Abordări recente în evaluarea formativă a utilizabilității

Costin Pribeanu, Ruxandra Dora Marinescu, Irina Cristescu, Lucreţia Pîrvu, Dragoş Daniel Iordache
Institutul Național de Cercetare-Dezvoltare în Informatică – ICI Bucureşti
Bd. Mareşal Averescu nr.8-10, Bucureşti
{pribeanu, doruma, irina.cristescu, lpirvu, iordache}@ici.ro

REZUMAT
Creşterea gradului de utilizare cât şi creşterea calitativă a informației digitale necesită o abordare centrată pe utilizator a procesului de proiectare, care să conducă la o îmbunătățire substanțială a utilizabilității aplicațiilor informatice. Scopul acestui articol este de a prezenta sintetic cele mai recente abordări în evaluarea formativă a utilizabilității sistemelor interactive. Problematica prezentată este diversă, incluzând indicatori folosiți în prezent pentru gradul de încredere şi validitatea rezultatelor obținute cu diferite metode.

Cuvinte cheie
Utilizabilitate, evaluare formativă, evaluare euristică.

Clasificare ACM
D.2.2: Design tools and techniques. H5.2 User interfaces.

INTRODUCERE

Dezvoltarea societății informaționale moderne se bazează pe accesul larg la tehnologia informatică într-o economie din ce în ce mai competitivă. In acest sens, creşterea gradului de utilizare a tehnologiei informatice şi creşterea cantitativă şi calitativă a informației digitale sunt două priorități majore. Atât creşterea gradului de utilizare cât şi creşterea calitativă a informației digitale necesită o abordare centrată pe utilizator a procesului de proiectare, care să conducă la o îmbunătățire substanțială a utilizabilității aplicațiilor informatice.

In funcție de momentul şi scopul evaluării, evaluarea utilizabilității poate fi formativă sau sumativă. Evaluarea formativă a utilizabilității este efectuată iterativ, pe parcursul ciclului de dezvoltare cu scopul identificării şi eliminării problemelor de utilizabilitate cât mai devreme (Theofanos & Quesenberry, 2005).

Deşi pe plan internațional s-au făcut eforturi considerabile de cercetare în ultima decadă, probleme serioase de utilizabilitate apar în continuare cu o frecvență ridicată la toate tipurile de produse software. Acest fapt denotă că evaluarea utilizabilității nu este încă integrată în ciclul de dezvoltare a sistemelor software.

Pe plan național, evaluarea utilizabilității este practicată sporadic, în cadrul unor proiecte de cercetare. Utilizabilitatea produselor informatice este scăzută, întrucât nu există încă o practică curentă de evaluare a produselor informatice înainte de a fi instalate la utilizator.

Scopul acestui articol este de a prezenta sintetic cele mai recente abordări în evaluarea formativă a utilizabilității sistemelor interactive. Problematica prezentată este diversă, incluzând indicatori folosiți în prezent pentru gradul de încredere şi validitatea rezultatelor obținute cu diferite metode.

SCOP ŞI METODE UTILIZATE

Scopul evaluării formative

Evaluarea formativă a utilizabilității are ca scop identificarea problemelor de utilizabilitate, analiza acestora şi formularea de sugestii pentru remediere. O preocupare de interes actual este analiza comparativă a problemelor de utilizabilitate care pot fi identificate cu diferite metode.

O problemă de utilizabilitate a fost definită de Nielsen (1993) ca orice aspect al interfeței cu utilizatorul despre care se presupune (se observă) că ar crea utilizatorului dificultăți / nemulțumiri raportat la un indicator important al utilizabilității (uşurință în învățare, uşurință în operare, rată de erori, satisfacție subiectivă) şi care poate fi atribuit unui singur aspect de proiectare.

Problemele de utilizabilitate se pot clasifica după gradul de severitate în trei categorii:

- Severe: probleme care nu permit ca utilizatorul să îndeplinească sarcina sau care rezultă în pierderi importante de date sau timp.
- Moderate: probleme care au un impact semnificativ asupra îndeplinirii sarcinii dar pentru care utilizatorul poate găsi o soluție.
- Minore: probleme care sunt iritante pentru utilizator dar care nu au un impact semnificativ asupra îndeplinirii sarcinii.

Pierderea de date este considerată catastrofală fie dacă datele nu pot fi refăcute, fie dacă refacerea necesită costuri foarte mari. Pierderea de timp este relativă la durata sarcinii.

Practica standard cere remedierea problemelor severe de utilizabilitate înainte de furnizarea produsului / aplicației software. În cel mai rău caz, acestea nu trebuie tolerate în sarcinile critice sau în cele cu o frecvență ridicată de execuție.

