相関ルールを用いたシステム障害対応データの傾向分析

十九川 博幸† 森崎 修司‡, 松村 知子‡, 門田 暁人‡, 松本 健一‡

株式会社日立システムアンドサービス 生産技術部†
国立大学法人奈良先端科学技術大学院大学 情報科学研究科‡

1. はじめに

ソフトウェアの社会への浸透に伴い，リリース時点での高信頼化，高品質化だけでなく，リリース後，サービスイン後に発生した問題（障害）の早期復旧，早期解決も求められている．リリース後に発生した障害の属性や対応期間の記録（障害対応データ）をもとに検討することにより，障害発生時の迅速な問題切り分け手順，対応工数の見積り，対応プロセスの改善の手がかりとできる．

障害対応データの分析は全体的な傾向の把握（平均対応日数や平均対応工数）と特定の障害の属性に共通する傾向の分析に大別できる．後者に着目することにより，共通した属性をもつ障害対応の改善に効果的な対策を練ることができる．しかしながら，全体的な傾向の把握と比較して特定の障害の属性に共通する傾向の分析は手がかりが少なく，データを眺めるだけでは分析が進まない．本稿ではそのような傾向分析を目的として，障害対応データを対象に相関ルール分析[1]を適用する．具体的には，株式会社日立システムアンドサービスで蓄積された障害対応データを対象として相関ルールを抽出する．結論部に障害の重要度をもつもの，対応日数をもつもの，に着目し，対象データに知識を持つ作業者と研究グループとで特徴的なルールの選出を行い，考察した．

2. 相関ルール分析

相関ルール分析は対象データに含まれる規則性を抽出することを目的としたデータマイニングアルゴリズムの1つであり，Agrawalらが提案したものである[1]．具体的には，出現頻度などの与えられた指標値を満たす規則性を「X ⇒ Y」の形式で全て抽出する．X を前提部，Y を結論部と呼び，X にあてはまる対象トランザクション(データベースの1エントリ)が同時に Y もあてはまることを表す．

X，Y は複数の項目を含むことができ，それぞれ「かつ("&")」で結合することができる．たとえば「(重要度=大) & (設計部門=A) ⇒ (対策所要日数=長)」という相関ルールから，重要度が大きく，かつ，A 設計部門で設計されたシステムで起きた障害の対策所要日数が長い障害が存在することを知ることができる．

ここで対象となる障害対応データ(データベース)を D，データベース中の個々の障害(トランザクション)を T とする．相関ルールの指標値として以下がある[1]．

支持度

対象データが相関ルールを満たす割合（出現頻度）であり，$support(X \Rightarrow Y)$と表記し，$support(X \Rightarrow Y) = s/n$ である．ここで，

$$s = \left|\{T \in D \mid X \subset T \cap Y \subset T\}\right|,\ n = \left|\{T \in D\}\right|$$

信頼度

前提部Xが満たされたときに同時に結論部Yも満たされる割合であり，$confidence(X \Rightarrow Y)$ と表記し，$confidence(X \Rightarrow Y) = s/y$ である．

ここで，$y = \left|\{T \in D \mid X \subset T\}\right|$

表1 障害の属性

属性名	説明
発生日	障害発生年月日
対策日	障害対策年月日
対策所要日数	障害発生から対策までの日数
顧客名	顧客名称
システム	システム名称
バージョン	障害が発生した，システム／プロダクトのバージョン番号
現象コード	全面停止や結果不正などの，現象を示すコード
重要度	障害の重要度を大／中／小で区別する
設計本部/設計部	障害が起きたシステムを開発した部門の名称
回答内訳コード	ソフト不良，SE 作業ミスやドキュメント不良など，障害の内訳を示すコード

An Analysis of System Issue Tracking Data by Association Rule Mining
† Hiroyuki Tokugawa, Hitachi Systems & Services, Ltd.
‡ Shuji Morisaki, Tomoko Matsumura, Akito Monden, Kenichi Matsumoto, Graduate School of Information Science, Nara Institute of Science and Technology

リフト値

前提部Xによりどのくらい結論部Yが満たされやすくなっているかを示しており，$lift(X \Rightarrow Y)$と表記し，$lift(X \Rightarrow Y) = confidence(X \Rightarrow Y)\, z/n$である．ここで，$z = \left|\{T \in D | Y \subset T\}\right|$.

