

# Towards Performance Modeling and Performance Prediction across IR/RecSys/NLP

Edited by

Nicola Ferro<sup>1</sup>, Norbert Fuhr<sup>2</sup>, Gregory Grefenstette<sup>3</sup>, and Joseph A. Konstan<sup>4</sup>

1 University of Padova, Italy [ferro@dei.unipd.it](mailto:ferro@dei.unipd.it)

2 University of Duisburg-Essen, Germany [norbert.fuhr@uni-due.de](mailto:norbert.fuhr@uni-due.de)

3 Institute for Human Machine Cognition, USA [ggrefenstette@ihmc.us](mailto:ggregenstette@ihmc.us)

4 University of Minnesota, Minneapolis, USA [konstan@umn.edu](mailto:konstan@umn.edu)

---

## Abstract

This report briefly describes the organization and the plenary talks given during the Dagstuhl Perspectives Workshop 17442. The goal of this workshop was to investigate the state-of-the-art and to delineate a roadmap and research challenges for performance modeling and prediction in three neighbour domains, namely information retrieval (IR), recommender systems (RecSys), and natural language processing (NLP).

**Seminar** 30 October–03 November, 2017 – [www.dagstuhl.de/17442](http://www.dagstuhl.de/17442)

**1998 ACM Subject Classification** H.3 Information Storage and Retrieval, I.2.7 Artificial Intelligence – Natural Language Processing

**Keywords and phrases** Information Systems, Formal models, Evaluation, Simulation, User Interaction

**Digital Object Identifier** 10.4230/DagRep.7.10.139

## 1 Executive Summary

*Nicola Ferro*

*Norbert Fuhr*

*Gregory Grefenstette*

*Joseph A. Konstan*

**License** © Creative Commons BY 3.0 Unported license

© Nicola Ferro, Norbert Fuhr, Gregory Grefenstette, and Joseph A. Konstan

Information systems, which manage, access, extract and process non-structured information, typically deal with vague and implicit information needs, natural language and complex user tasks. Examples of such systems are information retrieval (IR) systems, recommender systems (RecSys), and applications of natural language processing (NLP) such as e.g. machine translation, document classification, sentiment analysis or search engines. The discipline behind these systems differs from other areas of computer science, and other fields of science and engineering in general, due to the lack of models that allow us to predict system performances in a specific operational context and to design systems ahead to achieve a desired level of effectiveness. In the type of information systems we want to look at, we deal with domains characterized by complex algorithms, dependent on many parameters and confronted with uncertainty both in the information to be processed and the needs to be



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Towards Performance Modeling and Performance Prediction across IR/RecSys/NLP, *Dagstuhl Reports*, Vol. 7, Issue 10, pp. 139–146

Editors: Nicola Ferro, Norbert Fuhr, Gregory Grefenstette, Joseph A. Konstan



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

addressed, where the lack of predictive models is somehow bypassed by massive trials of as many combinations as possible.

These approaches relying on massive experimentation, construction of testbeds, and heuristics are neither indefinitely scaled as the complexity of systems and tasks increases nor applicable outside the context of big Internet companies, which still have the resources to cope with them.

The workshop was organized as follows. The first day was devoted to plenary talks focused on providing a general introduction to IR, RecSys, and NLP and on digging into some specific issues in performance modeling and prediction in these three domains. The second day, participants split into three groups – IR, RecSys, and NLP – and explored performance modeling and prediction issues and challenges within each domain; the working groups then reconvened to present the output of their discussion in a plenary session in order to cross-fertilize across disciplines and to identify cross-discipline themes to be further investigated. The third day, participant split into groups which explored these themes – namely measures, performance analysis, documenting and understanding assumptions, application features, and modeling performance – and reported back in plenary sessions to keep all the participants aligned with the ongoing discussions. The fourth and fifth days have been devoted to the drafting of this report and the manifesto originated from the workshop.

This documents reports the overview of the talks given by the participants on the first day. The outcomes of the working groups – both within-discipline themes and cross-discipline themes – as well as the identified research challenges and directions are presented in the Dagstuhl Manifesto corresponding to this Perspectives Workshop [1].

