Neuro-physiological data as a source of evaluation metrics for personalized IR

Jacek Gwizdka
School of Information, University of Texas at Austin
USA
wepir2018@gwizdka.com

ABSTRACT

The thesis of this position paper is that neuro-physiological (NP) signals can provide metrics useful in evaluating personalization in IR (EPIR). While it should be clear that NP signals provide personalized metrics, it is not clear which higher-level constructs should be considered in EPIR, and, thus, which specific NP measurements could be used. In addition to the "traditional" constructs reflected in measures that use result ranking and levels of information relevance, more recently investigated evaluation metrics include learning, emotions, affect and an overarching concept of user experience. I speculate on the usefulness of some of these higher-level constructs and on their measurement using NP methods and outline the main challenges.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)** → HCI design and evaluation methods → User studies; • Information systems~Users and interactive retrieval

KEYWORDS

Neuro-physiological data, evaluation metrics, personalized IR

1 INTRODUCTION AND MOTIVATION

Neuro-physiological (NP) signals are by their nature personal; they reflect person's brain activity and physiological processes. These processes arise, in part, as a result of changes in cognitive and affective states of a person. Thus, at least in principle, these signals can be used to infer cognitive and affective reaction to and perception of information. NP signals can be collected using a range of different methods and tools. The most common include functional magnetic resonance imaging (fMRI), functional near-infrared spectroscopy (fNIRS), electro-encephalography (EEG), galvanic skin response (GSR), and eye-tracking. Mostafa & Gwizdka argued in [30] for the use of NP evidence as the next frontier in interactive information retrieval (IIR) research. The use of NP measures can allow for a richer consideration of users and their interaction with information. NP measures can capture and help explain phenomena that are verbally not expressible and otherwise not observable. One other reason why NP measures are attractive is the possibility for collecting some of them unobtrusively. Not surprisingly, a number of measures derived from NP signals have been used in IIR. For example, detected by fMRI changes in activation of brain regions in a response to relevance levels [20, 28], and information need [29]; detected by EEG changes in brainwave frequencies in response to search result relevance [34]; changes in eye movement, fixation duration and pupil dilation in relation to relevance levels [19]; changes in reading eye movements in relation to searchers' interest [2, 10]. While these measures were not necessarily used in the context of evaluation, I suggest that since these and other NP measures are personalized they could be useful in evaluating personalization in IR (EPIR). Before discussing them, however, we need to take a step back. While it should be clear that NP signals provide personalized metrics, it is less clear which higher-level constructs should be considered in EPIR, and, thus, which specific NP measurements could be used. Therefore, the first challenge is to select higher-level constructs to use in EPIR, the second is to map the higher-level constructs onto NP measures. In this position paper, I speculate on the usefulness of some of the higher-level constructs and illustrate their measurement using NP methods. My goal is to briefly discuss pertinent aspects of each construct, rather than discussing them in depth. The illustrative NP measures and supporting citations to prior work will be, in the short space of this paper, rather selective.

2 EVALUATION CONSTRUCTS AND NP-METRICS

I suggest and review a few higher-level constructs from the perspective of their potential for use in EPIR and outline their established, or prospective, NP measures.

2.1 Relevance

The notion of relevance is present in most, if not all, "traditional" system and user-centered IR evaluation measures. Measures which involve precision, recall or some aspect of ranking (e.g., MAP@*N*) are based on the concept of document relevance, on counting relevant documents in combination with their presence (or absence) or with their position on the list of retrieved documents. In the context of personalized IR, it is reasonable to suggest that we focus on relevance as a relationship between information objects and some aspects of information need [8, 23,

WEPIR'18, March 15, 2017, New Brunswick, New Jersey USA

DOI:

35] and, on perceived relevance. Perceived relevance is user-centered and (potentially) expressed as multiple aspects of relevance in relation to the person's information need and her/his cognitive and affective state. These aspects can be articulated as, for example, three of the five manifestations of relevance described by Saracevic [33]: Cognitive relevance or Pertinence, Motivational/Affective relevance, and Situational relevance or Utility. Each of these relevance components reflects an aspect of perceived relevance and thus should be considered in EPIR.

