C4.5를 이용한 CMS 잡 오류 예측 모델

Zhenshun Xu, Shangsu Zuo, 최희수, 박대희†, 정용화, 조충호
고려대학교 컴퓨터정보학과
{jssoon77, mining2015, chlgmltn420, dhpark, ychungy, chcho}@korea.ac.kr

Prediction Model of CMS Job Failures using C4.5

Zhenshun Xu, Shangsu Zuo, Heesu Choi, Daihee Park, Yongwha Chung, Choong-ho Cho
Department of Computer & Information Science, Korea University

요 약

복잡한 그리드 컴퓨팅 환경에서 수행한 잡의 실패율을 낮추는 것은 그리드 환경의 효율성과 선순환을 위한 필수적인 요건이다. 본 논문에서는 데이터마이닝의 대표적인 방법인 결정트리의 C4.5 알고리즘을 이용하여 WLCG에서 수행한 CMS 잡 모니터링 결과에 대한 오류를 예측하는 모델을 설계하고 구현하였다. 제안한 예측 모델은, 1) CMS 대시보드에서 모니터링 결과 데이터를 추출하여 오라클 테이블에 로딩한다. 2) 결정트리인 C4.5 알고리즘을 기반으로 Oracle Data Miner에서 예측 모델링을 수행한다. 3) C4.5의 파라미터를 조절하여 적절한 예측결과 값을 도출한다.

1. 서론

스위스 제네바 근교의 유럽입자물리연구소(CERN) 기지에서는 대형강입자충돌기(LHC, Large Hadron Collider)의 실험용 검출기(CMS, ATLAS, ALICE, LHCb 등)가 존재하며, 해당 검출기들은 매년 약 25 페타 바이트의 정보를 수집한다[1-2]. WLCG(World LHC Computing Grid)는 세계 40여개 나라, 170여개 컴퓨터 센터의 리소스로 컴퓨팅 그리드환경을 구성하여 세계 각지의 연구자들에게 고에너지 물리 실험 환경을 제공한다[2]. LHC 실험중 하나인 CMS(Compact Muon Solenoid)에서는 하루에 65만 건의 CMS 잡(job)이 WLCG에서 스케쥴링 되며 그리드 환경에서 수행되는 잡을 모니터링 한다. 이러한 모니터링 결과는 해당 대시보드 데이터 저장소에 저장된다[2-3].

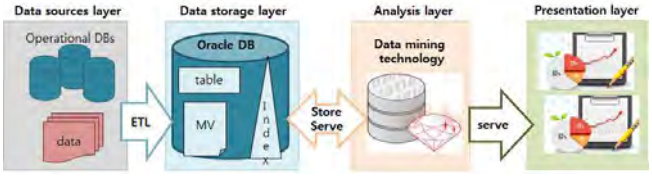
복잡한 그리드 컴퓨팅에서 오류를 핸들링 하는 것은 매우 어려운 작업이다[4-5]. 최근에는 이와 같은 오류의 실패 원인을 분석하기 위하여, 여러 가지 모니터링 도구들을 기반으로 오류코드와 연관된 실패한 잡들을 리포팅 하고 있다. Maier[4]등은 이와 같은 리포팅 정보를 이용하여 연관성 규칙 마이닝을 기반으로 오류코드와 정확한 고장원인에 대한 연관성을 분석하였으며, 그리드 워크로드에서 로그를 분석해서 잡 오류를 예측하는 연구도 발표되었다[5].

본 논문에서는 데이터 마이닝의 대표적인 예측기법인 결정트리의 C4.5 알고리즘을 이용하여 CMS의 잡 오류 발생 영향요인을 분석하고자 한다. 이는 추후 잡 오류 예방 대책에 대한 의사결정에 큰 도움이 될 것으로 판단한다. CMS 잡 모니터링 결과를 예측하기 위하여 Oracle Data Miner[8]에서 결정트리(C4.5) 알고리즘을 적용하여 예측모델을 생성하고 CMS 잡 모니터링 예측결과를 확인한다.

† 교신저자

2. CMS 잡 모니터링 오류 예측 시스템

본 논문에서 제안하는 CMS 잡 모니터링 오류 예측 시스템의 구조는 그림 1과 같다. 시스템은 크게 4개의 계층으로 구성된다. 첫째, 데이터 소스 계층에서는 CMS 대시보드의 상세페이지[3]에서 필요한 데이터를 추출한다. 둘째, 데이터 스토리지 계층에서는 추출한 데이터를 오라클 DB(DataBase)에 로딩 하고 C4.5알고리즘에 적용 가능하게 전처리 과정을 거친다. 셋째, 분석 계층에서는 데이터마이닝 도구인 Oracle Data Miner에서 C4.5알고리즘을 적용하여 예측 모델을 생성하고 특정 파라미터를 조절함으로 예측결과를 도출한다. 마지막으로, 표현 계층에서는 Oracle Data Miner의 모델보기와 결과보기를 통하여 예측모델과 예측결과를 확인할 수 있다.



(그림 1) CMS 잡 모니터링 오류 예측 시스템의 구조

2.1 데이터 수집 및 전처리

본 연구에서는 CMS 대시보드에서 2013년 7월부터 2014년 6월까지 1년 동안 수행한 잡 모니터링 결과 데이터를 샘플링한 약 30만 건의 데이터를 사용하였다[3].

