다각적 접근법에 의한 소프트웨어 평가 및 그 적용사례

권원일, 이상덕, 신석규
한국정보통신기술협회 (TTA)
e-mail : {wonil, sdlee, skshin}@tta.or.kr

Multilateral Approach for Software Evaluation

Wonil Kwon, Sang-duck Lee, Seok Kyoo Shin
TTA (Telecommunications Technology Association)
ITTL (IT Testing Laboratory)

요 약

소프트웨어가 매우 다양하고 측정하기 어렵다는 특성 때문에 객관적인 소프트웨어 평가를 위한 지속적인 연구가 이루어지고 있으며 논의의 대상이 되고 있다.
평가 대상 소프트웨어 사용자 설문 분석에 의한 평가, 표준 평가모듈에 근간한 평가, 사용 패턴에 근간한 평가를 병행하는 평가는 보다 객관성을 확보할 수 있다. 세 가지 각각의 평가가 국제표준인 ISO/IEC 9126 의 품질 특성에 기반하고 있다. 사용자 설문 분석에 의한 평가는 사용자 들의 설문을 통계 처리하여 도출하고, 표준 평가모듈에 근간한 평가는 국제표준의 내용을 구체화한 규격서를 포함하는 평가 모듈에 의한 소프트웨어 시험을 의미한다. 사용 패턴에 근간한 평가는 주로 사용하는 행태나 방법의 평가를 통해 이루어진다.

1. 서론

소프트웨어는 최종 제품은 물론 개발과정 단계의 제품도 가시적으로 볼 수 있는 형태가 아니어서 품질을 개선하는 것은 물론이고 그 기본이 되는 품질 측정 조차 쉽지 않은 일이다. 소프트웨어의 품질을 측정하기 위한 노력은 다양하게 이루어지고 있으나 소프트웨어의 종류가 다양하고 같은 종류의 소프트웨어도 제각기 다른 특성을 보여, 품질 측정 기법 연구가 소프트웨어 품질 보장에 확답을 주지 못하고 있는 실정이다. 이러한 이유로 소프트웨어 제품을 평가하는 것 보다는 개발 프로세스를 평가하고 심사하는 이론적, 실증적 연구가 상대적으로 많은 실정이다.

소프트웨어 제품에 대한 품질 시험· 평가에 있어 고려하여야 하고 해결하여야 하는 과제는 다수 존재한다. 우선 다수의 평가자가 관여되는 것이 일반적이고 이들의 품질이해와 중요도 인식이 다를 수 있다. 그리고, 측정항목이 다양하고 항목별로 다른 측정방법을 선택할 수 있어 측정값을 총합 (aggregation)하는데 문제점이 발생할 수 있다. 또한 모든 소프트웨어의 평가를 가능하게 하는 품질모형이 소프트웨어 분야에는 존재할 수 없다는 것과, 품질특성과 해당속성, 측정 메트릭의 선택들도 평가 대상과 환경에 따라 다양하다는 어려움이 있다. 이 가운데 국제표준에 근거할 경우, 객관적인 품질모형과 특성의 문제는 어느 정도 해결되지만, 이 역시 모든 소프트웨어를 다루지는 못할 뿐더러 나머지 과제들의 해결방법도 제시하지 있지는 않다.[7]

본 논문에서는 여러 해결과제 들을 고려하여 가능한 객관적으로 소프트웨어를 평가하기 위한 방안과 그 적용사례를 제시하고자 한다.

2. 기존 연구

대부분 S/W 제품 품질에 대한 연구는 특정 테스팅 기법의 적합성에 대한 내용이나, 특정 분야의 S/W 를 평가하는 방법에 대한 연구가 주류를 이루고 있다.