本稿では，同一の結論部を持つものに対して，同一の結論部を持ち，1箇所だけ異なるものについて，"|"(論理和)で相関ルールを結合する．たとえば「$(A = a_1)$ & $(B=b_1) \Rightarrow (C=c_1)$」と「$(A = a_2)$ & $(B=b_1) \Rightarrow (C=c_1)$」とを「$(A = a_1 | a_2)$ & $(B=b_1) \Rightarrow (C=c_1)$」とする．

3. 障害対応データ

対象となる障害対応データは株式会社日立システムアンドサービスで 1999 年から 2006 年までに蓄積された 601 件の障害対応データである．各障害は表 1 に示すような属性を持ち，対策所要日数以外はカテゴリ変数である．実際に集められた障害対応データには，自由記述の"現象"，"原因"，"発生条件"等が含まれるが，相関ルール分析の対象としていないので，表 1 には含めていない．

対策所要日数は障害が起きてから解決されるまでの日数である．本稿では対策所要日数を"即日"(1 日以内で解決)，"3 日以内"(2 日以上 3 日以内で解決)，"10 日以内"(4 日以上 10 日以内で解決)，"長期"(11 日以上)の 4 つに分割した．

表 2 重要度を結論部とするルール例

番号	ルール	支持度	信頼度	リフト値
1-1	(顧客名＝顧客 1)＆ (システム＝Xシステム)＆ (設計本部＝P 本部) ⇒(重要度＝中)	0.383	0.861	1.399
1-2	(設計本部＝Q 本部) ⇒(重要度＝大)	0.062	0.822	3.088
1-3	(設計本部＝R 本部) ⇒(重要度＝中)	0.052	0.705	1.144

表 3 対策所要日数を結論部とするルール例

番号	ルール	支持度	信頼度	リフト値
2-1	(設計部＝C 部署) ⇒(対策所要日数＝ 3 日以内)	0.105	0.625	1.183
2-2	(回答内訳＝SE作業ミス) ⇒(対策所要日数＝即日)	0.045	0.467	1.811
2-3	(設計部＝C 部署)＆ (回答内訳＝Program 不良) ⇒(対策所要日数＝即日)	0.035	0.367	1.423
2-4	(設計部＝A 部署) ⇒(対策所要日数＝長期)	0.031	0.760	2.776

4. 結果

最低支持度を 3%に設定し相関ルールを抽出した．得られたルールのうち"重要度"と"対策所要日数"に着目し，結論部にこれら 2 つを含むものを選択し，対象の障害対応データに詳しいメンバで検討した結果，表 2, 3 に示すようなルールに対して現実をあらわしていると判断された．表 2 は結論部を障害の重要度としたものであり，表 3 は結論部を対策所要日数としたものである．

- 特定部署で開発され，特定顧客に納入された特定システムで重要度"中"の障害が多い．(1-1)
- 特定部署で重要度の大きな障害が多い．(1-2)
- 特定部署で設計されたシステムでの発生障害は 3 日以内に対策される傾向にあるものの(2-1)，特定部署が設計したシステムの障害の対策所要日数が長期化するものがある．(2-4)
- 特定の回答内訳コードの障害は，即日解決されるものがある．(2-2, 2-3)

5. おわりに

リリース後に発生した問題の記録(障害対応データ)を対象として，相関ルール分析を適用し，得られた相関ルールから知見を得た．得られた知見には相関ルールを見るまで明らかでなかったもの，相関ルールを見る前から知っていたが，出現頻度(支持度)等の指標値については明らかでなかったものがあった．それらのルールを対策日数の見積りの根拠としたり，プロセス改善のきっかけとしたりできる．今後は，相関ルールをきっかけとする具体的なデータ分析手順を検討する予定である．

謝辞

本研究の一部は、文部科学省「eSociety 基盤ソフトウェアの総合開発」の委託に基づいて行われた．

参考文献

[1] R. Agrawal, T. Imielinski and A. Swami: Mining Association Rules between Sets of Items in Large Databases, In Proceedings of ACM SIGMOD Conference on Management of Data, pp. 207-216 (1993)