워크플로우 마이닝 기법(휴리스틱접근)

이명희° 장영원 유철중 장옥배
전북기능대학, 전북대학교
leemh@kopo.ac.kr°{ywchang, cjyoo, okjang}@chonbuk.ac.kr

Workflow Mining Technique(Heuristic Appproach)

Myoung-Hee Lee° , Young-Won Chang, Cheol-jung Yoo, Ok-bae Jang
Korea Foundation of Polytechnic Colleges ,Chonbuk National University

요 약

최근들어 기업의 업무가 더욱 전문화되고 복잡해짐에 따라 워크플로우 시스템도 복잡해지고 다양해 지고 있다. 이러한 문제로 인하여 실제 필요로 하는 프로세스의 관리 및 도출이 요구된다. 본 논문에서는 영향력있는 프로세스를 도출하고 지원하기 위한 워크플로우 마이닝에 관하여 분석한 후 분석을 바탕으로 상관관계분석과 주성분분석을 통하여 워크플로우를 보다 효율적으로 관리할 수 있는 마이닝 규칙을 제시한다.

1. 서 론

본 논문에서는 워크플로우의 로그파일을 이용하여 보다 효율적인 프로세스를 도출하고 지원하기 위한 워크플로우 마이닝에 관하여 분석한다. 이러한 분석을 바탕으로 마이닝 알고리즘을 적용하여 워크플로우를 보다 효율적으로 관리할 수 있도록 규칙을 제시한다. 마이닝 규칙에 따라 마이닝을 한후, 얻을수 있는 가시성과 업무간 효율성을 모색 한후 한 예로 기업의 워크플로우 마이닝에 대해 설명하고 분석한다.

2. 관련연구

2.1 워크플로우

WMC에서 정하고 있는 워크플로우는 비즈니스 프로세스를 전체 또는 부분적으로 컴퓨터에 이롭게 하거나 자동화한 것을 의미한다. 즉 워크플로우는 비즈니스 프로세스의 자동화를 통해 정확하면서도 신속한 업무 처리를 지원하는 정보기술이다[3].

2.2 워크플로우 마이닝

워크플로우 마이닝이란 워크플로우의 모니터링을 이용하여 워크플로우 시스템의 프로세스의 경향을 분석하여 또 다른 정보를 제공하는 기법을 의미한다. 즉, 워크플로우 마이닝은 업무 프로세스의 분석을 통하여 기업이나 고객에게 또 다른 정보를 제공할 수 있어야 한다[4].

2.3 데이터 마이닝 알고리즘

데이터 마이닝이란 방대한 데이터 속에 내재된 의미있는 상관관계, 패턴, 경향 등을 찾아내는 일련의 프로세스로 통계 및 패턴인식, 신경망 등 여러 기법이 사용되고 있다.

2.4 상관관계분석

상관 관계 기능을 사용하여, 데이터 영역내의 모든 필드(범주형 데이터 필드 제외)에 대해 서로간의 상관관계를 나타낸다.

2.5 PCA(Principal Component Analysis) 방법

주성분분석(principal component analysis)은 차원축소를 통하여 저 차원상에서 변수의 관계를 규명하는 다변량 자료분석기법이다.

3. 연구목적 및 방법

본 연구의 목적은 여러 가지 마이닝 접근 중 통계분석과 휴리스틱 접근을 이용하여 보다 효율적으로 최적화할 수 있는 매우 영향력 있는 프로세스를 도출해내는 것이다.
전체적인 순서는 다음과 같이 4단계로 구성된다.

1단계 : 데이터 필터링 및 전처리
2단계 : 순서정의
- *연속형 데이터 통계분석*
- *주성분 분석*
- *그룹데이터를 지정하여 그룹 간 분석(그룹설정)*
- *상관관계분석*
- *상관계수의 설정*
- *상관관계와 주성분분석을 이용한 데이터 분석*

3단계 : 규칙정의
4단계 : 사례연구

4. 워크플로우 마이닝의 휴리스틱 접근

워크플로우 마이닝 방법은 기본적인 데이터로 만들어 활용하기 위한 단계인 데이터필터링 및 전처리, 주성분 분석과 상관관계 분석을 이용한 데이터분석 규칙 정의, PCA 휴리스틱 적용 사례에 대한 연구 그리고 마지막으로 성능평가와 같은 순서로 진행된다.

4.1 데이터 필터링 및 전처리 과정

PCA(Principal Component Analysis) 분석에서 필요가 없는 node를 제거한다.
'Type node'에는 PCA분석이 독립변수(X)만으로 분석을 하게 되므로 모든 변수들을 독립변수와 연속형 변수로 지정하여 준다.

4.2 순서 정의

순서는 다음과 같이 진행된다.

4.2.1 연속형 데이터 통계분석

연속형 데이터 통계분석은 변수명과 변수형태 그리고 변수의 입출력 형태에 관해 정의한다. PCA분석이 독립변수(X)만으로 분석을 하게 되므로 변수명을 정하고 변수형태는 연속형 변수의 형태로 하고 입출력의 형태는 각각 독립변수로 설정한다.

4.2.2 주성분 분석

주성분 분석은 데이터를 몇 개의 주 성분으로 나눌지를 결정한다. 일반적으로 주성분의 수는 Eigen value(고유값)에 따라 정하고 Eigen Value(고유치)가 급강하를 보이고 난 후의 Eigen Value에 해당하는 인자 수를 선택한다.

4.2.3 그룹데이터를 지정하여 그룹 간 분석(그룹설정)

앞서 설정한 주 성분에 대하여 그룹데이터를 지정하여 상호간 관계를 분석한다.