Problemele de utilizabilitate pot fi identificate atât prin testare cu utilizatori cât şi prin evaluare euristică. Aşa cum arată Law & Hvannberg (2002), cele două metode sunt complementare.

Evaluarea euristică

Evaluarea euristică – HE (Heuristic Evaluation) este o metodă de inginerie a utilizabilității realizată de un număr redus de evaluatori care examinează o interfață cu utilizatorul, judecă respectarea unui set de principii de

utilizabilitate (euristici), elaborează o listă de probleme de utilizabilitate (UP) clasificate pe categorii de severitate, corespunzător impactului estimat asupra performanţelor utilizatorului sau acceptanţei (Molich & Nielsen, 1990; Nielsen, 1994).

Evaluarea euristică este, în esenţă, o inspecţie de utilizabilitate. Se recomandă utilizarea acestei metode înaintea testării cu utilizatori, astfel încât să se elimine un număr cât mai mare de probleme de utilizabilitate cu costuri cât mai reduse.

De regulă, în evaluarea euristică se colectetază:

- Măsuri cantitative: numărul de probleme de utilizabilitate pe două niveluri de severitate: major şi minor.
- Măsuri calitative: descrierea detaliată a UP individuale şi a contextului în care apar.

HE este cea mai răspândită metodă de evaluare, folosită de 76% din specialiştii din domeniul utilizabilităţii (UPA Survey, 2005).

Este necesară luarea în consideraţie a expertizei evaluatorilor pentru a estima corect gradul de încredere. Nu numai expertiza generală în utilizabilitate este utilă, dar şi specializarea acestora pe anumite aspecte ale sistemelor interactive, atât la nivel general (probleme cognitive, design grafic, sisteme web, intraneturi) cât şi de detaliu (arhitectura informaţiei, navigare, procesarea tranzacţiilor). În prezent există preocupări de atestare a evaluatorilor pe niveluri de expertiză, în scopul perfecţionării şi alocării optime a resursei umane.

Testarea cu utilizatori (UT)

Evaluarea formativă a utilizabilităţii prin testarea cu utilizatori (UT – User Testing) este o metodă care necesită un număr relativ restrâns de utilizatori (minim 5, ideal 15-25), care testează aplicaţia pe baza unor sarcini predefinite.

Testarea cu utilizatori este definită ca o metodă empirică intensivă de evaluare a utilizabilităţii implicând participanţi având caracteristici apropiate de cele ale utilizatorilor reali ai produsului care va fi evaluat.

Prin testarea cu utilizatori se înregistrează comportamentul utilizatorului cu ajutorul unor tehnici specifice cum sunt observarea, protocoalele de « gândire cu voce tare » (TA – Think Aloud) sau înregistrările video. Avantajul principal al tehnicii « gândire cu voce tare » o constituie mai buna înţelegere a modelului mental al utilizatorului şi interacţiunea cu produsul. În funcţie de logistica existentă, se recomandă combinarea mai multor tehnici.

Principala diferenţă între evaluarea euristică şi testarea cu utilizatori, o reprezintă absenţa utilizatorului în primul caz.

De regulă, în testarea cu utilizatori măsurile colectate sunt:

- Măsuri ale eficacităţii şi eficienţei : durata, erorile (numărul de alegeri greşite din meniu, numărul de selecţii greşite, alte erori), solicitarea ajutorului (numărul de cereri explicite de asistenţă din partea experimentatorului şi numărul de solicitari de ajutor online)
- Măsuri ale satisfacţiei : expresiile emoţionale (observarea frustrării, observarea confuziei), percepţiile, opiniile şi aprecierile participanţilor.

Măsurile eficacităţii şi eficienţei sunt colectate cu ajutorul unui stopwatch şi/sau al fişierelor de log. Măsurile satisfacţiei subiective sunt colectate cu : a) chestionare post-test b) Protocoale de « gândire cu voce tare ».

UT este o metodă de evaluare utilizată frecvent, fiind folosită de 75% din specialiştii în domeniul utilizabilităţii (UPA Survey, 2005).

