The Axiometrics Project

Eddy Maddalena and Stefano Mizzaro

Department of Mathematics and Computer Science
University of Udine
Udine, Italy
eddy.maddalena@uniud.it, mizzaro@uniud.it

Abstract. The evaluation of retrieval effectiveness has played and is playing a central role in Information Retrieval (IR). To evaluate the effectiveness of IR systems, more than 50 (maybe 100) different evaluation metrics have been proposed. In this paper we sketch our Axiometrics project, that aims to a formal account of IR effectiveness metrics.

1 Introduction

Effectiveness evaluation is of paramount importance in Information Retrieval (IR). Several effectiveness metrics have been proposed so far. In a survey in 2006 [6], more than 50 metrics have been collected, taking into account only the system oriented effectiveness metrics; it is likely that about one hundred systems oriented metrics exist today, let alone user-oriented ones or metrics for tasks somehow related to IR, like filtering, clustering, recommendation, summarization, etc. As stated for example in [8], there is nothing close to agreement on a common metric that everyone will use. It is a diffuse opinion that different metrics evaluate different aspects of retrieval behavior [4,8]. Each of these metrics has its own advantages and limitations. Metric choice is neither a simple task, nor it is without consequences: an inadequate metric might mean to waste research efforts improving systems toward a wrong target. It is clear that a better understanding of the formal properties of effectiveness metrics would help. This paper describes the Axiometrics project [5,7]: we propose an axiomatic approach to effectiveness metrics and we aim at defining some basic axioms that any reasonable metric should satisfy and that are formulated in a general way.

2 Related Work

Although formal approaches have high importance in the IR field, they have mainly focussed on the retrieval process rather than on effectiveness metrics themselves. However, some research specific to effectiveness metrics does exist, and it is briefly discussed here. An early attempt has been made by Swets [10] who lists some desirable properties, as quoted in [11, p.119-120]. Also van Rijsbergen himself in [11, Chapter 7] follows an axiomatic approach. In [3], Bollmann

proposes the Axiom of monotonicity and the Archimedean axiom, and their implication is presented as a theorem. In Yao [13] a new effectiveness metric that compares the relative order of documents is proposed and proved to be appropriate through an axiomatic approach. More recently, Amigó et al. in [1] focus their formal analysis on evaluation metrics for text clustering algorithms finding four basic formal constraints and in [2] present a unified comparative view of proposed metrics for the task of document filtering.

3 Measurement and Similarity

We propose to rely on measurement theory [12] to formalize IR effectiveness metrics. Measurement is defined as a process aimed at determining a relationship between a physical quantity and a unit of measurement. A particularly discussed issue is how the measurement is expressed. Stevens proposed the four standard *measurements scales* [9]: Nominal, Ordinal, Interval, Ratio. This classification has become a tradition in various field and has provides useful insights, although alternatives exist.

The evaluation process in IR is based on two quantities: (i) an automated estimation, by an IR system, of the relevance of a document, and (ii) a human (user or assessor) estimation of the relevance of a document. We can exploit measurement to model these quantities: given a query, a system tries to *measure* the relevance of the documents to the query, for example to rank the documents; given (a description of) an information need, an assessor tries to *measure* the relevance of the documents to the need. We therefore have two kinds of *relevance measurements* (and *measures* as well): one made by a system and referred to in the following as *system relevance measure(ment)*, and one made by a human and referred to in the following as *human (or user/assessor) relevance measure(ment)*.

By using a notion of measure(ment) that is common to both system and human, we can define a notion of similarity among them. Ideally, an IR system should both: use the same measurement scale of the human assessor, and provide the same measurement of the human assessor. However, systems are far from being perfect, and therefore the very same measurement is almost never provided. The aim of an IR system is thus to provide the measurement σ that is most similar to the assessor / user measurement α. Moreover, often the scales are different: $\mathsf{scale}(\alpha)$ can be fixed a priori, e.g., when a test collection provides human relevance assessments, and $\mathsf{scale}(\sigma)$ depends on the retrieval algorithm at hand, and different approaches have different scales. Of course, two measurements expressed on two different scales can not be identical (e.g., a rank can not be identical to a measurement expressed on a category scale, the usual ad-hoc retrieval situation). Thus, similarity needs to be defined over different scales.

