Report on the SIGIR 2017 Workshop on Axiomatic Thinking for Information Retrieval and Related Tasks (ATIR)

Enrique Amigó[1] **Hui Fang**[2]
Stefano Mizzaro[3] **ChengXiang Zhai**[4]

[1] National Distance University , Spain
[2] University of Delaware, USA
[3] University of Udine, Italy
[4] University of Illinois at Urbana-Champaign, USA

Abstract

The SIGIR 2017 workshop on Axiomatic Thinking for Information Retrieval and Related Tasks took place on August 11, 2017 in Tokyo, Japan. The workshop aimed to help foster collaboration of researchers working on different perspectives of axiomatic thinking and encourage discussion and research on general methodological issues related to applying axiomatic thinking to information retrieval and related tasks. The program consisted of one keynote talk, four research presentations and a final panel discussion. This report outlines the events of the workshop and summarizes the major outcomes. More information about the workshop is available at `https://www.eecis.udel.edu/~hfang/ATIR.html`.

1 Introduction

Recently, axiomatic thinking has been adopted for the development of both retrieval models and evaluation metrics with great promise. The general idea of axiomatic thinking is to seek a set of desirable properties expressed mathematically as formal constraints to guide the search for an optimal solution; the explicit expression of desirable properties makes it possible to analytically address issues that would otherwise appear to be purely empirical, provide theoretical guidance on how we might be able to optimize a retrieval model or evaluation metric, and apply any identified constraints directly in many practical applications.

The growth of research on axiomatic thinking for Information Retrieval (IR) can be easily seen from the increasing number of publications on this general topic (over 40 papers published recently), mostly in two lines. The first line is the application of axiomatic analysis to retrieval model development, where relevance is modeled based on a set of retrieval constraints that any reasonable retrieval functions need to satisfy. These constraints are then used to diagnose deficiencies of existing retrieval functions and guide the search for more

effective retrieval functions. Axiomatic analysis has been shown to be effective for diagnosing deficiencies of basic retrieval models and improving them [3, 5, 6], including particularly the development of BM25+, an improvement of a long-standing state of the art model (i.e., BM25) to fix its deficiency in document length normalization [9]. The second line is the successful application of axiomatic analysis to evaluation, in particular, to formalizing evaluation metrics. In general, these studies focus on setting particular situations in which metrics should behave in a certain manner, that are specified by means of constraints or axioms that metrics should satisfy. This theme recurs in the literature since the 70s, but it has received an increasing interest in the last five or ten years, when most of the papers have been published [1, 2, 4, 7, 8, 10, 12–14]. Moreover, the formalization attempts of evaluation metrics have gone well beyond the IR field: several results concern related areas like classification [13] or clustering [1, 11].

This workshop has three goals. The first goal is to go deeper into the natural connection between axiomatic IR models and axiomatic metrics accounts since, in both cases, researchers are attempting to model relevance with axioms, specifically model how users would make relevance judgments on documents with respect to a query. However, so far the two research lines have proceeded (almost) independently, and the broad IR community has not yet gained benefit from any synergy that is quite likely to be present. Second, we would like to engage a broader discussion of "axiomatic thinking" in general in terms of its potential broad impact on optimizing evaluation metrics for any tasks (not just IR tasks) and optimizing ranking for many other tasks, and identify important future research directions, especially promising interdisciplinary topic areas. This is an aim where the still missing synergy is likely to become concrete: indeed, whereas axiomatic IR models have been focusing on IR, the research on axiomatic metrics has been broader and has already included other tasks beyond ranking. Finally, it is necessary to discuss and summarize the basic open questions in the applicability of axiomatics in IR: (i) How to interpret *axiomatic* in the context of IR? (ii) How to identify potential constraints? (iii) How to verify those constraints? and (iv) How to combine axiom- and data-oriented approaches?

2 Topics of Interest and Program Committee

The workshop aimed to bring together researchers and practitioners from a broader community to exchange research ideas and results and foster collaborations across sub-communities. Some of the specific topics we envisioned to be covered by the workshop theme include, but not limited to:

- What constraints are effective to improve retrieval performance independent of the underlying model?
- What constraints were expected to be useful but have not been effective in practice? Why not? In the case of evaluation metrics, why some metric constraints do not affect the system comparison or the user satisfaction?
- How can we potentially unify the axiomatic analysis of IR models and evaluation metrics given that both lines of work aim at formally modeling relevance?
- Have new languages, media, or domains suggested new constraints for established domains?

