文書類似度に基づくパターン間関連解析手法

久保淳人[†] 鷲崎弘宜[††]
高須淳宏[††] 深澤良彰[†]

本稿では，計算機を用いてパターン文書からパターンの情報を抽出し，パターン間の関連を自動的に分析する手法を提案する．

Relation Analysis Technique for Patterns Based on Document Similarity

Atsuto Kubo,[†] Hironori Washizaki,[††] Atsuhiro Takasu[††]
and Yoshiaki Fukazawa[†]

In this paper, we propose a technique for extracting information of patterns from pattern documents, and a technique for identifying the appropriate relations between patterns automatically.

1. はじめに

Alexander らは，ある文脈における問題の解決策をパターンと定義し[1)]，Gamma らはパターンを用いてソフトウェア開発における知識を記述した[2)]．ソフトウェアパターンとは，ソフトウェア開発において繰り返し発生する問題について，解決されるべき問題，制約条件 (フォース) の検討，および，解法を記述したものである．ソフトウェアパターンを用いることで，問題解決に関する知識を効率的に共有および再利用できる．World Wide Web(WWW) における代表的なソフトウェアパターンリポジトリである 3) では 2004 年 12 月時点で 700 個以上のソフトウェアパターンが公開されている．

ソフトウェア開発においては，柔軟性と性能のトレードオフなど，制約条件の違いにより異なる解決策をとることがあるが，その場合，いくつかの解決策について，それぞれの制約条件や結果を比較する必要がある．また，特定の分析モデルに対応する設計が存在する場合 (例:組織階層のモデルを Composite パターン[2)] で設計する) 等では，開発フェーズを横断して知識を関連づける必要がある．このとき，知識を記述する単位としてパターンを利用し，パターン間の関連を明らかにすることで，組み合わせたパターンの利用を円滑に行うことができる．

これまでに，パターン間関連分析の試みはいくつか報告されている[4)5)] が，いずれも小規模である．パターン間関連分析は人的コストが高いので，多数のパターン間関連をすべて人手で分析するのは困難であり，計算機による支援が必要になる．そこで，本稿では，計算機を用いてパターン間関連を自動的に分析する手法を提案する．提案手法では，パターンを記述した文書からパターンの要素を文書片として抽出し，文書片間の類似度からパターン間関連を得る．副次的に，提案手法を実装したシステムによる，人手による分析では気付かなかったパターン間関連に関する示唆が期待される．

2. 提案手法

パターンを記述した文書がパターン文書であり，一連のパターン文書をまとめて閲覧できるようにしたものがパターンカタログである．パターン形式は，パターン文書の書式とパターン文書に記述される情報を規定する．

本稿では，パターン適用を文脈の遷移と捉え，パターン適用前後の文脈をそれぞれ開始文脈，結果文脈とした．このとき，パターンは開始分脈を始点，結果文脈を終点とするラベル付き有向グラフ (**図 1**) となる．なお，本稿のパターンモデルにおける開始文脈および結果文脈は問題も含むものとする．

† 早稲田大学
Waseda University
†† 国立情報学研究所
National Institute of Informatics

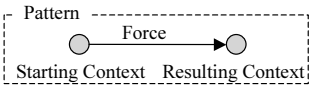

図 1 パターンモデル

提案手法の概要を**図 2** に示す．WWW 等からあらかじめ取得したパターン文書群を入力とする．HTML 解析部では，有名なパターン形式に従ったパターン文書の構造がほぼ共通である[6] ことを利用し，HTML タグの出現で状態遷移する状態機械を用いた HTML 解析器を使って見出しと本文の組を抽出する．パターン形式判定部では，パターン文書の見出しの種類がパターン形式に依存することを利用し，見出し集合として記述したパターン形式と，入力パターン文書の見出し集合との適合度を測定することで，パターン文書が従うパターン形式を判定する．パターン文書 d の見出し集合を $h(d)$ とし，パターン形式 f を $f = \{h_1, h_2, \ldots, h_m\}$ とするとき，適合度は

