AI-Assisted Document Tagging - Exploring Adaptation Effects among Domain Experts

Sebastian Müller[1], Matthias Baldauf[1] and Peter Fröhlich[2]

[1]OST - Eastern Switzerland University of Applied Sciences, IPM Institute for Information and Process Management, Rosenbergstrasse 59, 9001 St.Gallen, Switzerland

[2]AIT Austrian Institute of Technology, Center for Technology Experience, Giefinggasse 4, 1210 Wien, Austria

Abstract

Keeping a professional knowledge database with scientific publications up to date requires continuous scanning and annotating of newly published articles by domain experts. To shorten this time-consuming process, we study experts' assistance through a domain-specific tag recommender. We introduce the real-life case of a knowledge management system for nursing practitioners and present its architecture and user interface for creating and assigning tag recommendations. While the original tagging interface was thought to assist experts in the tagging process and possibly challenge them to reconsider their tag selections or assign more tags to a document, a preliminary evaluation shows an uncritical adoption of the provided recommendations by experts. We conclude that future design iterations of the recommendation user interface should try to prevent blind trust in the system and encourage reflection on tag suggestions.

1. Introduction

Healthcare professionals are required to ensure safe and cost-efficient care. One way to achieve this is through evidence-based practice, consisting of evidence from research, context, patient preferences, and clinical expertise. To provide evidence-based knowledge to nursing practitioners at the point of care, the Eastern Switzerland University of Applied Sciences offers a knowledge management system that informs about the latest scientific work for practical use [1].

While the system has been attracting great interest among nursing practitioners, scanning and preparing relevant scientific work is challenging. One of the core tasks of the editorial team is the annotation of scientific publications which is both time-consuming and cost-intensive. To this day, the team labeled 1'515 nursing care or medical publications with one or several of 24 tags to categorize the works for later recommendation to practitioners.

In order to support the work of the editorial team, we have been implementing a tag recommender system for health- and care-related scientific publications. Using a BERT-based feature engineering approach [2] combined with a standard radial basis function support vector machine, we were able to achieve Recall@3 values of around 90%.

Besides such performance measures, we particularly focused on the user interaction with the tag recommender: Instead of automatically assigning suggested tags and thus overriding the experts' opinion, we aimed at assisting the experts in the labeling task and reconsidering potential tag selections. We followed user-centered design principles and conducted a co-design workshop with the two nursing experts responsible for tagging the documents [3]. The results of this workshop were finally implemented in the tagging user interface of the knowledge management system 1.

In this position paper, we present the current version of the tag recommender system and report on first expe-

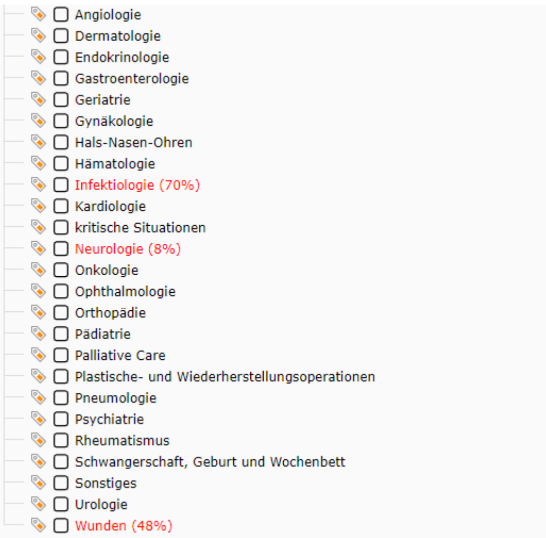

Figure 1: Screenshot of the application's tagging interface, where the three recommended tags are highlighted in red. The predicted tag probabilities are given in brackets.

AutomationXP23: Intervening, Teaming, Delegating - Creating Engaging Automation Experiences, April 23rd, Hamburg, Germany

sebastian.mueller@ost.ch (S. Müller); matthias.baldauf@ost.ch (M. Baldauf); peter.froehlich@ait.ac.at (P. Fröhlich)
0000-0002-1876-5082 (M. Baldauf); 0000-0002-2502-5947 (P. Fröhlich)

riences of the experts involved with the corresponding tagging user interface.

2. System Architecture and Tagging Interface

The knowledge management system *FIT-Nursing Care*[1] is a TYPO3-based website that makes relevant publications available to nursing practitioners. The website as well as our Python-based recommender system are hosted on a Linux server. Data such as relevant scientific publications are managed in a MariaDB database and made accessible via a REST-API. Novel suitable publications are identified by the editorial team and added to the system's literature database. The recommender system is implemented as an external component that checks the database for newly added publications once a day and determines tags to be recommended.

Nursing care experts administrate these publications in the backend of the website. This includes uploading new publications and assigning one or several of the 24 available tags, selected from an alphabetically sorted list. For the last task, the recommender system comes into play. Based on results from a previous co-design workshop with the experts and as a consequence of the distribution of the number of tags per document, three tags are recommended and presented in red with the associated probabilities (see Fig. 1) [3]).

The design shown in fig. 1 was deliberately chosen and implemented to not mislead experts into just confirming the assignment of the tags recommended. With them still having to select tags actively, we aimed at supporting the experts' decision-making process without the system being too authoritarian. In the human factors literature, communicating the uncertainty (or reliability) of a system's perceptions or predictions has proven to support the formation of long-term trust [4, 5]. Ease of use and seamless integration into the existing tagging process was achieved, in order to foster the continuous use of the system without any additional efforts for the experts.