**Acknowledgements.** We thank Schloss Dagstuhl for hosting us.

## References

- 1 N. Ferro, N. Fuhr, G. Grefenstette, J. A. Konstan, P. Castells, E. M. Daly, T. Declerck, M. D. Ekstrand, W. Geyer, J. Gonzalo, T. Kuflik, K. Lindén, B. Magnini, J.-Y. Nie, R. Perego, B. Shapira, I. Soboroff, N. Tintarev, K. Verspoor, M. C. Willemsen, and J. Zobel. Manifesto from Dagstuhl Perspectives Workshop 17442 – Towards Performance Modeling and Performance Prediction across IR/RecSys/NLP. *Dagstuhl Manifestos, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany*, 7(1), 2018.

**2 Table of Contents**

**Executive Summary**  
*Nicola Ferro, Norbert Fuhr, Gregory Grefenstette, and Joseph A. Konstan . . . . .* 139

**Overview of the Talks**

The Validity Problems of IR  
*Norbert Fuhr . . . . .* 142

Recommender Systems: The Evaluation Challenge  
*Joseph A. Konstan . . . . .* 142

Evaluation in Natural Language Processing  
*Gregory Grefenstette . . . . .* 142

Bad for IR, Worse for Recommenders: Missing Data and the External Validity of  
Offline Evaluations  
*Michael D. Ekstrand . . . . .* 143

Advanced Performance Modelling (and Prediction?) Techniques in IR  
*Nicola Ferro . . . . .* 143

Objective or Subjective measures?  
*Martijn C. Willemsen . . . . .* 144


User Utterance Understanding in Conversational Systems  
*Bernardo Magnini . . . . .* 145

**Participants . . . . .** 146

### 3 Overview of the Talks

#### 3.1 The Validity Problems of IR

Norbert Fuhr (*University of Duisburg-Essen, DE*)

License  Creative Commons BY 3.0 Unported license  
© Norbert Fuhr

Current IR experiments often suffer from flaws that affect the internal validity, such as e.g. invalid or inappropriate metrics, poor test design, multiple testing without correction, or lack of reproducibility. External validity deals with the extent to which the findings of a study can be generalized. For addressing this issue, we must deepen our understanding of the models used, especially their underlying assumptions, and devise methods for checking these assumptions in a new setting. Furthermore, we need to investigate the relationship between application properties and performance, i.e. characteristics of the controlled variables (documents, topics and relevance assessments) of an IR experiment and the evaluation result.

#### 3.2 Recommender Systems: The Evaluation Challenge

Joseph A. Konstan (*University of Minnesota – Minneapolis, US*)

License  Creative Commons BY 3.0 Unported license  
© Joseph A. Konstan

Recommender systems have become ubiquitous, helping businesses market and users find desired information and products. They employ a variety of techniques including non-personalized summary statistics, content-based information filtering, and personalized collaborative filtering, often using latent-factor models based on or approximating matrix factorization. Evaluating recommender system performance is challenging because the most accessible measures such as predictive accuracy, rank performance, etc., all fail to capture the actual utility of the system—recommending items the user would not have selected anyway without the aid of the recommender. We review a variety of algorithms, offline and online evaluation metrics, and the challenge of effectively evaluating performance of recommender systems in the context of actual use.

#### 3.3 Evaluation in Natural Language Processing

Gregory Grefenstette (*Institute for Human Machine Cognition, US*)

License  Creative Commons BY 3.0 Unported license  
© Gregory Grefenstette

In this talk, I present the two main ways that Natural Language Processing (NLP) systems are evaluated. One way is calculating the improvement in some applications that use NLP processes to produce their results. Examples of these applications are Summarisation, Question Answering, Plagiarism Detection, Speech Recognition, Entity Extraction, Classification, Machine Translation, Author Identification, Image Labeling, Information Retrieval and Recommendation, among others. The second way is intrinsic evaluation of individual NLP

modules, such as Language Identification, Tokenisation, Morphological Analysis, Part-of-Speech Tagging, Chunking, Shallow Parsing and Semantic Role Labelling, Deeper Parsing, Co-reference resolution, Topic Detection and Taxonomy/Thesaurus Extraction. We will explain how automated evaluation systems are set up, run and results reported, based upon gold standards and common metrics. For prediction, we will also describe some ways to characterize collections (used for training or testing). Finally, we will give an example of how much data is needed to produce expected results for analogy tests in word embeddings systems.

