基づいて構築した実用レベルのシステムが、関係データベースの利点を損なわず、十分な検索性能を発揮し得ることを示し、更に、ユーザのアクセス管理、複数ユーザーに対するデータベースの集中管理等の副次的なメリットについても述べる。

2　実用システムの概要

当社では、ソフトウェアの品質向上・生産性向上のためのQC活動として、SWQC(SoftWare Quality Control)を行なっており、活動の事例は、全社を対象として定期的に論文報告されている。我々はこれらの事例に基づいて品質向上・生産性向上のためのアドバイスを行なうSQUAD(Software Quality Control Advisor)システムを提案し、構築をすすめている [3]。ソフトウェア構築上の問題を抱えた利用者が、自分の問題を幾つかの属性の値の集合の形でシステムに提示すると、SQUADシステムが、その問題に類似した事例を事例ベースから検索し、類似度の高い順に類似事例を利用者に提示するというものである。SQUADシステムでは、従来の事例ベース推論研究と異なり、事例の適応フェーズは持たない。これは、実用システムでは、適応フェーズの実現は困難であるし、本応用の場合、事例の提示で十分であるという理由によっている。

3　新しいSQUADシステムの構成

図2に新しいSQUADシステムの構成を示す。事例ベースは、商用関係データベースシステム上に構築されている。1つの事例は、約50個の属性の値の集合で表現される。個々の属性は、関係データベースの1カラムに相当し、1つの事例は関係データベースの1つのレコードの形で表現されている。利用者がメニューインタフェースを通して、自分の問題をシステムに入力すると、システムはその問題を解釈し、複数のSQL検索式に変換し、それらSQL検索式を順々に実行していく。それらSQL検索式の結果が与えられた問題の類似事例として利用者に提示される。

検索条件は、各々の属性に関して、種々な抽象度で指定することが可能である。属性ごとの類似度は、属性ごとに定義された類似度定義記述を使って算出される。例えば、(開発言語=C, 開発マシン=EWS4800, 開発OS=SVR4) に対して、(開発言語=C++, 開発マシン=MIPS, 開発OS=BSD4.3) は図1と式1に従い、類似度が

(0.3*0.7+0.4*0.8+0.3*0.6)/(0.3+0.4+0.3)=0.71

と計算される。

$$\text{問題と事例の間の類似度} = \frac{\sum_{i=1}^{n} \text{属性 i の重み} \times \text{問題と事例の属性 i での類似度}}{\sum_{i=1}^{n} \text{属性 i の重み}}$$

式 1: 類似度計算式

従来のSQUADシステムでは、与えられた検索条件に対し、すべての事例を照合させて類似度の計算を行ない、類似度の高い順に事例を選び出すようにしていた。しかし事例ベースが大規模になってくると、検索速度向上のために、事例ベースの構造や、検索アルゴリズムを工夫する必要がでてくる。商用関係データベースには、様々な高速化が施されており、実用面で魅力がある。しかしながら、関係データベースの検索言語SQLには、類似検索の概念がなく、そのままではつかえない。そこで [5] で述べたように、類似検索と等価なSQL式を生成する機構の付加によって、関係データベース上で類似検索を可能にした(図2)。

4　実験と評価

関係データベースを使った事例ベース検索システムにおいて重要なことは実行速度である。利用者の1つの問い合わせを受け取ると、それを解釈して複数のSQL検索式に展開し、それらSQL検索式を実行する、という2段階の処理に分けることが出来るが、処理の大きさは、解釈・展開部分に比べてSQL

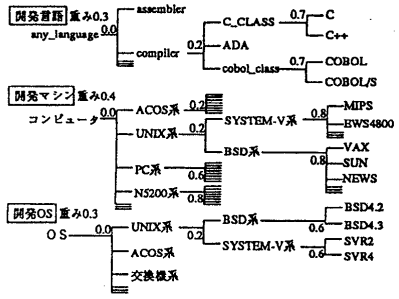

図 1: 類似度定義

* Case-base Retrieval Using Standard Relational Database (2) - Real System

†Akihiro SHIBATA, Hideo SHIMAZU , Hiroaki KITANO, Atsumi SATO and Juichiro KAJIHARA

‡C & C Information Technology Reseach Labs., NEC Corp.

§Software Engineering Development Lab., NEC Corp.

¶Software Planning Office, NEC Corp.