PREOCUPĂRI ACTUALE ÎN EVALUAREA FORMATIVĂ

Descrierea problemelor de utilizabilitate

Hvanberg & Law (2003) au propus o schemă de clasificare şi un model de descriere a problemelor de utilizabilitate porinind de la schema de clasificare a defectelor (DCS – Defect Classification Scheme), considerând că o problemă de utilizabilitate este similară unui defect (deficienţă) a unui sistem software. Schema cuprinde 10 atribute, pentru fiecare dintre ele fiind menţionate motivul pentru care a fost ales precum şi un set de valori posibile, după cum urmează : identificatorul problemei, descrierea problemei, metoda de identificare (HE / UT), declanşator (descrierea contextului), impact, etapa de dezvoltare în care se anticipează că va fi remediată, tipul problemei, etapa în care a fost remediată (pentru comparaţie cu etapa anticipată), sugestii pentru prevenire.

Într-o lucrare ulterioară, Vilbersdottir et al. (2006) au extins / modificat schema de clasificare cu câteva atribute. A fost introdus contextul (partea de interfaţă) separat de declanşator şi frecvenţa (număr de evaluatori care au identificat problema). Au fost adăugate atribute specifice UT : eficienţă (timp de execuţie sarcină), eficacitate (rată de succes), număr de frustrări şi număr de solicitări asistenţă.

Deşi aceste clasificări sunt utile pentru studiul utilizabilităţii, inclusiv din punct de vedere metodologic, acestea au o valoare practică redusă : majoritatea practicienilor rapoartează problemele de utilizabilitate într-un format mult mai succint.

Hvannberg et al. (2007), într-o lucrare mai recentă, au utilizat un format structurat care a fost adaptat după Cockton & Woolrich (2001) şi care conţine : identificatorul UP, descriere, dificultăţile întâmpinate de utilizator, contextul specific (localizarea problemei în interfaţă), cauzele posibile (ce anume este greşit în proiect), euristicile utilizate şi gradul de severitate (impactul, pe trei niveluri : major, moderat şi minor).

Consolidarea problemelor de utilizabilitate

Consolidarea problemelor de utilizabilitate este parte integrantă din evaluarea formativă în care sunt implicaţi mai mulţi evaluatori. Conform cu Law & Hvannberg (2008), procesul de consolidare a problemelor de utilizabilitate identificate de mai mulţi utilizatori (experimentul se referă la testarea cu utilizatori) se desfăşoară în două etape:

- *filtrarea:* eliminarea duplicatelor din lista de probleme identificate de către un singur evaluator

- *integrarea*: combinarea problemelor de utilizabilitate dintre diferite liste identificate de mai mulţi evaluatori, cu scopul de a reţine problemele unice, relevante.

Prima etapă are rolul de consolidare individuală: fiecare evaluator compară listele de probleme identificate de diferiţi utilizatori pentru aceeaşi sarcină pentru a produce o listă de probleme unice pentru acea sarcină.

A doua etapă are rolul de consolidare colaborativă: doi sau mai mulţi evaluatori, care au utilizat aceleaşi rapoarte de testare, cad de acord asupra unei liste unice de probleme de utilizabilitate pentru fiecare sarcină.

Un studiu asemănător, tot pe baza testării cu utilizatori, a fost raportat de Hornbaek & Frokjaer (2008), Scopul studiului a fost compararea tehnicilor utilizate pentru consolidarea problemelor de utilizabilitate. Tehnicile utilizate au fost următoarele:

- *Modificări similare* în aplicaţie, în vederea remedierii. Două probleme sunt identice, dacă remedierea uneia o rezolvă şi pe cealaltă.
- *Prioritizare practică* pentru proiectanţi sau dezvoltatori. Presupune o listă ordonată după prioritate şi are o valoare practică (aşa se procedează).
- *Caracteristicile de descriere a unei probleme de utilizabilitate*: cauză, problemă, rezultat şi modificare sugerată.
- *Cadrul de lucru al acţiunii utilizator* (UAF - User Action Framework) care constă într-o schemă de clasificare bazată pe 382 de categorii organizate pe 6 niveluri.

O concluzie a autorilor este că utilizarea de tehnici diferite produce rezultate diferite, atunci când sunt implicaţi evaluatori începători. Prima metodă a produs un număr mai mare de probleme de utilizabilitate. Cea de a doua metodă a întrunit cel mai mare coeficient de agreere între participanţi, în timp ce cea de a treia metodă a întrunit cel mai scăzut consens.

GRADUL DE ÎNCREDERE ŞI VALIDITATE

Indicatori de încredere

Încrederea se referă la măsura în care evaluările independente produc acelaşi rezultat. Validitatea se referă la măsura în care problemele detectate în cadrul evaluării sunt aceleaşi cu cele care apar în practica utilizării sistemului.

Pentru măsurarea gradului de încredere, Hertzum & Jacobsen (2001) recomandă doi indicatori: rata de detecţie şi coeficientul de agreere între oricare doi evaluatori (any-two-agreement).