### 3.4 Bad for IR, Worse for Recommenders: Missing Data and the External Validity of Offline Evaluations

*Michael D. Ekstrand (Boise State University, US)*

**License** © Creative Commons BY 3.0 Unported license  
© Michael D. Ekstrand

Missing data impedes the realistic offline evaluation of information retrieval and recommender systems. Data sets do not have complete data on the relevance of items or documents to users or queries. The information retrieval community has developed several techniques that attempt to address these problems, but these techniques are not applicable to evaluating recommender systems due to the personalized and entirely subjective nature of relevance in recommender applications. Further, the nature of recommendation tasks and the subjectivity of relevance mean that this missing data is particularly detrimental to the validity of recommender evaluations. In this talk, I review the problem of missing data in information retrieval and recommendation tasks, the methods IR has developed, and explain why those methods are not suitable for evaluating recommenders. I also describe some additional concerns in recommender system evaluation that arise from missing data, and demonstrate that proposed solutions depend on missing theoretical knowledge or unrealistic assumptions.

### 3.5 Advanced Performance Modelling (and Prediction?) Techniques in IR

*Nicola Ferro (University of Padova, IT)*

**License** © Creative Commons BY 3.0 Unported license  
© Nicola Ferro

Trying to explain the performance of a set of Information Retrieval (IR) systems across a set of topics is a preliminary step indispensable to start envisioning how to predict the performance of such systems. In this talk we discuss the different types of performance models which have been developed so far, which are all based on General Linear Mixed Models (GLMM) and ANalysis Of VAriance (ANOVA).

We start from the Topic and System effects models [1, 6]. We then consider the breakdown of the System effect into those of its components, namely stop lists, stemmers, and IR models [3, 4]. We discuss the use of simulation for showing the importance of the Topic\*System interaction effect [5] as well as very recent work on using random partitions of the document corpus to estimate this effect [7]. Finally, we report on preliminary results about the Sub-Corpus effect and System\*Sub-Corpus interaction effect [2].

We conclude by discussing how these explanatory models might be turned into predictive ones by using features describing these different factors and regression-like techniques.

## References

- 1 D. Banks, P. Over, and N.-F. Zhang. Blind Men and Elephants: Six Approaches to TREC data. *Information Retrieval*, 1(1-2):7–34, May 1999.
- 2 N. Ferro and M. Sanderson. Sub-corpora Impact on System Effectiveness. In N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, and R. W. White, editors, *Proc. 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*, pages 901–904. ACM Press, New York, USA, 2017.
- 3 N. Ferro and G. Silvello. A General Linear Mixed Models Approach to Study System Component Effects. In R. Perego, F. Sebastiani, J. Aslam, I. Ruthven, and J. Zobel, editors, *Proc. 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016)*, pages 25–34. ACM Press, New York, USA, 2016.
- 4 N. Ferro and G. Silvello. Towards an Anatomy of IR System Component Performances. *Journal of the American Society for Information Science and Technology (JASIST)*, 2017.
- 5 S. E. Robertson and E. Kanoulas. On Per-topic Variance in IR Evaluation. In W. Hersh, J. Callan, Y. Maarek, and M. Sanderson, editors, *Proc. 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012)*, pages 891–900. ACM Press, New York, USA, 2012.
- 6 J. M. Tague-Sutcliffe and J. Blustein. A Statistical Analysis of the TREC-3 Data. In D. K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, pages 385–398. National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA, 1994.
- 7 E. M. Voorhees, D. Samarov, and I. Soboroff. Using Replicates in Information Retrieval Evaluation. *ACM Transactions on Information Systems (TOIS)*, 36(2):12:1–12:21, September 2017.

## 3.6 Objective or Subjective measures?

*Martijn C. Willemsen (Eindhoven University of Technology, NL)*

License  Creative Commons BY 3.0 Unported license  
© Martijn C. Willemsen

Recommenders are traditionally evaluated using offline evaluation on historical data. More recently, focus has shifted to online evaluation of objective behavioral data using AB testing. However, such behavior is hard to interpret without using subjective measures that help interpreting the meaning of the behavior. For example lower click-rates might not be reflecting reduced interest, but increased engagement of a user consuming the recommended content from beginning to end without additional interactions. In this talk I first introduce our user-centric evaluation framework [3] and subsequently show in three cases how objective and subjective measures go hand in hand in predicting and understanding user behavior and system effectiveness. The first case demonstrates how we can build a better prediction model for user segments based on subjective survey data of only 3000 users than on the behavioral data of all 100k users [2]. In the second case I show how objective measures of similarity, obscurity and accuracy can be linked to subjective perceptions of diversity, novelty and satisfaction. These subjective measures can explain the different relative preferences of users for three classical recommender algorithms (item-item, user-user and SVD) [1]. In the final case I show how choice difficulty of recommendation lists can be reduced by using latent-feature diversification, which reduces similarity between items while maintaining


