

ÉVALUATION DE L'UTILISABILITÉ D'UN SITE WEB : TESTS D'UTILISABILITÉ VERSUS ÉVALUATION HEURISTIQUE

BOUTIN MARIO

Centre de recherche informatique de Montréal (CRIM), 550, rue Sherbrooke Ouest, Bureau
100, Montréal, Québec, Canada, H3A 1B9. mboutin@crim.ca

MARTIAL ODILE

Centre de recherche informatique de Montréal (CRIM), Montréal, Québec, Canada

Résumé

La conception et la réalisation d'un site Web nécessitent, entre autres, de tester l'utilisabilité du site avant sa mise en ligne. Deux techniques sont couramment utilisées pour tester les interfaces graphiques traditionnelles : les tests d'utilisabilité et l'évaluation heuristique. La présente étude vise à comparer ces deux techniques pour l'évaluation de sites Web. L'étude a été réalisée sur le site Web du CRIM. Les résultats obtenus avec chacune des deux techniques sont présentés et comparés en fonction des problèmes identifiés (type et quantité) et des coûts associés. Cette comparaison permet de conclure que l'évaluation heuristique ne peut être utilisée seule et que les tests d'utilisabilité donnent des résultats supérieurs surtout pour l'identification de problèmes sérieux.

Mots clés : site Web, tests d'utilisabilité, évaluation heuristique.

USABILITY EVALUATION OF A WEB SITE : USABILITY TESTING VERSUS HEURISTIC EVALUATION

Abstract

The design and production of a Web site require, among other things, to test the usability of the site before putting it online. Two techniques are commonly used to test traditional graphical user interfaces: usability testing and heuristic evaluation. The aim of this study is to compare the two techniques for evaluating Web sites. The study was performed on CRIM's Web site. The results of each technique are presented and compared considering the problems identified (type and quantity) and the relative cost of the techniques. The conclusions drawn here are that heuristic evaluation cannot be used by itself and that usability testing gives superior results especially for uncovering serious problems.

Key words : Web site, usability testing, heuristic evaluation.

INTRODUCTION

Depuis son arrivée au début des années 1990, le Web a connu une croissance exponentielle. Cependant, qui d'entre nous, internautes, n'a pas fait l'expérience frustrante de naviguer sur un site mal conçu, où l'information désirée est quasi impossible à trouver? En effet, bien qu'il soit facile de programmer une page Web, la conception et la réalisation d'un site Web ne sont pas une mince affaire et impliquent notamment de tester l'utilisabilité du site avant de le mettre en ligne. Pour les entreprises qui s'en préoccupent, les questions qui se posent alors sont : « Comment tester l'utilisabilité d'un site? », « Quelle technique donne les meilleurs résultats au moindre coût? ». C'est pour répondre concrètement à cette dernière question que le CRIM a mené une étude comparative de deux techniques d'évaluation de sites Web : les tests d'utilisabilité et l'évaluation heuristique.

PROBLÉMATIQUE

Les tests d'utilisabilité d'une interface consistent à mettre un échantillon représentatif des futurs utilisateurs dans une situation simulée d'utilisation réelle de cette interface. La performance des utilisateurs peut ainsi être quantifiée. Cette technique est relativement coûteuse et demande du temps (conception et réalisation d'une maquette fonctionnelle, élaboration de scénarios, recrutement, passation des tests, dépouillement et traitement des données, analyse, rapport). L'évaluation heuristique, quant à elle, fait appel à un expert qui analyse le produit en fonction de principes d'utilisabilité ou d'ergonomie provenant de la recherche et de la littérature. Cette technique est rapide et peu coûteuse. La littérature (4) (5) (8) préconise de faire appel à plusieurs experts pour accorder une validité à l'évaluation heuristique qui reste une technique relativement subjective. Les études existantes qui comparent spécifiquement les tests d'utilisabilité et les évaluations heuristiques ont été menées sur des interfaces graphiques traditionnelles (1) (3) (4) (5) (6) (8). Elles démontrent que chaque technique a ses forces et ses faiblesses. Nous avons voulu valider ces résultats en ce qui concerne les sites Web. Par ailleurs, dans le cycle du développement d'une interface graphique traditionnelle, l'ergonomie préconise d'utiliser ces deux techniques de façon complémentaire (4) (6). Cependant, compte tenu de l'explosion actuelle de l'offre de sites Web d'une part et de leur courte durée de vie d'autre part, l'évaluation d'un site tend à se résumer à une évaluation heuristique, souvent effectuée par un seul expert. Nous avons voulu également, par cette étude mesurer la validité d'un tel choix, en évaluant le nombre et l'importance des problèmes d'utilisabilité identifiés par les deux techniques.