- To what extent is a valid constraint in one domain also valid in other domains? More generally, which constraints for retrieval methods or evaluation metrics are core ones, and which constraints are highly scenario dependent?
- How can axiomatic thinking be combined with machine learning techniques to learn more effective retrieval functions?

All the submissions are peer-reviewed. The program committee consists of the following experts:

- Pablo Castells, Universidad Autonoma de Madrid, Spain.
- Fabio Crestani, University of Lugano, Switzerland.
- Ronan Cummins, University of Cambridge, UK.
- Norbert Fuhr, University of Duisburg-Essen, Germany.
- Julio Gonzalo, UNED, Spain.
- Yuanhua Lv, Microsoft Research, USA.
- Fabrizio Sebastiani, ISTI-CNR, Italy.
- Azadeh Shakery, University of Tehran, Iran.

3 Keynote

The program started with a keynote talk by Prof. Jian-Yun Nie from University of Montreal. The talk is about **Inference in Axiomatic Approaches to IR**. The studies on axiomatic analysis in IR have found a set of intuitive constraints that a ranking function should satisfy. Up to now, the constraints are mainly related to term distributions in the document, the collection and the pseudo-feedback documents. They provide a good explanation of the success and failure of the traditional models. However, few studies have coped with the problem of inference and reasoning in IR, which is an important aspect in the current and future IR. Indeed, when we judge the relevance of a document, we not only look at the distribution of query terms in it, but also consider how the other terms are semantically related to the query. In this keynote talk, Prof. Jian-Yun Nie argued that this inference aspect should be covered in the axiomatic analysis, and he examined a few possible avenues to do it. The talk has stimulated interesting discussions on new directions in axiomatic analysis especially in applying axiomatic analysis to assist inference of relative relevance and the measurement of utility (or power) of an axiom.

4 Research Presentations

The program continued with four research presentations by Laura Dietz, Paul Kantor, Rocío Cañamares and Enrique Amigó, covering the theoretical ground of generative retrieval models, the connection between Quantum Mechanics and IR, the probabilistic foundations of ranking functions in recommendation, and the axiomatics of similarity measures.

- **On the Equivalence of Generative and Discriminative Formulations of the Sequential Dependence Model:** This work was developed by Laura Dietz (University of New Hampshire) and John Foley (University of Massachusetts). This work

paper details the axiomatic underpinning of the SDM model as discriminative and generative probabilistic model. They demonstrate that the sequential dependence model (SDM) based on the theory of probabilistic graphical models is equivalent to a generative probabilistic model, and that the only difference arises whether model parameters are estimated in log-space or Multinomial-space. They demonstrate that parameter-estimation with grid-tuning is negatively impacting the generative formulation, an effect that vanishes when parameters are estimated with coordinate- gradient descent. This is concerning, since empirical differences may be falsely attributed to improved models. From the *axiomatic thinking's* point of view, this contribution gives an example of how different approaches to the same problems are grounded on the same basic assumptions.

- **QM and IR: Another Perspective:** This work was developed by Paul Kantor. This work explores some possible ways in which the Quantum Mechanics theory of measurement might be of use in axiomatizing the processes by which a person gains information from the system and vice versa. The key innovation is to propose both a model, as seen by the system, of the discontinuous change of the state of the user, and a coupled representation of the states of the user and the system, which might ultimately be used to guide the selection of materials to be presented to the user. From the *axiomatic thinking's* point of view, this work opens the door of considering other ways of defining the user behavior principles on which relevance scoring functions are grounded.
- **On the Optimal Non-Personalized Recommendation: From the PRP to the Discovery False Negative Principle:** This work was developed by Rocío Cañamares and Pablo Castells (Universidad Autónoma de Madrid, Spain). They revisit the Probability Ranking Principle in the context of recommender systems, finding a key difference in the retrieval protocol with respect to query-based search, that leads to the identification of a different optimal ranking principle for non-personalized recommendation. Based on this finding, they explore the definition of practical ranking functions that may lean towards the optimal ranking. Their theoretical analysis is corroborated by empirical results. From the *axiomatic thinking's* point of view, this work finds a simple probabilistic principle for the optimal ranking in the context of recommender systems.
- **An Axiomatic Account of Similarity:** This work was developed by Enrique Amigó, Julio Gonzalo, Fernando Giner and Felisa Verdejo (Uned, Spain). In this presentation, they show how axiomatic explanations of similarity from other fields (i.e., Tversky's axioms from the point of view of cognitive sciences, and metric spaces from the point of view of algebra) do not completely fit the problem of similarity in Information Access, and they propose a new set of axioms which can be synthesized into a single Similarity Information Monotonicity (SIM) axiom. Directly grounded on this axiom, they then introduce a new similarity model, the Information Contrast Model, which generalizes both Tversky's linear contrast model and Pointwise Mutual Information, and, unlike previous similarity models, satisfies the SIM axiom for a certain range of values of its parameters. From the *axiomatic thinking's* point of view, this work revisits the traditional similarity axiomatics based on metric spaces and feature sets, defining a new axiom set that supports the design of similarity measures on information theory.