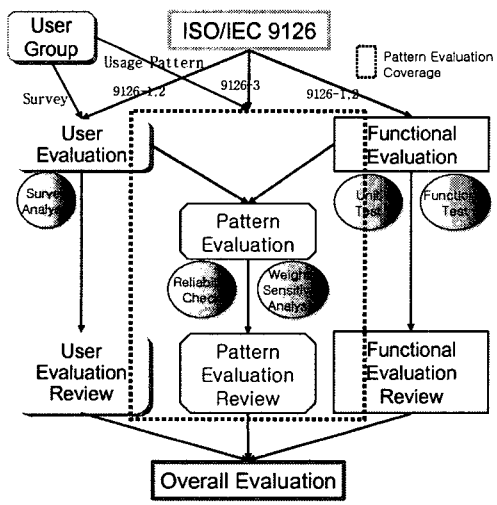
Emam (2001)은 S/W 품질 인증을 위한 패키지 S/W 제품 품질 평가에서는 사용자가 중심이 되어야 하는 시나리오 기반의 품질 평가가 필요하다고 주장한다. 이를 위해, 사용자가 중심이 되어 시나리오를 각 소프트웨어 분야 별로 만들어야 하며, 시나리오를 평가자가 평가하는데 있어서도 국제표준인 ISO/IEC 9126 (소프트웨어 품질 특성과 매트릭에 관한 국제 표준)의 사용품질 (Quality in Use)의 기준이나 표준 (criteria)을 사용하여 평가하여야 한다. 이때, 시나리오에 따라 중요도가 달라질 수 있고 이는 사용자에 따라 가변적이기 때문에 가중치는 범위값 (최소값, 최적값, 최고값)으로 주고 있다. 그 결과 최종 평가값도 일정 신뢰도 내에서의 범위값으로 주어진다. 또한 평가원간의 주관적인 판단을 최소화하고자 두명이나 두팀이 독립적으로 평가를 수행하고, 이를 통계적으로 처리하여 평가의 신뢰성 (reliability)를 측정하여 신뢰성이 통계적으로 보장되도록 한다. [5]

웹기반 S/W 의 시험 및 검증 기술 연구에서 웹기반 S/W 의 특성을 분석하고 이를 반영하기 위해 웹기반 S/W 를 구성하는 다양한 컴포넌트를 분리하여

단위테스트하고 이를 통합하여 시험 한다. 그리고, 시험 목적에 따라 다른 시험 기법을 적용한다. 웹기반 S/W 는 변동이 잦고 방대하게 분산되어 있으므로 리그레션 테스트와 링크오류를 찾기위한 정적분석이 필수적이고, 웹페이지를 상태기반으로 나타내고 이벤트에 의한 상태의 변환을 점검하는 방법이 중요하다.[8]

3. 다각적 접근법에 의한 소프트웨어 평가

경험적 연구 (Empirical Study)에 근간을 둔 소프트웨어 평가 프레임워크는 (그림 1)에서와 같이 사용자 설문 분석을 통한 평가 (사용자평가), 표준 평가모듈에 근간한 평가 (기능평가), 사용 패턴 분석에 의한 평가 (패턴평가)를 모두 반영하는 다각적 접근법에 의한 소프트웨어 평가 (MASE - Multilateral Approach for Software Evaluation) 형태를 취한다. 사용자평가는 S/W 의 실 사용자 그룹에 ISO/IEC 9126-2 을 근거로 하는 설문을 수거 분석하여 평가하는 것이다. 기능평가는 ISO/IEC 9126-2 의 외부품질 (External Quality)에 근거하여 작성된 평가모듈과 규격서에 의한 기능 테스트를 의미하며 경험이 풍부한 평가 전문가 들에 의해 수행된다. [6] 패턴평가는 실 사용자가 S/W 를 사용하는 방식 (Pattern)을 분석하여 이를 ISO/IEC 9126-3 의 사용품질 (Quality in Use)을 기준으로 평가하는 것이다. S/W 를 사용하는 방식은 중요도에 따라 가중치를 부여하여 평가한다. 이때, 평가의 객관성을 확보하기 위해 통계적인 기법을 이용하여 신뢰도 측정과 민감도 분석을 한다.[5] 이러한 다각적 평가는 동시에 또는 순서에 따라 진행된다.



(그림 1) MASE 소프트웨어 평가 프레임워크 (S/W Evaluation Framework)

3.1 사용자 설문 분석을 통한 S/W 평가

사용자평가는 ISO/IEC 9126-1 과 ISO/IEC 9126-2 를 바탕으로 작성된 설문을 실제 사용하고 있는 사용자 그룹에 실시한 결과 분석을 통해 이루어진다. 분석한 결과는 ISO/IEC 9126-1 에서 정의한 품질특성별로 분석되어 진다.