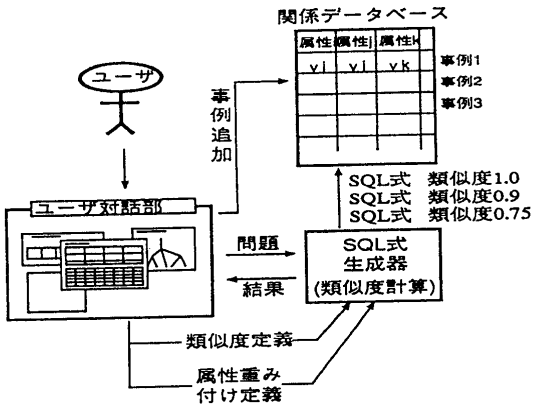

図 2: システム構成

検索式実行部分が圧倒的に大きいので、この手法の実行速度は、商用データベースの実行速度に完全に依存する。そこで現在、実用レベルの事例を持った事例検索システムで、実用に十分な性能が出るかどうかの評価を行なっている。但し、既に述べたように本手法の実行速度は、商用データベースの実行速度に完全に依存するが、商用データベースが提供している実行最適化の技法を完全に利用しているのではないので、ここでの評価は最終的なものではない。

検索性能測定のために 25、50、100、200、400、800、1600 件の事例レコードを持つテーブルを関係データベース中に作成した。与えられた問題を (開発言語 =ADA, 開発マシン =VAX) という条件にすると、以下の 4 個の SQL 式が生成される。

(a) 類似度 1.000 の式:
select * from 事例テーブル where (開発言語 ='ADA' and (開発マシン ='VAX'))
(b) 類似度 0.778 の式:
select * from 事例テーブル where (開発言語 ='ADA' and (開発マシン ='NEWS'))
(c) 類似度 0.644 の式:
select * from 事例テーブル where (開発言語 ='ADA' and (開発マシン in ('EWS4800','EWS4800/220', 'EWS4800/30','EWS4800/35','MIPS','SUN','SUN3', 'SUN4','SRARC','UP4800')))
(d) 類似度 0.556 の式:
select * from 事例テーブル where (開発言語 in ('HPL','SYSL','C','C++','COBOL','COBOL/S','IDL', 'FORTRAN','FORTRAN/S','BASIC','LISP','CHILL', 'S言語','PLC','PL/M') and (開発マシン ='VAX'))

これらの検索式は、類似度の高い順にデータベースに渡され実行される。実際に SQL 式生成器が生成する検索式の組合せ数の上限は

$$\prod_i^{(\text{指定条件数})}(\text{属性}i\text{の階層数})$$

であるが、通常これらすべてを検索する必要はなく、類似度の大きい検索式のみに限定して検索することになる。
図 3に、4 つの検索式の実行速度を示す。図 3の横軸は、格納レコード数の異なるテーブルであり、縦軸は、検索式の実行時間である。検索式の実行速度は、検索式の種類、テーブルの内容、ヒットしたレコード数、データベースインデックスの効果、等様々な原因に依存するので、図 3は普遍的なものではないが、一般的な利用者の類似検索要求に対する応答速度の目安

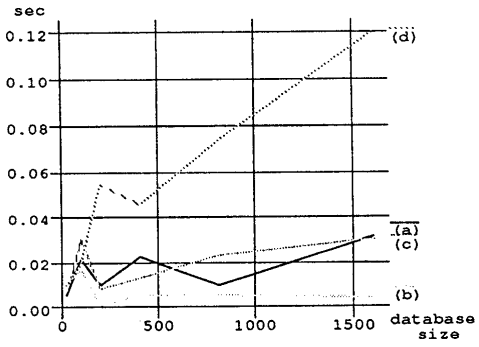

図 3: 事例ベース規模−検索時間

にはなる。仮に利用者が、類似度 0.5 以上の事例をすべて手に入れたいときには、これら個別の検索式実行時間を総和したものが応答速度になるが、それでも実用に耐え得る範囲と思われる。また、類似度の上位 10 件の事例が欲しいなどという場合には、類似度の高い検索式から実行していき、検索レコードの合計が 10 件を越えた所で、それ以降の検索式実行を中断することも可能なので、応答速度はさらに早くなる。

実用レベルの事例検索システムを構築する上で、商用の関係データベースシステムに備わっているさまざまな機能が有効に利用できる。例えば、利用者の認証機構や、データへのアクセス制限、ネットワークを介したデータベースアクセス等の機能がある。データの機密性に応じたアクセス制限は、データベースを分割し、それぞれにアクセス権限を設定することで容易に実現できる。また、ネットワークを介して遠隔地から事例ベースを検索するという使い方も、商用関係データベースがサポートしているネットワーク透明化機能を用いて労せず実現できる。

5 おわりに

本論文では、関係データベースを使った事例ベース検索の応用システム実現例について述べた。ソフトウェアの品質向上・生産性向上事例を対象に構築したシステム上で性能評価を行なったところ、大規模事例の類似検索において十分な検索能力が示せた。

参考文献

[1] 島津ほか, "事例ベース推論システム構築用ツールの開発", 人工知能学会, 研究会,SIG-KBS-9102-7,1991 年 9 月.

[2] 柴田ほか, "事例ベース検索システム構築用ツールの開発", 情処 43 全大,1991.

[3] 北野ほか, "*SQUAD* : 事例検索システム", 情処 44 全大,1992.

[4] H.Kitano,A.Shibata,H.Shimazu,J.Kajihara,A.Sato:"Buildin Large-Scale and Corporate-Wide Case-Based Systems:Integration of Organizational and Machine Excutable Algorithms",AAAI-92.

[5] 島津ほか, "関係データベースを使った事例ベース検索 *(1)* - アルゴリズム", 情処 45 全大,1992