MÉTHODOLOGIE

Les évaluations pour la comparaison des techniques ont été faites avec le site Web du CRIM (<http://www.crim.ca/>, février 2000). Au total, 14 utilisateurs ont participé aux **tests d'utilisabilité**. L'échantillon d'utilisateurs a été constitué de manière à ce que les participants aux tests représentent la population cible du site (i.e. personnel de recherche, personnel technique, gestionnaires) et les utilisateurs ont été recrutés sur une base bénévole parmi les employés du CRIM, les clients de CRIM Formation et certaines entreprises de Montréal. Quatre scénarios ont été élaborés de manière à couvrir les principales utilisations présumées du site. Deux expérimentateurs ont assisté aux tests pour noter les actions et commentaires des utilisateurs (« Penser tout haut »). Les utilisateurs ont été enregistrés et l'écran d'ordinateur a été filmé.

Pour l'**évaluation heuristique**, trois experts ont été contactés. Ils ont tous reçu les mêmes consignes par le biais d'un document électronique. Ils devaient évaluer 28 pages précises du site Web et les scénarios qui avaient été utilisés pour les tests d'utilisabilité leur ont été fournis de manière à offrir un contexte d'exploration. Les critères d'évaluation choisis pour

cette évaluation étaient ceux de Bastien et Scapin (2).

RÉSULTATS

Le nombre total de problèmes du site Web identifiés à l'aide des deux techniques est de 90. Pour permettre une comparaison des techniques, deux analystes ont classé les problèmes par niveaux de gravité (G4 à G0) selon les critères suivants, traduits et adaptés de Nielsen (8) :

- G4 = Catastrophe d'utilisabilité : impératif à corriger avant la mise en ligne du site.
- G3 = Problème majeur d'utilisabilité : important à corriger, priorité élevée.
- G2 = Problème mineur d'utilisabilité : faible priorité pour sa correction.
- G1 = Problème cosmétique seulement : à régler s'il reste du temps.
- G0 = Ceci n'est pas, selon moi, un problème d'utilisabilité.

Après un classement effectué indépendamment, les deux analystes ont discuté de celui-ci pour en arriver à un consensus et donc à un classement unique.

Les **tests d'utilisabilité** ont permis de découvrir 57 problèmes de différents niveaux de gravité. La figure 1 montre le nombre de problèmes découverts (selon le niveau de gravité) en fonction du nombre de participants aux tests. Pour tous les niveaux de gravité, sauf G2, un plateau existe après les sept ou huit premiers utilisateurs. La participation de 14 utilisateurs a permis d'identifier seulement 10 % de problèmes supplémentaires. Sur l'ensemble des 90 problèmes identifiés dans le site, les tests d'utilisabilité ont permis de trouver 88 % (22) des problèmes majeurs (G3, G4) et 54 % (35) des problèmes mineurs (G1, G2). Les tests d'utilisabilité ont permis, entre autres, d'identifier des problèmes qui étaient de nature à empêcher les utilisateurs de compléter une tâche importante, comme celle de s'inscrire à un cours de CRIM Formation.

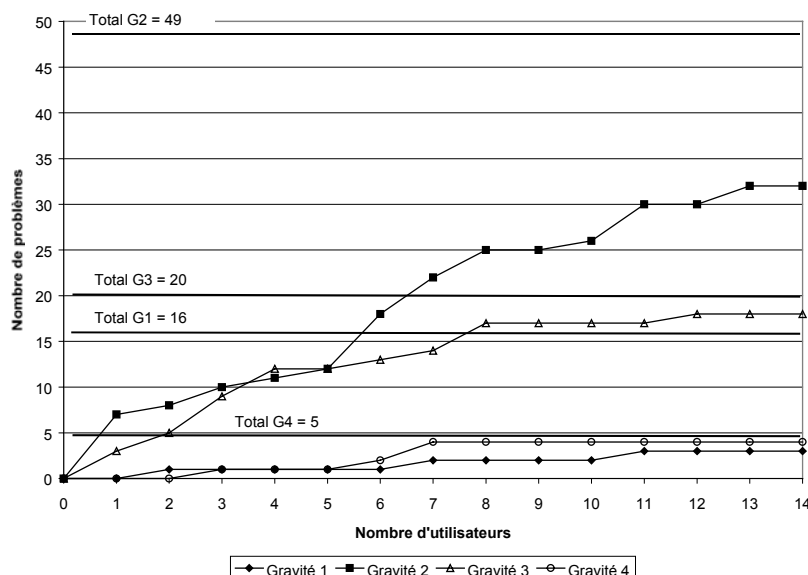


Figure 1 : Nombre de problèmes identifiés à l'aide des tests d'utilisabilité (classés par niveau de gravité : G1, G2, G3, G4) en fonction du nombre d'utilisateurs.

