Lernen zu Löschen: Vergessen digitaler Objekte als Gemeinschaftsaufgabe von Mensch und KI

Beitrag der Informatik

(Weiter-) Entwicklung eines partnerschaftlichen KI-Systems, das unterstützt, irrelevante Dateien zu identifizieren und zu löschen. Mit Induktiver Logischer Programmierung (ILP) werden wissensbasierte KI und maschinelles Lernen integriert. Lernen wird erklärungsbasiert und interaktiv umgesetzt, so dass der Mensch Systementscheidungen nachvollziehen und korrigieren kann.

Mensch-KI-Partnerschaft durch erklärbares und interaktives maschinelles Lernen

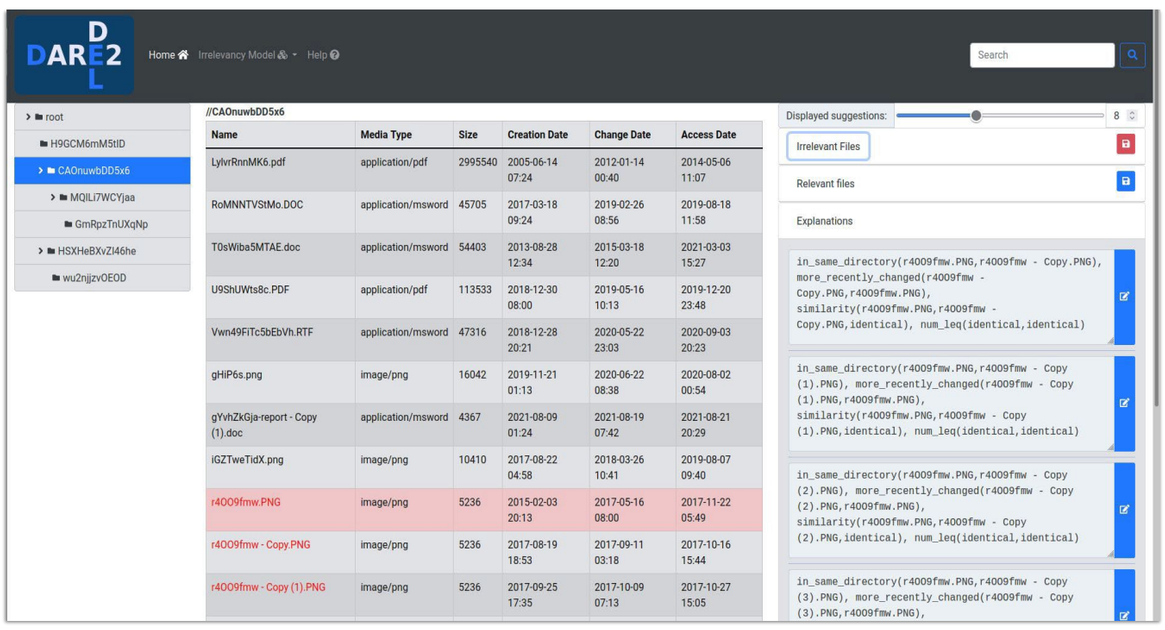

Abbildung: Datei *r4009fmw.png* wäre nicht als irrelevant klassifiziert, wenn sie im Vergleich zu *r4009fmw-Copy.png* neuer wäre.

Dare2Del nutzt

- **Metainformationen von Dateien**
- **domänenspezifisches Wissen und Hintergrundtheorie zum Lernen von Irrelevanzentscheidungen**

Dare2Del liefert Nutzenden Löschvorschläge

- **die akzeptiert oder abgelehnt werden können**

Dare2Del liefert Erklärungen für Irrelevanzentscheidungen

- **die korrigierbar sind**

Beitrag der Psychologie

Löschen soll so gestaltet und unterstützt werden, dass Personen sich tatsächlich damit auseinandersetzen. Dadurch sollen sich positive Konsequenzen sowohl für eigene als auch für organisationale Arbeitsabläufe ergeben.

Forschungsfragen Psychologie

- **Welche Gestaltungsmerkmale (z.B. Erklärungen) muss das Assistenzsystem Dare2Del haben, um bei Löschentscheidungen zu unterstützen?**
- **Inwieweit kann menschliches Vergessen durch Löschen unterstützt werden?**
- **Welche personenbezogenen Fähigkeiten (z.B. Fähigkeit zur Inhibition) gilt es zu beachten?**
- **Welche organisationalen Bedingungen (z.B. soziale Normen) haben einen Einfluss?**

Ergebnisse der zweiten Projektphase

Psychologie

AP-P-1: Experimente zur Gestaltung der Löschvorschläge (mit vs. ohne Erklärungen)

- **Erklärungen führen zu einer höheren Löschbereitschaft und helfen, das System glaubwürdiger und vertrauenswürdiger zu machen – vor allem, wenn sie überprüfbar sind**
- **Erklärungen reaktivieren kurzfristig aber auch Dokumentnamen und -inhalte (kein Vergessen)**

