Document Distance Metric Learning in an Interactive Exploration Process

Marco Wrzalik
RheinMain University of Applied Sciences
Wiesbaden, Germany
marco.wrzalik@hs-rm.de

ABSTRACT

Visualization of inter-document similarities is widely used for the exploration of document collections and interactive retrieval [1, 2]. However, similarity relationships between documents are multifaceted and measured distances by a given metric often do not match the perceived similarity of human beings. Furthermore, the user's notion of similarity can drastically change with the exploration objective or task at hand. Therefore, this research proposes to investigate online adjustments to the similarity model using feedback generated during exploration or exploratory search. In this course, rich visualizations and interactions will support users to give valuable feedback. Based on this, metric learning methodologies will be applied to adjust a similarity model in order to improve the exploration experience. At the same time, trained models are considered as valuable outcomes whose benefits for similarity-based tasks such as query-by-example retrieval or classification will be tested.

The measurement of inter-document similarities has been extensively studied in the past. There are various distance metrics using different representations such as weighted term vectors (e.g. TF-IDF, BM25) [9], distributions from topic models [7] or distributed representations from pre-trained language models [5].

Learning a metric can create improved similarity measures that fit specific domain characteristics or the requirements of a task at hand. *Learning to rank* has attracted much research towards this matter in the IR community. Related works form, together with other findings regarding metric learning, the groundwork for this research. In total, highly diverse approaches can be found: linear projections of term vectors [10]; pattern matching in sequences of word embeddings using convolutional neural networks [8]; word sequence learning using siamese recurrent neural networks [6]; to name a few.

Approaches using *online* feedback are particularly relevant to this research. There, collecting implicit feedback based on result lists such as observing clicks [3] or dwell times [4] are common feedback modalities. However, there is only little research on metric learning using feedback from interactions with rich visualizations of inter-document similarities such as proposed in [1]. We hypothesize that users can generate more valuable feedback while interacting with an explorable visualization than with a simple list of best hits. This can be argued with a more comprehensive understanding of underlying similarity relationships such visualizations can give and with the greater range of possible feedback modalities. In a spatial visualization, for example, feedback could be given by correcting datapoint positions, drawing lines as borders for desired clusters or rating the desirability of similarity relationships between result documents.

Following the above-mentioned considerations, the research questions we intend to pursue are: (i) Which feedback modalities enable users to express the desired similarity measure and how can interactive visualizations support users to generate feedback effectively? (ii) Which metric learning methodologies are applicable to improve a similarity model using the feedback from the proposed modalities? (iii) Can a visual exploratory search using the outcome of (i) and (ii) demonstrate arguable benefits over classic searches using result list presentations?

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; **Similarity measures**; • **Computing methodologies → Online learning settings**.

ACM Reference Format:
Marco Wrzalik. 2019. Document Distance Metric Learning in an Interactive Exploration Process. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3331184.3331420

REFERENCES

[1] Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic interaction for visual text analytics. In *Proc. of CHI '12.* ACM, 473–482.

[2] Florian Heimerl, Markus John, Qi Han, Steffen Koch, and Thomas Ertl. 2016. DocuCompass: Effective exploration of document landscapes. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST).*

[3] Katja Hofmann, Shimon Whiteson, and Maarten De Rijke. 2011. Balancing exploration and exploitation in learning to rank online. In *European Conference on Information Retrieval.* Springer, 251–263.

[4] Youngho Kim, Ahmed Hassan, Ryen W White, and Imed Zitouni. 2014. Modeling dwell time to predict click-level satisfaction. In *Proc. of WSDM '14.* ACM, 193–202.

[5] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of ICML'15.* 957–966.

[6] Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of AAAI'16.*

[7] Vasile Rus, Nobal Niraula, and Rajendra Banjade. 2013. Similarity measures based on latent dirichlet allocation. In *Proc. of CICLing '13.* Springer, 459–470.

[8] Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of SIGIR '15.*

[9] John S Whissell and Charles LA Clarke. 2013. Effective measures for inter-document similarity. In *Proceeedings of CIKM '13.* ACM, 1361–1370.

[10] Wen-tau Yih, Kristina Toutanova, John C Platt, and Christopher Meek. 2011. Learning discriminative projections for text similarity measures. In *Proceedings of CoNLL '11.* Association for Computational Linguistics, 247–256.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331420