Predicting Perceptual Speed from Search Behaviour

Olivia Foulds
olivia.foulds@strath.ac.uk
University of Strathclyde
Glasgow, Scotland

Alessandro Suglia
as247@hw.ac.uk
Edinburgh Centre for
Robotics, Scotland

Leif Azzopardi
leifos@acm.org
University of Strathclyde
Glasgow, Scotland

Martin Halvey
martin.halvey@strath.ac.uk
University of Strathclyde
Glasgow, Scotland

ABSTRACT

Perceptual Speed (PS) is a cognitive ability that is known to affect multiple factors in Information Retrieval (IR) such as a user's search performance and subjective experience. However PS tests are difficult to administer which limits the design of user-adaptive systems that can automatically infer PS to appropriately accommodate low PS users. Consequently, this paper evaluated whether PS can be automatically classified from search behaviour using several machine learning models trained on features extracted from TREC Common Core search task logs. Our results are encouraging: given a user's interactions from one query, a Decision Tree was able to predict a user's PS as low or high with 86% accuracy. Additionally, we identified different behavioural components for specific PS tests, implying that each PS test measures different aspects of a person's cognitive ability. These findings motivate further work for how best to design search systems that can adapt to individual differences.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI)*; • **Computing methodologies** → **Machine learning**; • **Information systems** → **Users and interactive retrieval**.

ACM Reference Format:
Olivia Foulds, Alessandro Suglia, Leif Azzopardi, and Martin Halvey. 2020. Predicting Perceptual Speed from Search Behaviour. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401210

1 INTRODUCTION

Information retrieval (IR) is a complex activity involving *human-computer interaction* (HCI) where users issue queries to search for and find relevant information depending on their task needs [2]. When conducting online searches, a lot of information is visibly presented to a user. As such, differences in a user's individual cognitive ability for processing visual stimuli are known to affect performance. Specifically, *Perceptual Speed* (PS) has been shown to be an influential factor affecting information-seeking [2, 11]. PS is defined by an individual's ability to accurately view, scan, and compare visual information that is presented to them [5]. Many tests that attempt to measure PS have been around for over 50 years, where the generalised format involves identifying how fast and accurately an individual can identify certain targets in a visual search task: `high` PS scorers successfully complete the task in the shortest amount of time with the fewest mistakes; whereas `low` PS scorers take longer and have poorer performance [10].

In IR, PS has been shown to significantly affect many parts of the search process. For example, compared to `high` PS, people with `low` PS have: engaged in less search activity through issuing shorter queries, clicking on fewer results and thus viewing fewer documents [8]; reported a more negative user experience, greater self-reported workload, perceived interfaces as less usable, and felt less satisfied with their search [8, 22]; preferred data to be visualised differently [9, 15]; and completed tasks with poorer performance, taking longer [2, 22] while learning less [5].

To understand why `low` PS users have such a negative search experience, a study was conducted where search interfaces with more components were considered less usable, more distracting and confusing to users with `low` PS – resulting in lower user engagement [22]. This makes sense considering other research has identified that people with `low` PS have a lower eye fixation rate, and so struggle with scanning what is in front of them [20, 21]. Consequently, these authors have stated the need for systems that can infer PS and subsequently adapt to help `low` PS scorers achieve a better search experience, for example, by adding tools such as highlighting or providing more space in less cluttered interfaces. These adaptations could theoretically allow `low` PS users to better navigate search results with less visual scanning required, while not reducing the information density for `high` PS users, which could lead to a degraded search experience.

However, current PS tests can take up to 20 minutes to perform [1]. This may be reasonable for a user to complete if their cognitive abilities are assumed to be stable over time [19]. Yet, other research has found that cognitive abilities are not stable, and instead can be affected by environmental factors such as tiredness levels or depressive symptoms [7, 16]. This would imply that PS is a changeable ability, requiring regular re-testing. Accordingly, it would be preferable if PS levels could be inferred automatically, from people's interactions, which would enable systems to adapt according to the user's current PS levels.

Despite the importance of PS levels in IR and HCI, there have been few attempts to infer the PS levels of users given their interactions with the system. In those studies [9, 20], eye gaze data from users undertaking information visualisation search tasks was used to build various Machine Learning (ML) models to predict whether the user had `low` or `high` levels based on one PS test. Having achieved the best accuracy scores of 57.1% in [20], and 60.6% in [9], this suggests that predicting PS levels from eye-gaze interaction is difficult. Furthermore, due to additional costly equipment that is sometimes physically uncomfortable, eye-tracking is hard

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401210

to deploy at scale, which limits its applicability in large scale user-adaptive systems.