The importance of relevance in IIR is reflected in significant interest in inferring relevance from NP measures (e.g., [1–4, 19, 20, 24, 28, 32]). Not surprisingly, many of these research projects focused on incorporating NP measures to improve retrieval (e.g., [6, 9, 17, 18]). Given this applied interest, NP measures of relevance published thus far are correlates of "relevance". It is not clear yet what are the underlying cognitive or affective processes, and, thus, why we observe these relationships. Are we, for example, observing decision making (relevance decision), or searcher's satisfaction from (a part of) information need?

Plausibly, each of the relevance manifestations is reflected in a different NP measure or in their interaction. Investigations which separate manifestations of relevance are very scarce. [6] is an example where thoughtful experimental design was used to separate measurement of cognitive relevance and affective relevance. More studies of relevance and NP measures with careful experimental design are needed.

2.2 Learning

Learning is a good candidate for evaluation construct, since it can be considered as a potential larger goal a person wants to achieve by interacting with a search system. Learning takes place at many levels, from learning a new word [25] to learning new subject. We observe recent interest in our research community in learning in the context of search. This interest has led to several international workshops and two special issues [16, 21] in well-regarded journals.

The majority of IIR research thus far considered observable search behaviors as indicators of learning (e.g., [14]). In cognate research areas, eye-tracking data has been used in reading research to predict comprehension [7, 15, 36], and in education to predict learners' performance [11, 12]. Mapping of learning onto NP measures on IIR tasks is still under-researched. One example includes [13], where differences between domain knowledge levels were found in eye-tracking data.

2.3 User Experience and Affect

User Experience (UX) seems another good candidate construct, albeit one that is multi-dimensional (like relevance) and may be difficult to precisely define. ISO defines UX (quoted after O'Brien [31]): "a person’s perceptions and responses that result from the use or anticipated use of a product, system or service". A definition from an HCI textbook: "User experience is the totality of the effect or effects felt (experienced) internally by a user as a result of interaction with, and the usage context of, a system, device, or product" [22]. The important components of user experience that have been historically under-considered in IIR but have received more focused interest in last decade [26] are affect and emotional reactions. These constructs should be of interest to EPIR for two reasons. First, they help to assess whether PIR makes searchers "happy". Second, they tap into one of the relevance manifestations – the Motivational/Affective relevance.

On IIR tasks, affect has been measured, for example, using galvanic skin response [6] as well as a combination of facial expressions, heart-rate variability, and galvanic skin response [5, 27].

3 CHALLENGES

The two grand challenges are to select higher-level constructs for EPIR and map them onto NP measures. The multi-dimensionality of constructs makes the mapping particularly challenging. Furthermore, one of the main benefits of NP measures, their "personal nature", makes their scales and also mappings (potentially) specific to a person. For example, [9] showed that scales of eye-tracking measures need to be "personalized" to improve relevance classification, while [37] applied similar procedure to EEG signals. The individual differences in mappings of constructs onto NP measures have not been addressed in IIR thus far. Their potential existence points to the need for introducing a "calibration" phase in interactive IR systems, whereby NP measures would be selected separately for each system user.

4 CONCLUSIONS

My aim in this position paper was to draw attention to personalized-by-definition NP measures as a potentially interesting and useful source of evidence for EPIR. In doing so, I outlined some possibilities and challenges associated with them. I believe that the use of NP measures offers a fruitful avenue in PIR, and that there is a lot of interesting work ahead of us on this path.

REFERENCES

[1] Ajanki, A. 2013. Inference of relevance for proactive information retrieval. (2013).

[2] Ajanki, A., Hardoon, D., Kaski, S., Puolamäki, K. and Shawe-Taylor, J. 2009. Can eyes reveal interest? Implicit queries from gaze patterns. *User Modeling and User-Adapted Interaction*. 19, 4 (2009), 307–339. DOI:https://doi.org/10.1007/s11257-009-9066-4.

[3] Allegretti, M., Moshfeghi, Y., Hadjigeorgieva, M., Pollick, F.E., Jose, J.M. and Pasi, G. 2015. When Relevance Judgement is Happening?: An EEG-based Study. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2015), 719–722.

[4] Arapakis, I., Athanasakos, K. and Jose, J.M. 2010. A comparison of general vs personalised affective models for the prediction of topical relevance. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2010), 371–378.