평가한 결과를 수식으로 표현해 보면 다음과 같고, 이러한 평가를 수행하는 방법과 유의해야 할 사항 등은 아래에 기술되어 있다.

$$E = \sum_{i=1}^{n} a_i U_i \quad \cdots\cdots (1)$$

E = 가중치 반영 S/W 품질 평가 값 (90% 신뢰도 내에서의 범위값)

a = 시나리오의 가중치 범위 값 (최소값, 최적값, 최대값)

U = 평가 값

N = 평가 항목 순서 (1, 2, 3, …)

3.2 표준 평가모듈에 근간한 평가

표준 평가모듈에 근간한 평가는 ISO/IEC 9126-1 의 품질모델과 ISO/IEC 9126-2 의 외부품질 (External Quality)에 근거하여 작성된 평가모듈과 규격서에 의한 기능 테스트를 수행하여 평가하는 것으로 경험이 풍부한 평가 전문가 들에 의해 수행된다. 평가모듈은 아래 [표 1]과 같다.[6]

이러한 평가를 위해 ISO/IEC 9126 이라는 국제 표준을 근간으로 한 평가 모듈(EM)에 정의한 소프트웨어 시험· 평가를 위한 품질 기준을 마련하였다. 그리고, 소프트웨어 별로 특성을 반영하여 전체 품질 특성 에서 점검 가능한 범위를 선정하기 위해 시험· 평가 대상 소프트웨어에 따라 다른 시험 규격서를 도출하여 사용한다. 범위가 선정된 품질 특성을 점검하기 위해 품질 검사표를 점검하기 위한 점검표를 도출하고 각 점검표 상에는 평가 항목 점검을 위한 상세 점검 항목들을 도출한다. 마지막으로, 점검표 상에 기술된 상세 점검 항목을 시험하기 위해 테스트케이스를 도출하여 시험을 수행한 후 보고서를 작성한다.

[표 1] 평가모듈

품질특성	부특성	평가 항목
일반적 요구사항		
기능성		
신뢰성		
효율성		
사용성	이해가능성	
	학습성	
	운영성	
	선호도	
	준수성	사용성 표준 준수 정보제공
		사용성 표준 준수율
유지보수성		
이식성		

3.3 사용 패턴 분석에 기반한 소프트웨어 평가

소프트웨어를 실제로 사용하는 패턴에 따라 평가하는 것은, 사용자평가 시 설문 자체가 부정확할 수 있다는 설문 분석의 한계와 기능평가 시 단위 기능테스트가 갖는 통합적이지 못한 한계점을 극복하기 위해 필요하다. 패턴평가는 단위 기능테스트가 문서작업만 남기고 끝나는 시점에서 시행하는 것이 적당하다. 이는 평가 대상 소프트웨어에 대해 평가원들이 충분한 지식과 경험을 갖게 되었을 때 효과적으로 사용패턴을 정의할 수 있기 때문이고 많은 실제 시험을 통해 경험적으로

입증되었다.

사용자의 사용 패턴 분석에 기반한 평가는 Emam (2001)이 제안하는 방법론을 근간으로 하고, 실제 적용 시 적절하게 변형하여 사용한다. 사용패턴은 평가자가 주관하여 사용자 그룹을 중심으로 의견을 취합하여 작성되어야 하며 필요에 따라 해당 소프트웨어 개발업체의 검토를 받을 수 있다. 사용패턴별로 중요도가 다를 수 있고 중요도에 불확실성이 존재할 수 있어 사용자 합의 하에 범위값 (0~100)으로 결정한다. 범위값은 삼각분포를 따르는 것으로 가정하여 최저값, 최적값, 최고값을 사용자로부터 이끌어 낸다. 사용패턴은 국제표준인 ISO/IEC 9126 의 사용품질 (Quality in Use)에 근간하고 소프트웨어의 특성을 반영한 평가기준항목 (criteria)을 설정하여 평가한다.[5]

