Axiomatic Thinking for Information Retrieval — And Related Tasks

Enrique Amigo
National Distance Education University
SPAIN
enrique@lsi.uned.es

Hui Fang
University of Delaware
USA
hfang@udel.edu

Stefano Mizzaro
University of Udine
Italy
mizzaro@uniud.it

ChengXiang Zhai
University of Illinois at Urbana-Champaign
USA
czhai@illinois.edu

ABSTRACT

We propose to organize the first workshop on the emerging interdisciplinary research area of applying axiomatic thinking to information retrieval (IR) and related tasks. The workshop would 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 IR and related tasks.

ACM Reference format:
Enrique Amigo, Hui Fang, Stefano Mizzaro, and ChengXiang Zhai. 2016. Axiomatic Thinking for Information Retrieval — And Related Tasks . In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 3 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 MOTIVATION

The main task of an Information Retrieval (IR) system is to return relevant documents to users in response to a query. However, this task is inherently an empirical task since the definition of *relevance* of a document to a query is not well defined, and in general, can only be judged by users who issued the query. Yet, defining relevance as rigorously as we can is essential to the development of both effective IR systems and sound evaluation metrics. As a result, modeling relevance has always been a central challenge in IR research for both retrieval model development and evaluation. Indeed, all information retrieval models developed so far (which are the basis of the algorithms used in all search engine applications) has explicitly or implicitly adopted one way or another to formalize the vague concept of relevance; similarly, all evaluation metrics of IR are meant to quantify the utility of the retrieved results from a user's perspective, and thus must also accurately reflect a user's view of relevance. Due to the importance of modeling relevance, the issue has been addressed since the early days of IR research, leading to multiple retrieval models such as the vector space models and probabilistic models as well as basic metrics such as precision, recall, average precision, and nDCG.

For many years, however, the notion of relevance has been modeled mostly as a "black box" without paying attention to the specific properties that a retrieval function or evaluation metric must satisfy from the perspective of relevance modeling; as such, there is often no way to analytically assess the soundness of a retrieval model or evaluation metric, hindering the study of their optimality. 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 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

Conference'17, Washington, DC, USA

DOI: 10.1145/nnnnnnn.nnnnnnn

[13] or clustering [1, 11], or even textual similarity, opinion mining, and so on.

The goal of the proposed workshop is to bring together researchers and practitioners interested in applying axiomatic analysis to all kinds of IR and IR-related problems, including particularly both those interested in developing retrieval models and those interested in developing evaluation measures, and to enable them to share their findings (both positive or negative), to present their latest research results, and to discuss future directions.

The proposed workshop is important and timely for several reasons. First, there is a 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. This workshop would be the first to bring together the two lines as well as the two communities — that, incidentally, are quite active in the last years as already mentioned above. Second, we believe that we need 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 accounts has been broder and has already included other tasks beyond ranking. Third, all published papers focused mostly on only one single problem (although there are some exceptions [2]), but the ideas, techniques, and lessons learned in solving an individual problem may be generalizable to other (related) problems. The proposed workshop would enable broadly engaging all researchers on this topic to discuss and summarize the general guidances on how to apply axiomatic thinking to a new problem, including particularly, (1) how to identify potential constraints; (2) how to verify those constraints; (3) how to apply these constraints to search for a solution; and (4) how to refine the constraints based on the evaluation results. The workshop would also facilitate sharing and discussing any useful negative findings across multiple problem domains.

2 THEME

As the title of the workshop suggested, the general theme of the workshop will be about all aspects of applications of axiomatic thinking to solve IR and IR-related problems. The basis of this general theme is the recent growth of work on applying axiomatic thinking to analyze and improve both retrieval models and evaluation metrics, which we expect to continue. The existing work has clearly demonstrated many advantages of axiomatic thinking, including particularly specific theoretical results in the form of novel constraints to be satisfied by retrieval functions or evaluation metrics and improved models or evaluation metrics. However, much more research is still needed in multiple directions.

Take, for example, the study of retrieval models. Retrieval constraints are often formalized based on retrieval heuristics, and the heuristics are strategies motivated by empirical observations. Unfortunately, there has been little attention on identifying and testing useful heuristics of retrieval models, and learning how to utilize them to develop effective retrieval models. Most heuristics lie in the minds of seasoned information retrieval researchers and exercising heuristics effectively is something of a black art. Clearly, it has come to a point where the constraints (i.e., heuristics) of retrieval models for various tasks should be revisited, organized, summarized and analyzed.