Consequently, detecting a user's PS level from other forms of interaction may be preferable. For example, how many documents are clicked on and how long they are viewed for may be strong indicators of PS levels as it has been shown that low PS users take longer completing search tasks [2, 22] and interact with search systems less [8]. Thus, it seems possible that search behaviour could be used to infer PS level. If this was the case then a user's online browsing patterns should allow systems to alter the visual display to dynamically accommodate individual abilities. Therefore, in this study we explore the following research questions in the context of information seeking:

(1) Given a user's search behaviour, can we accurately predict their perceptual speed?
(2) And, if so, what behaviours are most informative?

2 METHOD

We obtained search logs from a user study conducted on Prolific where 38 users undertook search tasks from the *TREC Common Core 2017* track [3]. After a practice task to familiarise users with the system, Topics 341, 347, 408, and 435 comprised the main search tasks and were counterbalanced in accordance with a previous IR study [17]. Users were given up to 8 minutes to find as many different and relevant articles that helped them learn about the given topic. A standard web search interface was used where users could query, inspect up to 10 result summaries per page, view articles, and bookmark them as relevant. The underlying retrieval system was composed from the *Whoosh* IR toolkit[1] with the *BM25* retrieval algorithm ($\beta = 0.75$). To imitate standard news article pages, graphical advertisements sourced from the *Ads of the World* database[2] were randomly allocated onto articles as a top and bottom banner, and four right-railed ads. Users were native English speakers (23M, 14F, and 1 other) ranging in age from 18 to 58 with a mean of 32 years, who received roughly US$13 compensation for participating.

While the logs contained many details, including user performance statistics, here, we only focus on behaviour. This allows for wider application to many different search tasks based on a user issuing one query. By focusing on per-query, differently from previous research using session-based information [9], a new user would only have to issue one query to determine their PS, as opposed to completing an entire search session. We therefore used the following behavioural metrics associated with 575 queries: *Length of query; № of words per query; Time spent issuing the query; Time on search engine result page (SERP); Mean time per result snippet; Total time viewing articles; Mean time per article; Total time of entire search session; № of unique articles clicked/viewed; № of SERPs viewed; Estimated № of results inspected; Depth of the last result snippet clicked in the SERP; Depth of the last result snippet hovered over in the SERP; № of mouse hovers over result snippets; № of mouse hovers over unique result snippets; № of mouse hovers over all adverts (ads), top-positioned ads, bottom-positioned ads, and side-positioned ads; № of ads clicked; and № of user-triggered events during a query session (E.g. № of articles/ads clicked, hovered etc.).*

[1] https://pypi.org/project/Whoosh/ – last accessed January, 2020.
[2] See https://www.adsoftheworld.com/ – last accessed January, 2020.

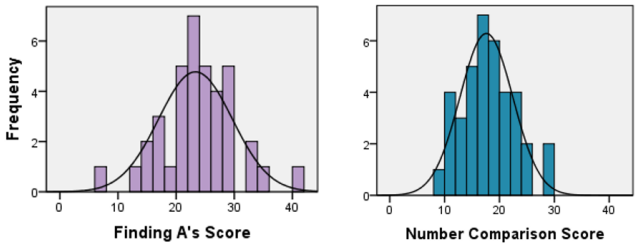

Figure 1: Individual scores for both Perceptual Speed tests

2.1 Perceptual Speed Tests

For each user, we also had their corresponding scores from two computerised PS tests, which were administered in the same session as the search tasks and based on *Ekstrom's Kit of Factor Referenced Cognitive Tests* [10].

Finding A's (FA). Users searched for two minutes through lists of words identifying any that contained the letter 'a'.

Number Comparison (NC). Users had two minutes to inspect pairs of numbers, presented in a list, and select non-identical pairs.

In both *FA* and *NC*, a user's score was the number of correctly identified targets, minus how many they incorrectly identified. Figure 1 shows the normal distribution of scores for each test. For *FA*, scores ranged from 7-40 with mean 23.32 ($\sigma = 6.351$), and median 23. For *NC*, scores ranged from 9-28 with mean 17.61 ($\sigma = 4.824$), and median 17. The Pearson correlation coefficient between *FA* and *NC* was 0.047. Similar to [4], this suggests that they were measuring two different aspects of PS. We therefore ensured that low and high PS scorers were calculated separately for each test. Previous research divided users into low and high PS groups based on a median split [8, 22]. Subsequently, for *FA*, any user with a score of 23 or below was coded as low (20) and the rest were coded as high (18), while for *NC* users, any with the score of 17 or below were coded as low (20) and the rest were coded as high (18).