AP-P-2: Experimente zu den Auswirkungen des Löschens mit und ohne Assistenzsystem

- **Löschvorschläge des Assistenzsystems werden von etwa 50% der Nutzenden angenommen**
- **Die Vorgabe sozialer Normen zum Löschverhalten hat vergleichbare Effekte**
- **Löschen geht tendenziell mit geringerer Beanspruchung einher**
- **Individuelle Inhibitionsfähigkeit geht (unabhängig vom Löschen) mit höherer Aufgabenleistung einher**

Informatik

Erklärbarkeit

- **Verbale Erklärungen in unterschiedlichen Detaillierungsgraden**
- **Near-Miss Erklärungen**

Interaktives Lernen

- **Klassifikationsentscheidung und Erklärung durch Nutzer/innen modifizierbar**
- **Feste Regeln (Vorschriften) dürfen nicht durch Lernen änderbar sein**
- **‘Human-in-the-loop’: Kompetenzerhalt, kein blindes Vertrauen**

Ausgewählte Publikationen

Göbel, K., Niessen C., Seufert, S., & Schmid, U. (2022). Explanatory machine learning for justified trust in human-AI collaboration: Experiments on file deletion recommendations. *Frontiers in Artificial Intelligence, 5*, Article 919534. https://doi.org/10.3389/frai.2022.919534

Göbel, K., Hensel, L., Schultheiss, O. C., & Niessen, C. (2022). Meta-analytic evidence shows no relationship between task-based and self-report measures of thought control. *Applied Cognitive Psychology, 36(3)*, 659-672. https://doi.org/10.1002/acp.3952

Kiefer, S., Hoffmann, M., & Schmid, U. (2022). Semantic interactive learning for text classification: A constructive approach for contextual interactions. *Machine Learning and Knowledge Extraction, 4*(4), 994-1010.

Rabold, J., Siebers, M., & Schmid, U. (2022). Generating contrastive explanations for inductive logic programming based on a near miss approach. *Machine Learning, 111*(5), 1799-1820.

Ai, L., Muggleton, S.H., Hocquette, C., Gromowski, M., & Schmid, U. (2021). Beneficial and harmful explanatory machine learning. *Machine Learning, 110*(4), 695-721. https://doi.org/10.1007/s10994-020-05941-0

Göbel, K., & Niessen, C. (2021). Thought control in daily working life: How the ability to stop thoughts protects self-esteem. *Applied Cognitive Psychology, 35*(4), 1011 - 1022. https://doi.org/10.1002/acp.3830

Schmid, U. (2021). Interactive learning with mutual explanations in relational domains. In S. Muggleton and N. Charter, editors, Human-like Machine Intelligence, pp. 337–353. Oxford University Press. https://doi.org/10.1093/oso/9780198862536.001.0001

Gromowski, M., Siebers, M., & Schmid, U. (2020). A process framework for inducing and explaining Datalog theories. *Adv. Data Anal. Classif. 14*(4), 821-835. https://doi.org/10.1007/s11634-020-00422-7

Niessen, C., Göbel, K., Lang, J., & Schmid, U. (2020). Stop thinking: An experience sampling study on suppressing distractive thoughts at work. *Frontiers in Psychology, 11*, Article 1616. https://doi.org/10.3389/fpsyg.2020.01616

Niessen, C., & Lang, J. W. B. (2020). Cognitive control strategies and adaptive performance in a complex work task. *Journal of Applied Psychology*. Advance online publication. https://doi.org/10.1037/apl0000830

Niessen, C., Göbel, K., Siebers, M., & Schmid, U. (2020). Time to forget: Intentional forgetting in the digital world of work. *Zeitschrift für Arbeits- und Organisationspsychologie, 64*, 30-45. https://doi.org/10.1026/0932-4089/a000308

Siebers, M., & Schmid, U. (2019). Please delete that! Why should I? KI-Künstliche Intelligenz, 33(1), 35-44. https://doi.org/10.1007/s13218-018-0565-5

INTENTIONAL FORGETTING IN ORGANISATIONEN SPP 1921

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Friedrich-Alexander-Universität Erlangen-Nürnberg

Lehrstuhl für Psychologie im Arbeitsleben

Prof. Dr. Cornelia Niessen

Projektmitarbeitende

Dr. Kyra Göbel[1]

Angelina Olivia Wilczewski, B.Sc.[2]

Durgesh Nandini, M.Sc.[2]

(Projektmitglied 2020 – 2023)

Dipl.-Psych. Michael Siebers, B.Sc.[2]

(Projektmitglied 2016 – 2019)

[1] *Friedrich-Alexander-Universität Erlangen-Nürnberg*
[2] *Otto-Friedrich-Universität Bamberg*

Otto-Friedrich-Universität Bamberg

Professur Angewandte Informatik, insbes. Kognitive Systeme

Prof. Dr. Ute Schmid