Also, it has been shown that, in most existing retrieval models, there exist some underlying heuristics that are closely related to the empirical performance. For example, many models include TF-IDF style term weighting. In addition, extensions to these models often also rely upon very similar techniques [15]. These sorts of extensions include term proximity and pseudo-relevance feedback. Ideally, as we discover these shared constraints, we would have more guidance on developing effective retrieval models. Since there exist some commonality among various retrieval tasks, discovering and summarizing the effectiveness of different constraints would deepen our understanding and open some new research directions.

Opportunities of applying axiomatic thinking also go beyond analyzing the basic retrieval functions; in fact, understanding constraints is also beneficial to many IR tasks that use machine learning techniques. Instead of having a designer carefully choose a set of assumptions to make when designing a formal model, these approaches use machine learning to weight items in a pool of features derived from many retrieval heuristics. However, this potentially results in a bloated backend which computes many features irrelevant to the task or collection. Having knowledge about relevant features would help slim down backends and speed learning and ranking. An important strength of the axiomatic methodology is that evaluation data sets become resources used to check motivated hypotheses instead of optimization mechanisms, which are at risk of overfitting.

There are even more opportunities for new research on applying axiomatic thinking to evaluation as has already been happening where researchers have done axiomatic analysis of metrics for tasks such as text categorization as we mentioned earlier in this proposal. In general, an understanding of how to apply axiomatic thinking to IR problems may become increasingly important as information retrieval continues to broaden into new areas. New tasks often require new constraints, and an understanding of these constraints can provide guidance on how to adapt existing methods or how to develop mew methods for the new tasks. For example, domain-specific IR tasks such as medical record search might require new retrieval constraints that can capture the domain knowledge.

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

1. What constraints are effective to improve retrieval performance independent of the underlying model?

2. 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?

3. 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?
4. Have new languages, media, or domains suggested new constraints for established domains?
5. 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?
6. How can axiomatic thinking be combined with machine learning techniques to learn more effective retrieval functions?

REFERENCES

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

[2] Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. 2013. 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)*. ACM, New York, NY, USA, 643–652.

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

[4] Luca Busin and S. Mizzaro. 2013. Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics. In *Proceedings of ICTIR 2013: 4th International Conference on the Theory of Information Retrieval*. ACM, New York – USA, 22–29.

[5] Hui Fang, Tao Tao, and Chengxiang Zhai. 2011. Diagnostic Evaluation of Information Retrieval Models. *ACM Trans. Inf. Syst.* 29, 2, Article 7 (April 2011), 42 pages. DOI: http://dx.doi.org/10.1145/1961209.1961210

[6] Hui Fang and ChengXiang Zhai. 2005. An exploration of axiomatic approaches to information retrieval. In *SIGIR '05*. 480–487. DOI: http://dx.doi.org/10.1145/1076034.1076116

[7] Marco Ferrante, Nicola Ferro, and Maria Maistro. 2015. Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. In *Proceedings of ICTIR 2015*. ACM, New York, NY, USA, 21–30. DOI: http://dx.doi.org/10.1145/2808194.2809452

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

[9] Yuanhua Lv and ChengXiang Zhai. 2011. Lower-bounding Term Frequency Normalization. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*.

[10] Eddy Maddalena and S. Mizzaro. 2014. Axiometrics: Axioms of Information Retrieval Effectiveness Metrics. In *Proceedings of the Sixth EVIA Workshop*. National Institute of Informatics, Tokyo, Japan, 17–24. ISBN: 978-4-86049-066-9.

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

[12] Alistair Moffat. 2013. Seven Numeric Properties of Effectiveness Metrics. In *AIRS'13*. 1–12.

[13] Fabrizio Sebastiani. 2015. An Axiomatically Derived Measure for the Evaluation of Classification Algorithms. In *Proceedings of ICTIR 2015*. ACM, 11–20. DOI: http://dx.doi.org/10.1145/2808194.2809449

[14] Marina Sokolova. 2006. Assessing Invariance Properties of Evaluation Measures. In *Proceedings of NIPS'06 Workshop on Testing Deployable Learning and Decision Systems*.

[15] ChengXiang Zhai. 2008. Statistical Language Models for Information Retrieval A Critical Review. *Found. Trends Inf. Retr.* 2, 3 (March 2008), 137–213.