사용패턴평가는 두명이 독립적으로 수행하거나 두팀 이상이 독립적으로 수행하고, 이들 독립적인 평가간의 연관성을 카파 계수 (Kappa coefficient)를 구하여 파악하도록 한다. 카파 계수 값을 통해 두 독립적인 평가가 긴밀하게 연계되어 있어 3 번째 평가자나 평가팀이 평가하여도 동일한 결과를 얻을 수 있다는 것을 밝혀 평가의 신뢰성 (reliability)를 확보한다. 이때, 카파 계수는 0.44 보다 작으면 두 평가사이에 동의하는 강도 (Strength of Agreement)가 약하고, 0.44 에서 0.62 사이이면 중간정도 이고, 0.62 이상이면 강하다고 보며, 이를 신뢰도 측정의 근거로 한다.[3] [4]

평가항목인 시나리오를 측정하기 위한 기준은(Criteria)은 ISO/IEC 9126 의 사용품질 (Quality in Use)의 만족도 (Satisfaction), 생산성 (Productivity), 효과성 (Effectiveness), 안전성 (Safety)을 근거로 평가대상 소프트웨어의 특성을 반영하여 평가자가 결정한다. 즉, 각각의 시나리오에 대해서 몇 개의 평가기준항목 (Criteria)이 있을 수 있으며 이는 4 점 척도 (0, 1, 2, 3)로 구하는 것을 기본으로 한다.[2]

가중치를 반영한 평가 결과의 범위값은 몬테칼로 시뮬레이션 (Monte Carlo Simulation)으로 처리하여 평가기준항목 (Criteria)별로 통계적 평균을 구하는 것을 의미한다. 이렇게 산출된 최종 평가값은 평가의 불확실성을 반영한 값이다.[5]

4. 다각적 접근법에 의한 S/W 평가(MASE) 적용사례

평가 대상 소프트웨어는 솔루션은 온라인 상의 모든 고객 상담 채널을 통합하여 관리하고, 고객의 문의사항에 대해서 신속하고 편리하게 응대를 해주는 솔루션이다. 고객들의 e-mail 문의에 대한 자동화 답변 도구, 단순하고 반복적인 질문을 관리하는 도구, 고객과의 일대일 문자 및 음성 채팅을 실시간으로 제공하는 도구 등을 포함하고 있다.

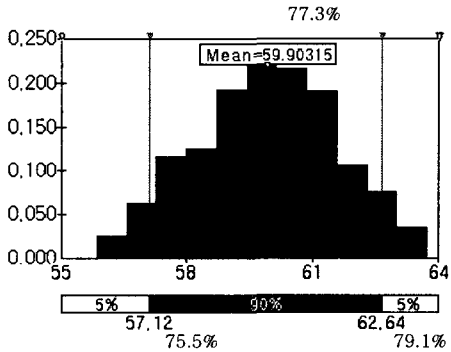
사용자평가를 위해 사용자 설문서를 앞서 [표 1]에서 설명한 평가모듈과 같은 형태로 사용자들이 알기 쉽게 변형하여 사용하였다. 설문 답변자는 라이코스 코리아, 아시아나항공, 야후 코리아 등의 업체에서 해당 소프트웨어를 직접 실무에 사용하고 있는 20 명의 실무자로 구성되어 있다.

가중치는 평가 항목별로 모든 사용자가 범위값으로 부여한 중요도를 평균하여 도출한다. 이때, 이상치를 제거하고자 최소값과 최대값을 제외하고 평균하였다. 그리고, 평가는 4 점 척도 (0, 1, 2, 3)로 이루어 졌으며, 평가값에 가중치를 반영하는 방법은 아래의 식 (2)와 같으며 이 값은 평가 값이 최대점수이고 가중치가 최대치일 때 최고값 100 점이 주어지도록 고안되어있다.[5] 이러한 평가값을 식 (1)에 적용하여 전체 평가값을 도출 하였다. 전체 평가값을 몬테칼로 시뮬레이션을 통해 구한 결과는 아래 그림과 같고, 평가 결과는 90%의 신뢰도 내에서 57.12~62.64 의 평가 값을 가지며 평균치는 59.90 이다. 이를 백분율 값으로 변형하면, 50%의 만족이 (50 X 1.5) / 3 = "25"의 평가값을 의미하며, 75%를 만족시키는 평가값은 (75 X 2.25) / 3 = "56.25"이다. 결국 60 점일 경우 77~78%를 만족시키는 값임을 알 수 있다.