In addition, guidelines for administering PS tests state that at least two PS tests should be performed to determine a valid PS measure [10]. However, guidance is lacking for how to combine scores [11]. We therefore developed our own coding system to derive an **Overall Perceptual Speed (*OP*)** measure: A user was low if they were low on both PS tests, medium if they were low on one but not the other, and high if they did not score low on either. This resulted in 10-low , 20-medium, and 8-high users.

2.2 Models

To perform the classification tasks (i.e. predict low/high on *FA* and *NC*, and predict low/medium/high on *OP*), we employed several standard machine learning models as per [9, 20]: Support Vector Machine (SVM), Decision Tree, Random Forest, Logistic Regression, Neural Networks and a Majority Class baseline. For the Neural Network model, we experimented with three different architectures, but only present the best results which came from the 1-layer feed-forward network with 32 hidden units and ReLU activation function [18]. We used a softmax activation function in the last layer of the network to obtain a probability distribution over the class

labels. We trained the model for 50 epochs using the ADAM [14] optimiser with batch size 32 minimising the cross entropy loss. Network weights were initialised using the Glorot Uniform initialisation scheme [12].

For all the models, due to the reduced number of data points, we applied k-fold cross validation ($k = 5$) to obtain a better estimate of the model performance. Accuracy on the test set was our evaluation measure for the classification task. Features were also scaled uniformly to improve the convergence rate of the learning methods. The Neural Network was implemented using Keras [3] whereas all the others were implemented using scikit-learn [4]. Default model parameters were used, unless otherwise specified.

3 RESULTS

Table 1 provides the accuracy achieved by each model for each classification task. While most models outperformed the baseline, the Decision Tree had the best performance with accuracies of 74.96% on the ***Finding A's* (*FA*)** task, and 57.39% on ***Overall Perceptual Speed* (*OP*)** task, whereas the the ***Number Comparison* (*NC*)** task achieved it's best accuracy of 61.74% from a Neural Network.

To identify which behaviours contributed the most towards the class predictions for each task, we applied a feature selection algorithm based on mutual information [5]. Table 2 shows the five main predictive features for *FA*, *NC*, and *OP*. From this, we can see that the number of user-triggered events was important for all classification tasks, while the result hover depth, number of articles clicked on, and number of side-ads hovered over were important for at least two of the tasks. Additionally, different features correspond to each dataset. For example, time measures appear important for *NC*, but not *FA* or *OP* .

We then re-ran all models using only the top five features for each dataset, also shown in Table 1. While the accuracy of some models degraded (e.g. Logistic Regression), the performance for Random Forest and Decision Trees increased considerably. The Decision Tree consistently delivered the best performance with 86.09% accuracy on *FA* , 83.30% for *NC* , and 69.57% for *OP* .

We further analysed the data looking for possible patterns between the PS class values and the five most relevant features. For each feature, we computed the mean differences between classes, and identified the following behaviours:

Finding A's. Compared to `high` PS, `low` PS users triggered more events, inspected more results, clicked on more articles, had greater result hover depth, and hovered over more ads at the side.

Number Comparison. Compared to `high` PS, `low` PS users hovered over more side-ads but less top-ads, spent longer examining each snippet and article, but overall actually triggered less events.

Overall PS. Again in comparison to `high` PS users, `low` PS clicked on more articles, had greater result hover depth, hovered over more bottom-ads, but overall hovered over less ads resulting in less overall user-triggered events. Yet, `medium` PS users triggered the most events, hovered over the most bottom and overall ads, and were in between `low` and `high` PS for article clicks and result hover depth.

[3] https://keras.io/
[4] https://scikit-learn.org/
[5] As described in: https://tinyurl.com/mutual-info

4 DISCUSSION AND FUTURE WORK

This paper aimed to explore whether Perceptual Speed (PS), which affects information-seeking, can be predicted from people's search behaviour, and if possible, identify which features of search behaviour are most informative. Our results showed that we could classify the PS levels of our users with accuracy scores up to 86.1% depending on the type of PS test and model. These findings are particularly promising as they are substantial improvements over the majority class baselines of 54.3% and 50.1%. In previous attempts only minor improvements over the majority class baseline were reported [9, 20] – while it is not possible to directly compare studies as prior work was performed on different data – it is worth noting that our work only used behavioural log data rather than eye-tracking data. A clear benefit of using log-based features is that it can be deployed more widely, and at scale. However, it would be interesting to explore how combining the log-based features and eye-gaze data could be utilised to further improve accuracy.