Rata de detecţie se calculează după formula:

$$Rata\ detecţiei = Media\ |P_i|/|PU|\ pentru\ toţi\ n\ evaluatori$$

In această ecuaţie, P_i este setul de probleme detectate de evaluatorul E_i, PU este setul de probleme detectate colectiv de toţi evaluatorii, iar n este numărul de evaluatori. Calculul se face pe baza listei de probleme unice de utilizabilitate, adică după eliminarea duplicatelor care apar în listele individuale şi între aceste liste.

Coeficientul de agreere se calculează după formula:

$$Coeficientul\ de\ agreere = Media\ |P_i \cap P_j|/|P_i \cup P_j|\ pentru\ toate\ perechile\ 1/2n*(n-1)\ de\ evaluatori$$

In această ecuaţie, P_i şi P_j sunt seturile de probleme detectate de evaluatorii E_i şi E_j iar n este numărul de evaluatori. Calculul se face pe baza listei de probleme unice de utilizabilitate, adică după eliminarea duplicatelor care apar în listele individuale şi între aceste liste.

Efectul evaluatorului

Practica evaluării arată că atât problemele de utilizabilitate identificate, cât şi modul de descriere a acestora pot varia semnificativ de la un evaluator la altul. Această diferenţă este cunoscută sub numele de *efect al evaluatorului* şi este întâlnit atât în evaluarea euristică cât şi în testarea cu utilizatori. În mod asemănător, lista problemelor de utilizabilitate diferă semnificativ şi de la un utilizator la altul, diferenţă cunoscută sub numele de *efect al utilizatorului*.

Efectul evaluatorului în evaluarea utilizabilităţii este definit de Hertzum & Jacobsen (2001) ca situaţia în care mai mulţi evaluatori care apreciază aceeaşi interfaţă cu aceeaşi metodă de evaluare, identifică seturi de UP diferite şi grade de severitate diferite pentru aceeaşi problemă. Efectul evaluatorului este o măsură a gradului de încredere cu care pot fi privite rezultatele evaluării.

Analiza a 11 studii de utilizabilitate efectuate cu 3 metode de evaluare diferite: inspecţie cognitivă (CW – Cognitive Walkthrough), evaluare euristică (HE) şi « gândire cu voce tare » , făcută de Hertzum & Jacobsen (2001), arată că efectul evaluatorului se manifestă pentru:

- Evaluatori experţi şi începători.
- Probleme de utilizabilitate minore şi majore.
- Identificare de probleme şi apreciere grad de severitate.
- Evaluare de sisteme simple şi sisteme complexe.

Hertzum & Jacobsen (2001) arată că efectul utilizatorului nu poate fi atribuit slăbiciunii unei metode, particularităţilor unui studiu / artefact şi nici nu este un incident întâmplător. Ca atare, nu se pune problema existenţei acestui efect ci a cauzelor şi căilor în care poate fi diminuat. Principala cauză este faptul că o evaluare a utilizabilităţii presupune interpretare, care depinde, la rândul ei, de diferenţele individuale între evaluatori în ceea ce priveşte abilităţile cognitive, expertiza, motivaţia şi personalitatea.

Referitor la aplicarea metodelor de evaluare, Hertzum & Jacobsen (2001) au propus trei recomandări pentru a reduce efectul evaluatorului:

- Scopul analizei şi selectarea sarcinilor trebuie să fie explicite.
- Este important să fie găsite cele mai multe dintre problemele unui sistem şi se recomandă în mod expres să fie implicaţi mai mulţi evaluatori.

- Accent pe procedurile de evaluare şi pe criteriile de selecţie a problemelor în fiecare caz.

Intr-un alt studiu, Jacobsen et al. (1998) arată că 4 evaluatori au identificat în medie 52% din numărul total de probleme de utilizabilitate.

Indicatori de validitate

Evaluarea euristică produce un set de probleme de utilizabilitate care sunt considerate *anticipate* în timp ce testarea cu utilizatori identifică probleme de utilizabilitate *reale*. Diferenţa între cele două seturi o reprezintă problemele *fals pozitive* (alarme false, identificate de HE dar neconfirmate de UT) şi *fals negative* (problemele de utilizabilitate care nu au fost identificate de HE).

Pentru măsurarea validităţii, eficacităţii şi eficienţei evaluării, se folosesc 3 indicatori (Hartson et al., 2001) : validitate, completitudine (thoroughness) şi eficacitate generală (overall effectiveness).