4.2.4 상관관계분석

데이터 영역내의 모든 필드(범주형 데이터 필드 제외)에 대해 서로간의 상관관계를 볼 수 있다.

4.2.5 상관계수의 설정

상관계수의 설정은 0부터 1.0 사이로 설정하고 상관관계를 파악한다(0.0 <= p < 1.0).

4.2.6 상관관계와 주성분분석을 이용한 데이터 분석

주성분 분석의 그룹간 분석을 통한 contribute table과 상관관계를 이용한 correlation table을 이용하여 최적의 프로세스를 마이닝 한다.

4.3 규칙 정의

상관관계와 주성분분석을 이용한 데이터분석을 위한 알고리즘은 다음과 같다.

input : workflow event log
output : principal Component process

Rule 1. Definition Correlation table
Given Process A
IF(p > 0.9) THEN process candidate

Rule 2. Definition Contribution table
Given Process A
For each Principal Component and Group
(p > 0.8) THEN Process CANDIDATE
p is contribution value

Rule 3. Given ProcessA
For I = 0 To PrincipalComponentNumber{
For I = 0 To GroupNumber{
IF (p > 0.9) && (count > 5){
SELECT PROCESS
}
}}

p: 상관계수

4.4 사례연구

4.4.1. 도메인 선정

1) 실험데이터는 어느 생산공장의 54개의 조업 공정의 프로세스 데이터이다.

2) 비즈니스 프로세스간 편차를 분석한다.

3) 보다 영향력 있는 프로세스를 도출한다.

4.4.2. 가정 및 제약사항

가정 및 제약사항은 다음과 같다.

1) 데이터는 연속형 데이터이어야 한다.

2) 각각의 프로세스는 독립적이다.

3) 프로세스의 크기는 54개, 데이터의 수는 7596개이다.

4) 로그 데이터는 노이즈가 없다.

4.4.3 데이터분석

1) 연속형 데이터 통계분석

연속형 데이터 통계분석을 통해 평균과 분산, 표준편차, 첨도, 왜도에 대해 분석한다.

2) 주성분 분석

주성분의 개수를 정할 때 Eigen value를 큰 순서 대로 나타낸 막대그래프로 어느 Eigen Value(고유치)가 급강하를 보이고 난 후의 Eigen Value에 해당하는 인자 수 3개를 선택한다.

3) 그룹데이터를 지정하여 그룹간 분석(그룹설정)

4) 상관관계분석(p = 0)

5) 상관계수의 설정(p = 0.9)

변수명	A6	A7	A27	A31	A33
A6	1	0.902105		0.472261	
A7	0.902105		1	0.575979	
A27	0.472261		0.575979		1
A31	-0.17335		-0.203833		0.024007
A33	-0.0244526		0.169024		0.247617
A41	0.0182834		0.0525681		0.376466
A42	-0.369152		-0.489974		-0.77461
A45	0.0112642		0.0795395		0.188311
A50	0.957044		0.869718		0.437217

correlation table (p = 0.9)

6) 상관관계와 주 성분분석을 이용한 데이터 분석

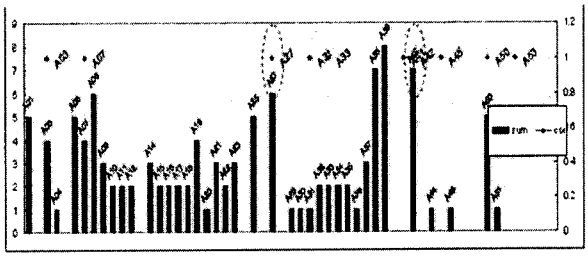
위 Rule 1.을 사용하여 나타난 프로세스는 다음과 같다.

A6, A7, A27, A31, A33, A41, A42, A45, A50, A53

위 Rule 3. 을 사용하여 작성한 테이블은 다음과 같다.

Relation	g1-g2			g1-g3			g2-g3			sum	Correlation
Group	g1	g2	g3	g1	g2	g3	g1	g2	g3	n>0.8	p>0.9
Name	1	2	3	4	5	6	7	8	9		
A01	1	1	1			1			1	5	
A02											
A03	1		1			1			1	4	1
A04				1						1	
A05											
A06	1	1	1			1			1	5	
A07	1		1				1		1	4	1
A08	1	1		1	1		1	1		6	
A09						1		1	1	3	
:				1	1					2	

다음은 위 Rule 3을 이용하여 만든 표를 이용하여 그린 그래프이다.



5. 결론 및 향후연구

본 논문에서는 워크플로우 마이닝을 위한 방법으로 규칙에 따라 적용할 수 있는 휴리스틱 접근을 사용하였다. 상관관계분석과 PCA를 통해 보다 영향력 있고 효율적으로 프로세스를 마이닝 하였다.

향후에는 실제 필드에서 적용되는 상세 비즈니스 데이터를 이용하여 마이닝 하는 것이 더 바람직하다고 생각되며 휴리스틱 적용을 위해 좀더 정형화된 규칙 및 알고리즘에 대한 연구가 필요하다고 본다

참 고 문 헌

[1] F. Casati, "Workflow Evolution. Data and Knowledge Engineering," 24(3):211-238, 1998.

[2] Bussler, C., "B2B Protocol Standards and their Role in Sementic B2B Integration Engines," March 2002, Vol.24, No.1. IEEE Computer Society.

[3] W.M.P van der Aalst, "Process Mining : Discovering Workflow Models from Event-Based Data," BNAIC 2001, pp.283-290, 2001.