Les **évaluations heuristiques** faites par trois experts ont permis d'identifier 52 problèmes du site Web. La majorité de ces problèmes appartiennent aux critères de guidage (67 %) et de charge de travail (19 %), critères décrits par Bastien et Scapin (2). La majorité des problèmes identifiés (43/52, soit 83 %) étaient différents d'un évaluateur à l'autre (Tableau 1).

Tableau 1 : Nombre de problèmes soulevés selon les évaluateurs (identifiés par A, B et C)

Niveau de gravité	Nombre de problèmes soulevés selon les évaluateurs							Nb total de pb évalué.	Nb total de pb du site	% de pb identifiés par 3 évaluateurs	% de pb (majeur / mineur)
	A	B	C	AB	BC	AC	ABC				
4		4						4	5	80 % (4)	52 % (13/25)
3	2	4	3					9	20	45 % (9)	
2	6	5	7	1	2		3	24	49	49 % (24)	60 % (39/65)
1	5	5	2		1	1	1	15	16	94 % (15)	
Total	13	18	12	1	3	1	4				
Total	43			5			4	52	90	58 % (52)	58 % (52/90)

Sur l'ensemble des 90 problèmes identifiés dans le site, les évaluations heuristiques ont permis d'identifier 52 % (13) des problèmes majeurs (G3, G4) et 60 % (39) des problèmes mineurs (G1, G2). Aucun problème majeur (G3, G4) n'a été identifié par plus d'un évaluateur.

Lors de la comparaison des deux techniques, on constate que celles-ci ont permis de trouver à peu près le même nombre de problèmes : 63 % pour les tests d'utilisabilité et 58 % pour les évaluations heuristiques (Tableau 2). La différence ne se situe pas dans le nombre, mais dans leurs niveaux de gravité. En effet, les tests d'utilisabilité ont permis de trouver deux fois plus de problèmes de niveau G3 que les évaluations heuristiques. Les évaluations heuristiques ont pour leur part permis de trouver cinq fois plus de problèmes de niveau G1 que les tests d'utilisabilité. Une faible proportion des problèmes (21 %) ont été identifiés à la fois à l'aide des tests d'utilisabilité et des évaluations heuristiques. On note toutefois que les problèmes communs aux deux techniques sont essentiellement des problèmes majeurs (G3, G4), proportionnellement trois fois plus nombreux que les problèmes mineurs (G1, G2).

Tableau 2 : Répartition des problèmes communs aux deux techniques

Niveau de gravité	Nb total de pb du site	% de pb tests (14 utilisateurs)	% de pb évalué. (3 évaluateurs)	% de pb communs	% de pb communs (majeur / mineur)
4	5	80 % (4)	80 % (4)	60 % (3)	40 % (10/25)
3	20	90 % (18)	45 % (9)	35 % (7)	
2	49	65 % (32)	49 % (24)	14 % (7)	14 % (9/65)
1	16	19 % (3)	94 % (15)	13 % (2)	
Total	90	63 % (57)	58 % (52)	21 % (19)	21 % (19/90)

Lorsque l'on compare les tests d'utilisabilité à une évaluation heuristique faite par un seul évaluateur, on note la supériorité des tests pour tous les niveaux de gravité de problèmes à l'exception des problèmes de niveau G1 (Tableau 3). Dans ce cas-ci, les tests d'utilisabilité ont permis de trouver onze fois plus de problèmes majeurs (G3, G4) qu'une évaluation heuristique faite par un seul évaluateur. Les tests d'utilisabilité sont également supérieurs lorsque comparés à l'évaluation heuristique faite par deux évaluateurs. Dans le cas présent, plusieurs nouveaux problèmes graves ont cependant été trouvés par le deuxième évaluateur. Les tests d'utilisabilité ont toutefois permis de découvrir deux fois plus de problèmes majeurs (G3, G4) que l'évaluation heuristique. Finalement, lorsque les tests d'utilisabilité sont

comparés à l'évaluation heuristique faite par trois évaluateurs, les tests sont encore supérieurs et révèlent une fois et demie plus de problèmes majeurs (G3, G4) que l'évaluation heuristique.