Given the potential to accurately classify the PS of users, this work paves the way for further investigations into PS and IR. If the PS levels of users can be inferred from search behaviour, then an open question that arises is: "how do we tailor and adapt IR search interfaces to provide the best user experience and search performance to users with varying cognitive abilities?". More research is clearly required as currently, only suggestions for how to accommodate different PS levels have been stated, such as highlighting tools and different data visualisations [20].

In this work, we have tried to infer the PS levels of users given two different PS tests. One of the interesting findings was that accurately classifying users into `low` or `high` for these tests required very different features. For example, measures of time were not helpful in inferring PS levels on the *FA* test, but time spent examining result snippets and articles were useful in inferring PS levels on the *NC* test. Additionally, although the number of user-triggered events was a main feature for both individual PS tests, the opposite pattern occurred where `low` PS in *FA* have *more* interaction, yet `low` PS users in *NC* have *less* interaction. This further suggests that different interface adaptations may be required for different types of PS aspects, depending on the task required. It also reinforces the suggestions in the guidelines for administering multiple PS tests [10, 11]. Thus, more work is required to investigate what types of PS tests are most appropriate in the context of IR.

Our work, however, is not without limitations. Firstly, our sample size was relatively small reducing both the ability to draw more reliable conclusions as well as training classifiers that better generalise. Secondly, our analysis was based on data from one particular search task (i.e. topic search) within the context of news (i.e. TREC Common Core) using a standard web search interface. Additionally, we trained our models based on query log data recorded in isolation and not in the context of an entire search session. Given that search behaviours can change over time [6], it may be possible to develop models which exploit this temporal element and the dependency between interactions (e.g. via Recurrent Neural Networks [13]) to improve PS predictions. Finally, although we used more PS tests than previous works, different PS tests measure different aspects of PS [4], and there are many other PS tests that we did not consider. Therefore, it is worth exploring how well other aspects of PS can

Table 1: Accuracy values (%) for each of the classification tasks and PS tests. Bold text indicates the highest score for each model. The Decision Tree generally performs the best overall.

Perceptual Speed	PS Class	Model Attributes	Baseline	Neural Network	SVM	Decision Tree	Random Forest	Logistic Regression
Finding A's	***Low/High***	**All features**	50.09	58.09	58.43	**74.96**	60.70	51.48
		5 main	50.09	54.43	56.52	**86.09**	71.13	47.83
Number Comparison	***Low/High***	**All features**	50.61	**61.74**	60.00	59.83	59.13	60.00
		5 main	50.61	61.91	63.83	**83.30**	73.39	55.30
Overall	***Low/Med/High***	**All features**	54.26	53.74	56.00	**57.39**	51.48	54.78
		5 main	54.26	52.70	54.61	**69.57**	59.30	53.91

Table 2: The top 5 behaviours (where 1 is most informative) for predicting Perceptual Speed in Finding A's, Number Comparison, and Overall Perceptual Speed.

	FA	NC	OP
Total sum of all user-triggered events	1	1	1
Number of unique articles clicked/viewed	5	*n/a*	2
Depth of last result snippet hover in SERP	2	*n/a*	4
Mean time spent per article	*n/a*	3	*n/a*
Mean time spent per result snippet	*n/a*	4	*n/a*
Estimated number of results inspected	3	n/a	*n/a*
Number of hovers over top-positioned ads	*n/a*	5	*n/a*
Number of hovers over bottom-positioned ads	*n/a*	*n/a*	3
Number of hovers over side-positioned ads	4	2	*n/a*
Number of hovers over all ads	*n/a*	*n/a*	5

be predicted from other PS tests, and whether these aspects may relate to different search behaviours. In summary, more research is required, with larger numbers of participants, in a variety of different search contexts, with other PS Tests, so that these findings can be further generalised.

In conclusion, we have shown that a user's Perceptual Speed classification can be predicted, with reasonably high accuracy, from search behaviour. Furthermore, different behaviours appear to correspond to different PS abilities. Our findings are highly encouraging and point to the need for a number of different lines for future research in pursuit of developing dynamic search interfaces and systems that are tailored and adapted to individual cognitive abilities.