CMS 잡 오류 발생 시 오류코드 예측을 위한 속성은 어플리케이션 버전, 제출도구버전, 분산엔진, 재시도 횟수, 수행 사이트, 제출시간, 제출화면버전, 수행상태, 오류코드 등을 포함한다. 우선 오라클 DB에 각각의 속성을 포함한 베이스 테이블을 생성하고 CMS 잡 모니터링 결과 데이터를 오라클DB의 해당 테이블에 로딩한 후, 그 입력변수들을 C4.5알고리즘 분석의 변수로 사용가능하게 전처리 과정을 거친다.

2.2 C4.5를 이용한 CMS 잡 오류 예측 모델

CMS 잡 오류 예측 모델링을 구현하기 위하여 Oracle SQL Developer의 확장패키지인 Oracle Data Miner를 사용한다[9]. 그림 2는 Oracle Data Miner에서 분류모델인 C4.5 알고리즘을 적용한 화면을 보여주고 있다.

(그림 2) Oracle Data Miner에서 예측모델을 적용한 화면

3. CMS 잡 오류 예측 실험 결과

본 연구의 실험에서는 CMS 잡 모니터링 결과에 관하여, 1) 성공 혹은 실패를 예측하거나, 2) 실패일 경우 오류코드를 예측하는 두 가지 예측모델링 실험을 수행하였다. 실험결과는 <표 1>과 같이 성공실패 여부의 예측결과 실험에서 예측신뢰도는 72%, 평균정확도는 86%를 보였으며 CMS 잡 오류발생 시 오류코드종류의 예측결과 실험에서 예측신뢰도는 76%, 평균정확도는 81%임을 확인하였다.

<표 1> 예측신뢰도와 평균정확도

예측모델	예측신뢰도	평균정확도
성공실패여부	72%	86%
실패오류코드	76%	81%

그림 3은 CMS 잡 모니터링 실패결과에 대한 오류코드종류를 예측한 실험결과의 왼쪽 트리를 보여주고 있다.

(그림 3) CMS 잡 오류코드 예측 결과 (왼쪽 트리)

그림 3의 트리에서 붉은색으로 표기한 경로에 관한 규칙을 확인해보면 <표 2>와 같다.

<표 2> 그림 3의 붉은색 경로에 대한 규칙

조건 & 속성	속성값 & 결과
IF ApplicationVersion is in	"CMSSW_4_2_8_patch7", "CMSSW_4_4_2_patch5", "CMSSW_5_3_11", "CMSSW_5_3_12_patch3", "CMSSW_6_2_0_patch1" , "CMSSW_6_2_7"
And TargetCE is in	"10_Selected_SE", "12_Selected_SE", "14_Selected_SE", "16_Selected_SE", "3_Selected_SE"
And Tperiod is in	"0" , "1", "10", "13", "14", "16", "18", "19", "2", "22", "23", "3", "4", "5", "8", "9"
And Site is in	"T0_CH_CERN", "T1_DE_KIT", "T1_ES_PIC", "T1_IT_CNAF", "T1_UK_RAL", "T1_US_FNAL", "T2_AT_Vienna", "T2_BE_IIHE", "T2_CN_Beijing", "T2_DE_DESY", "T2_EE_Estonia"
Then	**Error E0**

4. 결론

본 논문에서는 CMS 잡 모니터링 데이터를 기반으로 C4.5 알고리즘을 이용한 CMS 잡 모니터링 결과에 대한 오류를 예측하는 모델을 제안하였다. 이러한 예측 모델링의 결과는 추후 그리드 시스템 관리자에게 실제 사이트 관리, SW 버전관리, 오류 예방 대책, 오류 발생 후 상황 대처를 위한 중요한 자료로 활용될 수 있으며 더 나아가서 그리드 환경의 생산성 제고에 기여 할 것으로 기대한다.

5. 감사의 글

이 논문은 BK 21 Plus와 정부(미래창조과학부)의 재원으로 한국연구재단 기초연구 실험데이터 글로벌 허브 구축사업(N-14-NM-IR06) 및 KISTI 위탁과제의 지원을 받아 수행된 연구임.

참고문헌

[1] CERN information, http://home.web.cern.ch/about
[2] WLCG information, http://wlcg-public.web.cern.ch
[3] CMS dashboard information, http://dashboard.cern.ch/cms/index.html
[4] G. Maier, D. van der Ster, D. Kranzlmuller. “Finding Associations in Grid Monitoring Data”, 2009 10th IEEE/ACM International Conference on Grid Computing. 2009. pp.89-96.
[5] H. Saadatfar, H. Fadishei, H. Deldari. "Predicting job failures in AuverGrid based on workload log analysis", New Generation Computing, Vol.30 No.1, 2012. pp.73-94.
[6] 김영진, 류정우, 송원문, 김명원,"의사결정트리를 이용한 날씨에 따른 화재발생 확률 예측모델", 정보과학회논문지:소프트웨어 및 응용, Vol.40 No.11, 2013.11, pp.705-715.
[7] 김무수, 박건우, 이상훈, "의사결정트리를 이용한 적 주타격 방향 분석", 한국정보과학회, 한국정보과학회 학술발표논문집, Vol.39 No.1(B), 2012.6, pp.66-68.
[8] J. Han, M. Kamber, J. Pei, “Data Mining: Concepts and Techniques”, 3rd Ed., Morgan Kaufmann.
[9] Oracle Data Miner, http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/dataminerworkflow-168677.html.