Tableau 3 : Comparaison des résultats des tests d'utilisabilité versus l'évaluation heuristique

Niveau de gravité	Nb total de pb du site	% de pb tests (14 utilisateurs)	% de pb 1 évaluateur	% de pb 2 évaluateurs	% de pb 3 évaluateurs
4	5	80 % (4)	0 % (0)	80 % (4)	80 % (4)
3	20	90 % (18)	10 % (2)	30 % (6)	45 % (9)
2	49	65 % (32)	20 % (10)	35 % (17)	49 % (24)
1	16	19 % (3)	44 % (7)	81 % (13)	94 % (15)
Total	90	63 % (57)	21 % (19)	44 % (40)	58 % (52)

Cependant, les tests d'utilisabilité prennent beaucoup plus de temps (dans ce cas-ci, 33 jours-personnes) que les évaluations heuristiques (dans ce cas-ci, environ 3 jours-personnes). Globalement, le coût des tests d'utilisabilité a été évalué à dix fois celui des évaluations heuristiques. S'il avait été possible de connaître l'existence du plateau après les huit premiers utilisateurs, les coûts des tests auraient pu être diminués considérablement.

CONCLUSION

Les résultats de cette étude nous permettent de conclure que, dans le cas du site Web testé, les caractéristiques respectives de chacune des deux techniques d'évaluation sont analogues à celles qu'on retrouve dans la littérature concernant des interfaces graphiques traditionnelles. Ils confirment également que les tests d'utilisabilité et les évaluations heuristiques permettent d'identifier des problèmes de nature différente et par conséquent que ces deux techniques sont complémentaires.

Toutefois, notre comparaison met très clairement en évidence que, pour le site Web évalué :

- les tests d'utilisabilité permettent d'identifier des problèmes majeurs d'utilisabilité mais aussi des problèmes d'utilité de l'application (liés à la tâche);
- l'évaluation heuristique est une technique très subjective puisque les problèmes identifiés sont fortement influencés par l'expertise des évaluateurs et qu'il existe une grande variabilité dans les résultats d'un évaluateur à l'autre;
- une évaluation heuristique, même menée par plusieurs experts et a fortiori quand elle est faite par une seule personne, est insuffisante pour évaluer l'utilisabilité d'un site Web.

Cependant, dans le cas de sites Web à la clientèle répartie à travers le globe, il est le plus souvent quasiment impossible de réunir un échantillon représentatif d'utilisateurs pour mener des tests d'utilisabilité. En revanche, l'évaluation heuristique est pour sa part rapide, peu coûteuse et permet de découvrir des problèmes majeurs d'utilisabilité selon les critères ergonomiques adoptés. Sur la base des résultats de cette étude, le CRIM prépare la publication d'une grille pour codifier et guider une évaluation heuristique plus objective des sites Web (7). Une réflexion est également commencée afin de faire évoluer les tests d'utilisabilité en se servant au mieux des nouvelles possibilités de la technologie.

RÉFÉRENCES

- (1) BAILEY, ALLAN, RAIELLO. (1992). Usability testing vs. Heuristic evaluation: A head-to-head comparison. Proceedings of HFS'92, pp. 409-413.
- (2) BASTIEN, J. M. Christian., SCAPIN, Dominique.L. (1993). Ergonomic criteria for the evaluation of human-computer interfaces. RT-0156, INRIA, France.
- (3) DESURVIRE, H., LAWRENCE, D., ATWOOD, M. (1991). Empiricism versus judgement: Comparing user interface evaluation methods on a new telephone-based interface. ACM SIGCHI Bulletin, 23, 4 (October), pp. 58-59.
- (4) JEFFRIES, R., DESURVIRE, H. (1992). Usability testing vs heuristic evaluation : Was there a contest ? SIGCHI Bulletin, 24 (4), pp. 39-41.
- (5) JEFFRIES, R., MILLER, J.R., WHARTON, C., UYEDA, K.M. (1991). User interface evaluation in the real world: A comparison of four techniques. Proceedings of CHI'91, (New Orleans, LA, April 27 – May 2, 1991), ACM, New York, pp. 119-124.
- (6) KARAT, C.-M., CAMPBELL, R., FIEGEL, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. Proceedings of CHI'92, (Monterey, CA, May 3-7, 1992), ACM, New York, pp. 397-404.
- (7) MILLERAND, F. (2001). Guide pratique d'évaluation de sites Web. Centre de recherche informatique de Montréal.
- (8) NIELSEN, J. (1992). Finding usability problems through heuristic evaluation. In Proceedings of CHI '92, (Monterey, CA, May 3-7), ACM, New York, pp. 372-380.
- (9) NIELSEN, J. (1995). Severity ratings for usability problems.
<http://www.useit.com/papers/heuristic/severityrating.html>