5 Panel Discussions

The program continued with a panel moderated by ChengXiang Zhai. The topic of the panel discussion is **Past, Present and Future of Axiomatic Thinking**. The panelists include a group of experienced IR researchers:

- Ben Carterette, University of Delaware, USA
- Pablo Castells, Universidad Autonoma de Madrid, Spain
- Kevyn Collins-Thompson, University of Michigan, USA
- Paul Kantor, Rutgers University, USA
- Stephan E. Robertson, City University of London, UK (Professor Emeritus)

Zhai started the panel discussion with the two questions for the panelists: (1) *How is your work related to axiomatic thinking?* and (2) *What is your definition of axiomatic thinking?* It is interesting to hear from the panelists about their experiences on applying axiomatic thinking to different tasks. We briefly summarized their answers as follows.

- Castells shared his experiences of applying axiomatic thinking to evaluation metrics and diversity in the recommender systems. He mentioned that axiomatic thinking is a powerful way of communicating with others when justifying new methods, and it is also a tool for ourselves to check whether a method makes sense and discover additional advantages that we did not think of. He defined axioms as minimal assumptions that are needed to develop a theory. Axioms should be as few as possible.
- Kantor mentioned that he has worked on axioms, that are obviously true. For example, back in 1984, his work on discoverer's cross references was based on the axiom that a person who used a document knows more about what the document is about than the words in the document, although the axiom does not have a nice mathematical formulation.
- Collins-Thompson used his dissertation work as an example to explain how to apply axiomatic thinking to solve difficult problems such as finding reliable query expansion methods. In particular, he picked a simple representation of functions and formulated it as a convex optimization problem. He was able to encode axioms about query terms as set-level constraints. He mentioned that connecting axiomatic thinking with convex optimization problem allows him to do sensitivity analysis and perturbation analysis.
- Carterette's research focused on theoretical aspects of the IR evaluation. He mentioned that, in his dissertation work, he was looking at the bounds of performance differences between two systems under certain conditions. To him, axiom thinking is not about having a small set of assumptions needed to develop a theory, but it is about doing deduction on a set of principles you derived from evaluation measures and ranking lists.
- Robertson pointed out that it is unclear what are the axioms. Every retrieval model is based on assumptions, so in some sense we might say that the assumptions are axioms. In some existing studies, axioms seem to be above scoring functions, i.e., how a scoring function should behave. But he prefers axioms that are not tied to a particular scoring function. For example, term independence assumption is not based on a particular scoring function, but we can derive a scoring function from it. He also explained the two-stage process of deriving Okapi BM25. Finally, he pointed out one limitation of our

field: that we tend to concentrate on the retrieval effectiveness experiments even when retrieval methods are derived from more abstract models. We ignore the questions such as, given a model, what are all the predictions we can make from the model that might be testable.

Zhai summarized the insightful discussions into a few key points. (1) Axioms are defined as a minimum set of requirements or assumptions that we need for a task. Axioms depend on tasks, and there are no universal axioms. (2) There are discussions about informal vs. formal constraints. Informal constraints sometimes can be preliminary steps that will lead to the formal constraints. (3) Formal constraints can be tapped into optimization algorithms to solve the problem. Axiomatic thinking is a tool that allows us to see the problems or methods from a different angle. (4) Another potential benefit of axiomatic thinking lies on analytical comparison.

The panel then discussed *whether axiomatic is the right word here*. Castells suggested that axiomatics can be seen as an umbrella covering assumptions, desirable properties, requirement, etc. And he also suggested that we might want to create a glossary for different notions. Moreover, Robertson pointed out that assuming an axiom as a thing that we all agree on is a very strong assumption.

The panel concluded with a discussion of the future, i.e., *how can we grow an axiomatic thinking community?* Most participants expressed interests that they are interested in attending the workshop if we organize it again. In response to the question *how axiomatic thinking can be combined with other techniques such as machine learning*, Collins-Thompson pointed out that integrating both perspective is a challenging problem, which requires to be formulated in such a way that it can be solved in real time.