[5] Arapakis, I., Konstas, I. and Jose, J.M. 2009. Using Facial Expressions and Peripheral Physiological Signals As Implicit Indicators of Topical Relevance. *Proceedings of the 17th ACM International Conference on Multimedia* (New York, NY, USA, 2009), 461–470.

[6] Barral, O., Kosunen, I., Ruotsalo, T., Spapé, M.M., Eugster, M.J.A., Ravaja, N., Kaski, S. and Jacucci, G. 2016. Extracting relevance and affect information from physiological text annotation. *User Modeling and User-Adapted Interaction*. 26, 5 (Dec. 2016), 493–520. DOI:https://doi.org/10.1007/s11257-016-9184-8.

[7] Biedert, R., Dengel, A., Elshamy, M. and Buscher, G. 2012. Towards robust gaze-based objective quality measures for text. *Proceedings of the Symposium on*

Eye Tracking Research and Applications (New York, NY, USA, 2012), 201–204.

[8] Borlund, P. 2003. The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*. 8, 3 (2003), paper no. 152.

[9] Buscher, G., Dengel, A., Biedert, R. and Elst, L.V. 2012. Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. *ACM Trans. Interact. Intell. Syst.* 1, 2 (2012), 9:1–9:30. DOI:https://doi.org/10.1145/2070719.2070722.

[10] Buscher, G., Dengel, A. and van Elst, L. 2008. Eye movements as implicit relevance feedback. *CHI '08 extended abstracts on Human factors in computing systems* (New York, NY, USA, 2008), 2991–2996.

[11] Chen, S.-C., She, H.-C., Chuang, M.-H., Wu, J.-Y., Tsai, J.-L. and Jung, T.-P. 2014. Eye movements predict students' computer-based assessment performance of physics concepts in different presentation modalities. *Computers & Education*. 74, (May 2014), 61–72. DOI:https://doi.org/10.1016/j.compedu.2013.12.012.

[12] Chien, K.-P., Tsai, C.-Y., Chen, H.-L., Chang, W.-H. and Chen, S. 2015. Learning differences and eye fixation patterns in virtual and physical science laboratories. *Computers & Education*. 82, (Mar. 2015), 191–201. DOI:https://doi.org/10.1016/j.compedu.2014.11.023.

[13] Cole, M.J., Gwizdka, J., Liu, C., Belkin, N.J. and Zhang, X. 2013. Inferring user knowledge level from eye movement patterns. *Information Processing & Management*. 49, 5 (Sep. 2013), 1075–1091. DOI:https://doi.org/10.1016/j.ipm.2012.08.004.

[14] Collins-Thompson, K., Rieh, S.Y., Haynes, C.C. and Syed, R. 2016. Assessing Learning Outcomes in Web Search: A Comparison of Tasks and Query Strategies. *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (New York, NY, USA, 2016), 163–172.

[15] Copeland, L., Gedeon, T. and Mendis, S. 2014. Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error. *Artificial Intelligence Research*. 3, 3 (Aug. 2014), 35. DOI:https://doi.org/10.5430/air.v3n3p35.

[16] Eickhoff, C., Gwizdka, J., Hauff, C. and He, J. 2017. Introduction to the special issue on search as learning. *Information Retrieval Journal*. (Sep. 2017), 1–4. DOI:https://doi.org/10.1007/s10791-017-9315-9.

[17] Eugster, M.J.A., Ruotsalo, T., Spapé, M.M., Barral, O., Ravaja, N., Jacucci, G. and Kaski, S. 2016. Natural brain-information interfaces: Recommending information by relevance inferred from human brain signals. *Scientific Reports*. 6, (Dec. 2016), 38580. DOI:https://doi.org/10.1038/srep38580.

[18] Eugster, M.J.A., Ruotsalo, T., Spapé, M.M., Kosunen, I., Barral, O., Ravaja, N., Jacucci, G. and Kaski, S. 2014. Predicting Term-relevance from Brain Signals. *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval* (New York, NY, USA, 2014), 425–434.

[19] Gwizdka, J. 2014. Characterizing Relevance with Eye-tracking Measures. *Proceedings of the 5th Information Interaction in Context Symposium* (New York, NY, USA, 2014), 58–67.

[20] Gwizdka, J. 2013. Looking for Information Relevance In the Brain. *Gmunden Retreat on NeuroIS 2013* (Gmunden, Austria, Jun. 2013), 14.