3. First Evaluations

Throughout the implementation of the recommender system, its performance was continuously evaluated on a statistical level using common machine learning metrics (such as precision, recall, map@3).

Besides that, we were interested in how the novel tag recommendation is used by the experts and how their original tagging behavior is impacted by the novel function. We designed and conducted a small user study with two experts from the editorial team, which were asked to re-tag 60 publications. Besides the functional recommender described above, we implemented an additional version of the recommender which picked one of the 24 tags at random. Both experts were provided recommendations from both variants during the tests, without knowing which system recommended the tags to the publication they were currently reviewing. For each publication they tagged, they were asked to identify which of the two systems generated the notification.

Their success rate in identifying the functional recommender was 77% and 85%, respectively. This showed that the non-randomness of the recommendations is noticeable and the experts can assess whether they are shown a solid recommendation or not.

After the system had been in use for three months, we interviewed the key expert mainly tasked with tagging documents about his experiences. He reported that he had noticed a definite change in his tagging approach. While originally he scanned a publication first and then chose one or several adequate tags out of the list, he now started to first check on the three recommendations made and then check their plausibility against the publication, in many cases only by examining the title of the publication. This process of quickly validating the tags recommended led to ignoring potential other suitable tags. In many cases, the option of assigning a non-recommended tag was overlooked.

Furthermore, we found that the current version of the probability display did not fulfill its purpose. The expert reported to rather rely on the red highlighting, yet neglecting the actual probability presented. Whether the percentage provided was above 90% or between 60% and 90% would result in the same outcome - the assignment of the tag. As described by the expert interviewed, this was mainly done out of efficiency and convenience. This is an interesting finding, which tends to be obtained much more often in real-world trials like the one presented here, as compared to lab studies. Findings pointing in this direction have been presented [6], but only little empirical evidence has so far been gathered.

4. Design Challenges

While our first evaluation of the implemented system was informal and only involved the key expert mainly responsible for organizing and tagging the documents, we still found a crucial adaptation effect in his work behavior. Despite his knowledge on the statistical evaluation results and the overall functionality of the recommender, he quickly started to have blind faith in the system. Only focusing on the top three recommended tags and ignoring the probabilities shown, the expert accepted passing on the responsibility for assigning the correct tags to the system.

[1] https://www.fit-care.ch/

For further iterations of the tag recommender user interface, we identify two design challenges:

Prevent blind trust: Users seem to trust the recommender regarding the top three tags suggested while ignoring the probabilities displayed. How can we better point out uncertainties and create awareness for more accurate manual checking? A potential solution could include different color codes for visualizing the different levels of certainty. Another approach could be to also show the uncertainties of the other, non-recommended tags, in order to increase users' sensemaking of the data. Also, different levels of trust indications could be experimented with: additionally to the tag-specific reliability, also explanations for the quality of tag results could be provided, as well as overall system reliability [7].

Encourage reflection: The recommender was supposed to assist in the tagging process, yet not to overrule or replace the experts' opinion. However, recommended tags seem to influence the experts' choice very strongly. How can we encourage reflection on recommendations and combine automated recommendations and expert knowledge in the best possible way?

5. Conclusion and Outlook

The first evaluation of our recommendation system for tagging scientific publications led to interesting results. While the original user interface was thought to assist experts in the tagging process and possibly challenge them to reconsider their tag selections or assign more tags to a document, it seems to have led to an uncritical adoption of the provided recommendations by experts. Moreover, the displayed trust calibration cues did not seem to be considered as such. The displayed probabilities for the three top tags were rather used as a means to derive the preferred tag, based on the identification of the one with the highest reliability ranking.

In future work, we plan investigating experts' interactions with the recommender in more depth. We will prototype and evaluate alternatives for the current user interface, while particularly trying to encourage reflection on the tag suggestions.

References

[1] R. Ranegger, S. Haug, J. Vetsch, D. Baumberger, R. Bürgin, Providing evidence-based knowledge on nursing interventions at the point of care: findings from a mapping project, BMC Medical Informatics and Decision Making 22 (2022) 308. URL: https://doi.org/10.1186/s12911-022-02053-8. doi:10.1186/s12911-022-02053-8.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[3] S. Müller, B. Tödtli, J. Vetsch, M. Rickenmann, S. Haug, M. Baldauf, P. Fröhlich, Designing experts' interactions with a semi-automated document tagging system, Proceedings of the Workshop on Engaging with Automation co-located with the ACM Conference on Human Factors in Computing Systems (CHI 2022) (2022). URL: https://ceur-ws.org/Vol-3154/short4.pdf.

[4] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human factors 46 (2004) 50–80.

[5] F. Chen, A Longitudinal Study of Trust Calibration Methods with Individual Differences, Technical Report, Commonwealth Scientific and Industrial Research Organisation, 2022.

[6] A. R. Wagner, P. Robinette, An explanation is not an excuse: Trust calibration in an age of transparent robots, in: Trust in Human-Robot Interaction, Elsevier, 2021, pp. 197–208.

[7] P. Fröhlich, A. G. Mirnig, D. Falcioni, J. Schrammel, L. Diamond, I. Fischer, M. Tscheligi, Effects of reliability indicators on usage, acceptance and preference of predictive process management decision support systems, Quality and User Experience 7 (2022) 6.