6 Conclusions and Future

We have mostly realized the three goals that we have set for this workshop, including integration of research in axiomatic retrieval models and that in axiomatic evaluation metrics, engaging a broad discussion of the topic from multiple perspectives, and examining some basic questions and open research problems.

One conclusion that we draw from this workshop, and especially from the closing panel, is that there seems to be no clear distinction between axiomatics and other theoretical tools applied in IR researches, such as assumptions, desirable properties or even scoring function generalizations. However, there is agreement on the advantages of axiomatics as a tool that can help researches to find optimal functions or models, or at least, to guide the searching process. In addition, axiomatics can help to understand the nature of approaches and analyze their differences.

A major challenge in research on axiomatics is how to formally define relatively general constraints with consideration of individual scenarios and tasks. And the panelists have doubts about the feasibility of achieving agreement when setting fundamental assumptions.

The nuclear open question was suggested by Robertson, who pointed out that traditionally, science methodology is based on checking basic principles by means of focused experiments, while our community is working on competing on system effectiveness estimated by means of evaluation benchmarks. In this sense, axiomatic thinking provides a methodology to take more solid steps in the research process.

In conclusion, we hope that this workshop has been the first step to create a research community working on these interesting issues. To further help creating a research community on this topic, we are working on a proposal to edit a special issue of a major information retrieval journal on this topic.

Acknowledgments

We thank ACM and SIGIR for hosting this workshop, thank all the participants for attending the workshop, thank all the panelists for sharing their experiences and insights, and thank all the PC members for providing valuable comments for the submissions.

References

[1] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486, 2009.

[2] E. Amigó, J. Gonzalo, and F. Verdejo. A general evaluation measure for document organization tasks. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 643–652, New York, NY, USA, 2013. ACM.

[3] P. D. Bruza and T. W. C. Huibers. Investigating aboutness axioms using information fields. In *Proceedings of the 17th ACM SIGIR*, pages 112–121, New York, NY, USA, 1994. Springer-Verlag New York, Inc. ISBN 0-387-19889-X. URL `http://dl.acm.org/citation.cfm?id=188490.188521`.

[4] L. Busin and S. Mizzaro. Axiometrics: An axiomatic approach to information retrieval effectiveness metrics. In *Proceedings of ICTIR 2013: 4th International Conference on the Theory of Information Retrieval*, pages 22–29, New York – USA, Oct. 2013. ACM.

[5] H. Fang and C. Zhai. An exploration of axiomatic approaches to information retrieval. In *SIGIR '05*, pages 480–487, 2005. ISBN 1-59593-034-5. doi: http://doi.acm.org/10.1145/1076034.1076116.

[6] H. Fang, T. Tao, and C. Zhai. Diagnostic evaluation of information retrieval models. *ACM Trans. Inf. Syst.*, 29(2):7:1–7:42, Apr. 2011. ISSN 1046-8188. doi: 10.1145/1961209.1961210. URL `http://doi.acm.org/10.1145/1961209.1961210`.

[7] M. Ferrante, N. Ferro, and M. Maistro. Towards a formal framework for utility-oriented measurements of retrieval effectiveness. In *Proceedings of ICTIR 2015*, pages 21–30, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3833-2. doi: 10.1145/2808194.2809452. URL `http://doi.acm.org/10.1145/2808194.2809452`.

[8] C. Ferri, J. Hernández-Orallo, and R. Modroiu. An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38, 2009.

[9] Y. Lv and C. Zhai. Lower-bounding term frequency normalization. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, 2011.

[10] E. Maddalena and S. Mizzaro. Axiometrics: Axioms of information retrieval effectiveness metrics. In *Proceedings of the Sixth EVIA Workshop*, pages 17–24, Tokyo, Japan, 2014. National Institute of Informatics. ISBN: 978-4-86049-066-9.

[11] M. Meila. Comparing clusterings. In *Proceedings of COLT 03*, 2003. URL `http://www.stat.washington.edu/mmp/www.stat.washington.edu/mmp/Papers/compare-colt.pdf`.

[12] A. Moffat. Seven numeric properties of effectiveness metrics. In *AIRS'13*, pages 1–12, 2013.

[13] F. Sebastiani. An axiomatically derived measure for the evaluation of classification algorithms. In *Proceedings of ICTIR 2015*, pages 11–20. ACM, 2015. ISBN 978-1-4503-3833-2. doi: 10.1145/2808194.2809449. URL `http://doi.acm.org/10.1145/2808194.2809449`.

[14] M. Sokolova. Assessing invariance properties of evaluation measures. In *Proceedings of NIPS'06 Workshop on Testing Deployable Learning and Decision Systems*, 2006.