[21] Hansen, P. and Rieh, S.Y. 2016. Editorial: Recent advances on searching as learning: An introduction to the special issue. *Journal of Information Science*. 42, 1 (Feb. 2016), 3–6. DOI:https://doi.org/10.1177/0165551515614473.

[22] Hartson, R. and Pyla, P.S. 2012. *The UX Book: Process and Guidelines for Ensuring a Quality User Experience*. Elsevier.

[23] Huang, X. and Soergel, D. 2013. Relevance: An improved framework for explicating the notion. *Journal of the American Society for Information Science and Technology*. 64, 1 (2013), 18–35. DOI:https://doi.org/10.1002/asi.22811.

[24] Klami, A., Saunders, C., de Campos, T.E. and Kaski, S. 2008. Can relevance of images be inferred from eye movements? *Proceedings of the 1st ACM international conference on Multimedia information retrieval* (New York, NY, USA, 2008), 134–140.

[25] Kraiger, K., Ford, J.K. and Salas, E. 1993. Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*. 78, 2 (1993), 311–328. DOI:https://doi.org/10.1037/0021-9010.78.2.311.

[26] Lopatovska, I. and Arapakis, I. 2011. Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction. *Information Processing & Management*. 47, 4 (Jul. 2011), 575–592. DOI:https://doi.org/16/j.ipm.2010.09.001.

[27] Moshfeghi, Y. and Jose, J.M. 2013. An effective implicit relevance feedback technique using affective, physiological and behavioural features. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2013), 133–142.

[28] Moshfeghi, Y., Pinto, L.R., Pollick, F.E. and Jose, J.M. 2013. Understanding Relevance: An fMRI Study. *Advances in Information Retrieval*. P. Serdyukov, P. Braslavski, S.O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz, eds. Springer Berlin Heidelberg. 14–25.

[29] Moshfeghi, Y., Triantafillou, P. and Pollick, F.E. 2016. Understanding Information Need: An fMRI Study. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2016), 335–344.

[30] Mostafa, J. and Gwizdka, J. 2016. Deepening the Role of the User: Neuro-Physiological Evidence As a Basis for Studying and Improving Search. *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (New York, NY, USA, 2016), 63–70.

[31] O'Brien, H.L. 2011. Chapter 4 Weaving the Threads of Experience into Human Information Interaction (HII): Probing User Experience (UX) for New Directions in Information Behaviour. *New Directions in Information Behaviour*. Emerald Group Publishing. 69–92.

[32] Salojärvi, J., Puolamäki, K. and Kaski, S. 2005. Implicit Relevance Feedback from Eye Movements. *Artificial Neural Networks: Biological Inspirations – ICANN 2005*. W. Duch, J. Kacprzyk, E. Oja, and S. Zadrożny, eds. Springer Berlin Heidelberg. 513–518.

[33] Saracevic, T. 1996. Relevance Reconsidered 1996. *Proceedings of the second conference on conceptions of library and information science (CoLIS 2)* (1996).

[34] Scharinger, C., Kammerer, Y. and Gerjets, P. 2016. Fixation-Related EEG Frequency Band Power Analysis: A Promising Neuro-Cognitive Methodology to Evaluate the Matching-Quality of Web Search Results? *HCI International 2016 – Posters' Extended Abstracts*. C. Stephanidis, ed. Springer International Publishing. 245–250.

[35] Toms, E.G., O'Brien, H.L., Kopak, R. and Freund, L. 2005. Searching for Relevance in the Relevance of Search. *Context: Nature, Impact, and Role*. F. Crestani and I. Ruthven, eds. Springer Berlin Heidelberg. 59–78.

[36] Vo, T., Mendis, B.S.U. and Gedeon, T. 2010. Gaze Pattern and Reading Comprehension. *Neural Information Processing. Models and Applications* (Nov. 2010), 124–131.

[37] Wang, S., Gwizdka, J. and Chaovalitwongse, W.A. 2016. Using Wireless EEG Signals to Assess Memory Workload in the n-Back Task. *IEEE Transactions on Human-Machine Systems*. 46, 3 (Jun. 2016), 424–435. DOI:https://doi.org/10.1109/THMS.2015.2476818.