$$\frac{W_{range}}{3} \times U_n \quad \text{..... (2)}$$

W_{range} = 가중치 (Weight) 범위 값 (최소값, 최적값, 최대값)

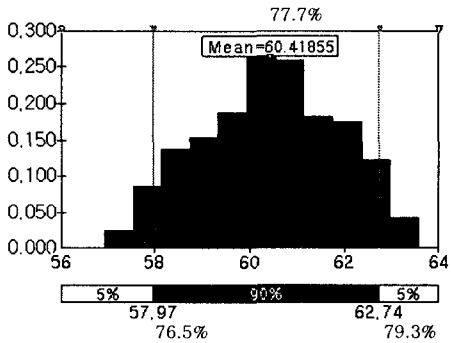
U_n = 평가 항목의 평가값



(그림 2) 대상 소프트웨어 전체 평가 값

평가 값에 가중치를 반영하여 품질특성 별로 평가한 결과는 평가 값의 평균을 몬테칼로 시뮬레이션을 통해 계산하였으며, 90%의 신뢰도로 범위값을 구하였다.

품질특성 별 평가 결과의 한 예로, 기능성의 경우, (그림 3)에서 보는 바와 같이 90%의 신뢰도 내에서 57.95~62.74 의 평가 값을 가지며 평균치는 60.42 이다.



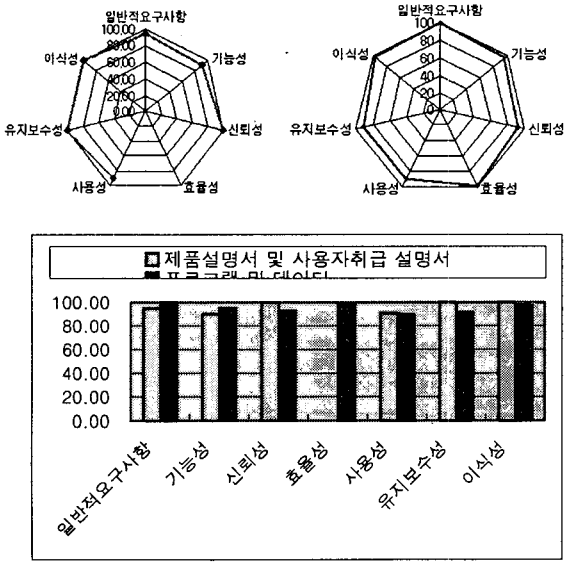
(그림 3) 기능성 평가 값

신뢰성 평가 결과는 90%의 신뢰도 내에서 59.86~64.85 의 평가 값을 가지며 평균치는 62.40 이다.

효율성, 사용성, 유지보수성, 이식성은 각각 범위값 58.2~63.55 에 평균치 60.87, 범위값 56.9~62.74 에 평균치 59.84, 범위값 50.7~55.89 에 평균치 53.32, 범위값 60.09~66.65 에 평균치 63.48 로 유지보수성이 상대적으로 낮은 점수를, 이식성이 높은 점수를 보이고 있다. 전체 평가값은 범위값 56.98~62.53 에 평균치 59.84 를 보이고 있다.

기능평가는 소프트웨어시험센터의 시험원 5 명이 평가 모듈에 근거하여 20 일간 진행되었다.[6] 문서 평가와 프로그램 평가로 분리되어 진행되었으며, 아래 (그림 4)에서 보는 바와 같이, 문서 (제품 설명서 및 사용자 취급 설명서) 평가는 일반적요구사항 (95.0%), 기능성 (90.1%), 신뢰성 (100%), 사용성 (90.9%), 유지보수성 (100%), 이식성 (100%)의 평가 요구 사항 만족도를 보였다. 여기서, 3 가지 항목에 대해 100%가 나온 이유는 사용자 메뉴얼에 해당 품질특성이 요구하는 사항이 적절히 기술되어 있기 때문이다.