sufficient levels of attractiveness. The study shows that a diverse 5-item set is experienced as more satisfactory than a top-5 item set, despite the lower predicted accuracy of the list and the lower average rank of the items chosen by the user [4].

## References

- 1 Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. User Perception of Differences in Recommender Algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 161–168, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2668-1. 10.1145/2645710.2645737.
- 2 Mark P. Graus, Martijn C. Willemsen, and Kevin Swelsen. Understanding Real-Life Website Adaptations by Investigating the Relations Between User Behavior and User Experience. In Francesco Ricci, Kalina Bontcheva, Owen Conlan, and Séamus Lawless, editors, *User Modeling, Adaptation and Personalization*, number 9146 in Lecture Notes in Computer Science, pages 350–356. Springer International Publishing, June 2015. ISBN 978-3-319-20266-2 978-3-319-20267-9. URL [http://link.springer.com/chapter/10.1007/978-3-319-20267-9\\_30](http://link.springer.com/chapter/10.1007/978-3-319-20267-9_30).
- 3 Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, March 2012. ISSN 0924-1868, 1573-1391. 10.1007/s11257-011-9118-4.
- 4 Martijn C. Willemsen, Mark P. Graus, and Bart P. Knijnenburg. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Modeling and User-Adapted Interaction*, 26(4):347–389, October 2016. ISSN 0924-1868, 1573-1391. 10.1007/s11257-016-9178-6.

## 3.7 User Utterance Understanding in Conversational Systems

Bernardo Magnini (Fondazione Bruno Kessler, IT)

License  Creative Commons BY 3.0 Unported license  
© Bernardo Magnini

In the context of the recent resurgence of Artificial Intelligence, Conversational Agents have been attracting the attention of the NLP community. Conversational systems offer an interesting scenario for cross-domain predictability in NLP, for two reasons: (i) task oriented conversational agents are being developed in a huge numbers of application scenarios (e.g. virtual coaching, personal assistant, e-commerce, etc.) in different domains (e.g. food, sport) and for different languages; (ii) there are very few conversational datasets available for training models. In this context predictability is crucial for successfully develop high quality conversational systems. However, it opens several fundamental research questions. Which are the characteristics of the language (e.g. specific terminology, typical conversational patterns) of a certain domain that mainly affect the system performance? Which are the relevant characteristics of the application domain (e.g. complexity of entities and properties)? Which are the characteristics of the task (i.e. the problem to be solved by conversation, like booking a restaurant, or recommending a book)? How these three levels are related one with the other to determine predictability?



## Participants

- Pablo Castells  
Autonomous University of Madrid, ES
- Elizabeth M. Daly  
IBM Research – Dublin, IE
- Thierry Declerck  
DFKI – Saarbrücken, DE
- Michael D. Ekstrand  
Boise State University, US
- Nicola Ferro  
University of Padova, IT
- Norbert Fuhr  
Universität Duisburg-Essen, DE
- Werner Geyer  
IBM TJ Watson Research Center – Cambridge, US
- Julio Gonzalo  
UNED – Madrid, ES
- Gregory Grefenstette  
IHMC – Paris, FR
- Joseph Konstan  
University of Minnesota – Minneapolis, US
- Tsvi Kuflik  
Haifa University, IL
- Krister Lindén  
University of Helsinki, FI
- Bernardo Magnini  
Bruno Kessler Foundation – Trento, IT
- Jian-Yun Nie  
University of Montréal, CA
- Raffaele Perego  
CNR – Pisa, IT
- Bracha Shapira  
Ben Gurion University – Beer Sheva, IL
- Ian Soboroff  
NIST – Gaithersburg, US
- Nava Tintarev  
TU Delft, NL
- Karin Verspoor  
The University of Melbourne, AU
- Martijn Willemsen  
TU Eindhoven, NL
- Justin Zobel  
The University of Melbourne, AU