$$fitness(d, f) = \frac{h(d) \cap f}{h(d) \cup f}.$$

パターン抽出部では，上述のパターンモデルの要素と見出しとの対応表を用いて，モデルの各要素に対応する文書片を得る．パターン間関連分析部では，パターン要素の文書片間の類似度からパターン間関連を得る．パターンモデルの要素である各文書片を TF-IDF 法[7] による単語重みのベクトルを用いて表現し，ベクトルの偏角の余弦値を類似度とする．$tf(t,d)$ を文書 t 中の単語 d の出現回数，$df(t)$ を単語 t が出現する文書数，パターン文書数を N とすれば，

$$tfidf(t,d) = tf(t,d)(\log \frac{3N}{df(t)} + 1).$$

文書片の各組み合わせについて，以下のパターン間関連が対応する．

開始文脈同士の類似: 開始文脈同士が類似する 2 つのパターンは，同じ問題に対する異なった解法を与える．

結果文脈同士の類似: 結果文脈同士が類似する 2 つのパターンは，類似した適用結果をもたらす．

結果文脈と開始文脈の類似: あるパターン p_1 の結果文脈と，別のパターン p_2 の開始文脈が類似する場合，p_1 の適用後に連続して p_2 を適用可能である．

リポジトリに格納されたパターンおよびパターン間関連は，**図 3** のようなパターン関連グラフ[8] や，検索システムを用いて利用者に提供される．

3. 実　　験

提案手法を実装したシステムは，異なる作者がそれぞれ Gamma らによるデザインパターン[2] を記述した 2 つのパターン文書群について，パターン文書群中同一のデザインパターンを記述したパターン文書の組を，11 点平均精度[7]0.789 で正しく検出した．また，分析の結果図 3 のパターン関連グラフを得た．

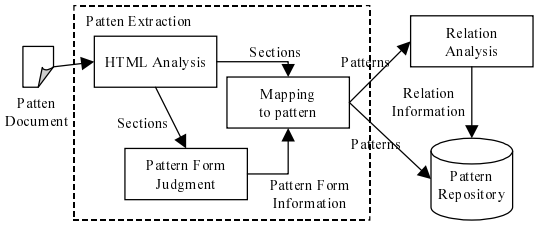

図 2 手法の概要

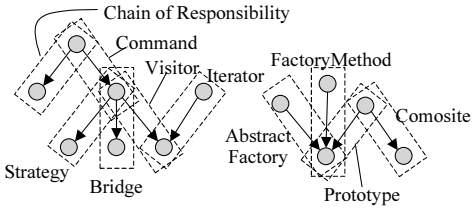

図 3 パターン関連グラフの例

4. 議　　論

提案手法について，特に以下の点で議論を行いたい: フォースを用いたパターン間関連分析，パターン作者毎の表現の異なりがもたらす精度低下への対処，日本語で書かれたパターン間の関連分析．

参 考 文 献

1) Alexander, C. et al.: *A Pattern Language*, Oxford University Press (1977).
2) Gamma, E. et al.: *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley (1994).
3) Cunningham & Cunningham, Inc.: Portland Pattern Repository. http://c2.com/ppr/.
4) Zimmer, W.: *Relationships between Design Patterns*, Pattern Languages of Program Design, Vol.1, Addison-Wesley (1995).
5) Ong, H.-Y. et al.: Rewriting a Pattern Language to Make it More Expressive, ChiliPLoP2003 (2003).
6) 鈴木純一ほか: ソフトウェアパターン再考, 日科技連 (2000).
7) 徳永健伸: 情報検索と言語処理, 東京大学出版会 (1999).
8) 久保淳人, 鷲崎弘宜, 高須淳宏, 深澤良彰: 文書中のパターン間の文書類似度による関連分析, *DBSJ Letters*, Vol. 3, No. 3 (2004).