ACKNOWLEDGMENTS

Data underpinning this publication will be available from The University of Strathclyde Research Information Portal at https://doi.org/10.15129/d6acf3d1-d64c-4cdd-9b41-e7edb5709f37. The authors would like to thank Dr David Maxwell for his contribution towards the search system developed for the user study and the anonymous reviewers for their helpful comments. This work was part funded by BAE Systems Maritime and EPSRC as part of an Industrial Cooperative Award in Science & Technology (CASE) Studentship (EP/S513908/1).

REFERENCES

[1] P. Ackerman and M. Beier. 2007. Further explorations of perceptual speed abilities in the context of assessment methods, cognitive abilities, and individual differences during skill acquisition. *J. of Exp. Psych.: Applied* 13, 4 (2007), 249.
[2] A. Al-Maskari and M. Sanderson. 2011. The effect of user characteristics on search effectiveness in information retrieval. *Information Processing & Management* 47, 5 (2011), 719–729.
[3] J. Allan, D. Harman, E. Kanoulas, D. Li, C. Van Gysel, and E. Voorhees. 2017. TREC 2017 Common Core Track Overview.. In *TREC*.
[4] B. Allen. 1994. Cognitive abilities and information system usability. *Information processing & management* 30, 2 (1994), 177–191.
[5] B. Allen. 1994. Perceptual speed, learning and information retrieval performance. In *Proc. of the 17th ACM SIGIR Conf.* 71–80.
[6] F. Baskaya, H. Keskustalo, and K. Järvelin. 2013. Modeling behavioral factors in interactive information retrieval. In *Proc. of the 22nd ACM CIKM Conf.* 2297–2302.
[7] A. Beaujean, S. Parker, and X. Qiu. 2013. The relationship between cognitive ability and depression: a longitudinal data analysis. *Social psychiatry and psychiatric epidemiology* 48, 12 (2013), 1983–1992.
[8] K. Brennan, D. Kelly, and J. Arguello. 2014. The effect of cognitive abilities on information search for tasks of varying levels of complexity. In *Proc. of the 5th Information Interaction in Context Symposium*. 165–174.
[9] C. Conati, S. Lallé, M. A. Rahman, and D. Toker. 2017. Further Results on Predicting Cognitive Abilities for Adaptive Visualizations.. In *Proc. of the 26th IJCAI Conf.* 1568–1574.
[10] R. Ekstrom, J. French, H. Harman, and D. Dermen. 1976. *Manual for kit of factor-referenced cognitive tests*. Vol. 102. Educational testing service Princeton.
[11] O. Foulds, L. Azzopardi, and M. Halvey. 2020. Reflecting upon perceptual speed tests in information retrieval: limitations, challenges, and recommendations. In *Proc. of the 5th ACM SIGIR CHIIR Conf.* 234–242.
[12] X. Glorot and Y. Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proc. of the 13th AISTATS Conf.* 249–256.
[13] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint:1511.06939* (2015).
[14] D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint:1412.6980* (2014).
[15] S. Lallé, C. Conati, and G. Carenini. 2017. Impact of Individual Differences on User Experience with a Real-World Visualization Interface for Public Engagement. In *Proc. of the 25th UMAP Conf.* 369–370.
[16] M. Lyons, T. York, C. Franz, M. Grant, L. Eaves, K. Jacobson, K. W. Schaie, M. Panizzon, C. Boake, H. Xian, R. Toomey, S. Eisen, and W. Kremen. 2009. Genes determine stability and the environment determines change in cognitive ability during 35 years of adulthood. *Psychological Science* 20, 9 (2009), 1146–1152.
[17] D. Maxwell, L. Azzopardi, and Y. Moshfeghi. 2019. The impact of result diversification on search behaviour and performance. *IRJ* 22, 5 (2019), 422–446.
[18] V. Nair and G. E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proc. of the 27th ICML Conf.* 807–814.
[19] R. Palmquist and K-S Kim. 2000. Cognitive style and on-line database search experience as predictors of Web search performance. *Journal of the American society for information science* 51, 6 (2000), 558–566.
[20] B. Steichen, C. Conati, and G. Carenini. 2014. Inferring visualization task properties, user performance, and user cognitive abilities from eye gaze data. *ACM TiiS* 4, 2 (2014), 1–29.
[21] D. Toker, C. Conati, B. Steichen, and G. Carenini. 2013. Individual user characteristics and information visualization: connecting the dots through eye tracking. In *Proc. of the SIGCHI CHI Conf.* 295–304.
[22] L. Turpin, D. Kelly, and J. Arguello. 2016. To blend or not to blend? Perceptual speed, visual memory and aggregated search. In *Proc. of the 39th ACM SIGIR Conf.* 1021–1024.