Validitatea se defineşte după formula:

```
Validitate = UP anticipate / UP
anticipate + UP fals pozitive
```

Completitudinea se defineşte după formula:

```
Completitudine = UP anticipate / UP
reale
```

Eficacitatea generală se defineşte după formula:

```
Eficacitate = Validitate *
Completitudine
```

Aşa cum se observă, calcularea acestor indicatori presupune evaluarea cu ambele metode, pe baza aceluiaşi scenariu de utilizare, fapt care măreşte costurile evaluării.

CONCLUZII

În ultimii ani asistăm la tendinţe de extindere a cadrului de lucru în evaluarea utilizabilităţii sistemelor informatice, ca urmare a provocărilor societăţii informaţionale:

- Extinderea caracterului multidisciplinar al activităţii de evaluare determinat de cerinţele unor tehnologii specifice, cum sunt cele de e-commerce, e-government şi e-learning.
- Studiul măsurilor adecvate pentru evaluarea utilizabilităţii şi pentru înţelegerea relaţiilor între diferitele măsuri utilizate în practica evaluării.
- Integrarea evaluării utilizabilităţii în metodologia de proiectare centrată pe om.
- Creşterea interesului pentru validitatea rezultatelor şi adecvarea metodelor de evaluare, inclusiv prin compararea rezultatelor obţinute cu diferite metode şi utilizarea de metode complementare.

Un aspect important, tot mai des referit în literatura de specialitate din ultimii doi ani, este utilitatea „în aval'" a metodelor de evaluare formativă (downstream utility): măsura în care plusul sau minusul de utilizabilitate (îmbunătăţirea sau deteriorarea) a unui sistem poate fi atribuit în mod direct remedierilor solicitate de evaluările efectuate (Law et al., 2007).

Confirmare

Această lucrare este finanţată din proiectul de cercetare PS MCSI 49/2008.

REFERINŢE

1. Cockton, G., Woolrych, A. (2001). Understanding inspection methods: lessons from an assessment of heuristic evaluation. Blandford, A., Vanderdonckt, J., Gray, P.D. (Eds.), *Proceedings of People and Computers XV*. Springer-Verlag, 171–182
2. Hartson, H.R., Andre, T.S., Williges, R.C., (2001). Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction 13*, 373–410
3. Hertzum, M., Jacobsen, N.E. (2001) „The evaluator effect: A chilling fact about usability evaluation methods". *International Journal of Human-Computer Interaction 13*. 421-443.
4. Hornbaek , K & Frokjaer, E. (2008). Comparison of techniques for matching of usability problem descriptions. *Interacting with Computers 20*. 505-514.
5. Hvannberg, E.T. and Law, E.L.-C. (2003) "Classification of Usability Problems (CUP) Scheme", *Proceedings of Interact 2003*.
6. Hvannberg, E.T. and Law, E.L.-C., Larusdotir, M.C. (2007) „Heuristic Evaluation: Comparing ways of finding an reporting usability problems, *Interacting with Computers 19, 255-240*.
7. Jacobsen, N.E., Hertzum, M., John, B. (1998) „The evaluator effect in usability tests" *Proceedings of ACM CHI'98*. ACM Press. 255–256.
8. Law, L-C. & Hvannberg, E.T. (2002) "Complementarities and convergence of heuristic evaluation and usability test: A case study of UNIVERSAL brokerage platform". *Proceedings NordiCHI Conference 2002*. ACM. 71-79
9. Law, E.C, Lárusdóttir, M.C, Norgaard, M. (Eds) (2007) *Downstream Utility 2007 : The Good, the Bad, and the Utterly Useless Usability Evaluation Feedback*. IRIT Press – Toulouse.
10. Molich, R., and Nielsen, J. (1990). Improving a human-computer dialogue, *Communications of the ACM 33, 3*, 338-348.
11. Nielsen, J. (1993). *Usability Engineering*. Academic Press, New York.
12. Nielsen, J. (1994). "Heuristic evaluation". Nielsen, J., and Mack, R.L. (Eds.), *Usability Inspection Methods*, John Wiley & Sons, New York.
13. Theofanos, M. & Quesenbery, W. (2005) "Towards the Design of Effective Formative Test Reports". *Journal of Usability Studies, Issue 1, Vol.1*. 27-45
14. *UPA 2005 Survey*. Usability Professionals Association.
15. Vilbersgtottir, S.G., Law, E.L.-C., Hvannberg, E.T.. (2006) „Classification of Usability Problems (CUP) Scheme: Augmentation and Exploitation, *Proceedings of NordiCHI 2006*, 281-290.