프로그램 평가는 일반적 요구 사항 (100%), 기능성 (95.2%), 신뢰성 (92.6%), 효율성 (98.5%), 사용성 (89.8%), 유지보수성 (91.5%), 이식성 (97.5%)의 평가 요구 사항 만족도를 보였다. 여기서, 일반적 요구사항이 만점을 받은 이유는 이 항목이 제공된 프로그램의 바이러스 감염여부만을 확인하기 때문이다. 이는 소프트웨어가 평가 대상이 되기 위한 기본항목이기도 하다.



(그림 4) 기능평가 결과

패턴평가의 경우, 기능평가 시 기능테스트가 커버할 수 없는 부분을 시험하기 위해 개개의 단순기능테스트를 연결하는 차원에서 시험원이 자의적으로 사용자 입장에서 패턴을 생성하여 평가하였다. 이는 제안된 패턴평가를 만족시키지 못하는 시도이다. 향후 평가에서는 사용패턴을 시험전문가가 기능테스트 한 경험과 지식을 바탕으로 사용자 설문 분석 결과를 반영하는 것은 물론, 직접적으로 사용자와의 인터뷰를 통해 사용자가 중요하다고 생각하는 내용을 반영하여 도출해낼 것이다. 이를 위해 체계적으로 사용패턴을 도출해 내는 방법론과 모델, 그리고 구체적인 가이드라인에 대한 연구를 수행할 계획이다.

5. 결론

소프트웨어의 평가를 사용자평가, 기능평가, 패턴평가로 평가하여 이를 종합하는 다각적 접근법에 의한 평가를 제안하고, 이를 이메일 관리 소프트웨어에 실제로 적용하여 그 실용성을 검증하고자 하였다.

개개의 평가 모두 국제표준인 ISO/IEC 9126 의 품질 특성과 사용품질에 기반하고 있다. 사용자평가는 사용자 설문을 분석하여 평가하는 것이고, 기능평가는 평가모듈에 근간하여 전문 시험기관이 평가하는 것이며, 패턴평가는 사용 패턴을 전문가와 사용자가 함께 만들어 내고 이를 기반으로 평가/분석하여 평가하는 것이다. 이는 평가에 소요되는 시간은 많을 수 있으나 정확하고 객관적인 평가에는 유용하다. 그리고, 한가지씩 개별적으로 평가하는 것 보다는 3 가지가 통합적으로 적용됨으로써 상대적으로 소요되는 시간도 상당부분 단축되는 것도 경험적으로 확인하였다.

평가 적용대상 소프트웨어에 대한 패턴 분석이 체계적으로 이루어지지 않아 패턴평가가 정식으로 이루어지지 않았으므로 제안된 평가 모델이 완전히 적용되지는 않았다. 이는 향후의 시험을 통해서 지속적으로 검증하여 나갈 것이다. 그리고, 향후에는 각각의 평가 간의 연관성에 대한 분석과 연구도 함께 진행할 예정이다.

6. Reference

[1] Andrew K. Rae, H. Hausen, P. Robert, Software Evaluation for Certification, McGRAW-HILL Book Company Europe, 1995

[2] ISO/IEC 9126: Software engineering- software product quality

[3] J. Cohen, A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement, vol. XX, no. 1, pp.37-46, 1960

[4] J. Cohen, Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Agreement or Partial Credit, Psychological Bulletin, vol. 70, pp. 213-220, 1968

[5] Khaled El Emam, A methodology for Scenario-Based Software Certification, ETRI report, 2001

[6] 박상욱 외, 패키지 소프트웨어 품질 인증을 위한 시험· 평가 프레임워크, 한국정보처리학회 추계학술대회, 2001.10

[7] 이종무 외, 측정척도를 고려한 패키지 S/W 품질인증 방법, 한국정보시스템학회 2001 추계 학술대회, 2001

[8] 최은만, 웹 기반 소프트웨어의 시험 및 검증 기술, 정보과학회지, 제 19 권 제 11 호, 2001.11