4 IR Effectiveness Metric

On the basis of the concepts of measurement, measurement scales, and similarity we now turn to modeling the effectiveness metrics itself. An effectiveness metric provides a numerical representation of the similarity between two relevance measurements. A metric is then a function that takes as arguments two measurements α and σ, a set of documents D, and a set of queries Q, and provides as output a numeric value (usually in $\mathbb{R}$): $\mathsf{metric} : \alpha \times \sigma \times D \times Q \mapsto \mathbb{R}$.

A metric is defined on the basis of five components: $\mathsf{scale}(\alpha)$ and $\mathsf{scale}(\sigma)$; a notion of similarity sim; how the values on single documents are averaged over a set of documents D (we denote the corresponding averaging function with avgD); and how these averages are averaged over a set of queries Q (avgQ). We can write: $\mathsf{metric}(\mathsf{scale}(\alpha), \mathsf{scale}(\sigma), \mathsf{sim}, \mathsf{avgD}, \mathsf{avgQ})$ to describe a metric.

By using suitable similarity functions, hopefully the framework can model most, if not all, known metrics [7].

5 Axioms and Theorems

In [7], we have proposed 13 axioms and 5 theorems: they define properties that, ceteris paribus, any effectiveness metric should satisfy. Given the space limits, we can only briefly sketch some of them.

Axiom 1 (Document monotonicity) *Let q be a query, D and D' two sets of documents such that $D \cap D' = \varnothing$, α a human relevance measurement and σ and σ' two system relevance measurements such that:*[1]

$$\underset{q,D}{\mathsf{metric}}(\alpha, \sigma) \underset{(=)}{\overset{>}{}} \underset{q,D}{\mathsf{metric}}(\alpha, \sigma')$$

and

$$\underset{q,D'}{\mathsf{metric}}(\alpha, \sigma) \underset{(=)}{\overset{>}{}} \underset{q,D'}{\mathsf{metric}}(\alpha, \sigma').$$

Then

$$\underset{q,D\cup D'}{\mathsf{metric}}(\alpha, \sigma) \underset{(=)}{\overset{>}{}} \underset{q,D\cup D'}{\mathsf{metric}}(\alpha, \sigma').$$

A similar axiom holds for query sets (omitted for space limits). Another axiom states that if system relevance measures on two documents d and d' are equally correct, system relevance of d is higher than system relevance of d', and d' is not less relevant than d, then the effectiveness metric should be more affected by d than by d' (represented by $\sqsupset$).

[1] In this axiom the equal = and greater than > signs have obviously to be paired in the appropriate way, "row by row". We use this notation for the sake of brevity.

Axiom 2 (System relevance) *Let q be a query, d and d' two documents, α and σ two (human and system) relevance measurements such that* $\mathsf{sim}_{q,d}(\alpha, \sigma) = \mathsf{sim}_{q,d'}(\alpha, \sigma)$*, $\sigma(d) > \sigma(d')$, and $\alpha(d) \geq \alpha(d')$. Then $d \sqsupset_{metric(\alpha,\sigma)} d'$.*

This entails as a corollary the often stated property that early rank positions affect a metric value more than later rank positions. A symmetric axiom can also be stated on user relevance measurement: a metric should weigh more, and be more affected, by more relevant documents. This is perhaps less intuitive than the previous one, but it does indeed seem natural in this framework.

6 Conclusions and Future Work

We propose a framework based on the notions of measure, measurement, and similarity to define axioms and to derive theorems on IR effectiveness metrics. Our approach aims to a threefold contribution: (i) the proposal of using measurement to model in a uniform way both system output and human relevance assessment, and the analysis of the different measurement scales used in IR; (ii) the notions of similarity among different measurement scales and the consequent definition of metric; and (iii) the axioms and theorems. In the future, we will seek for new axioms and theorems that can allow us to define and discover new metrics property. We will also focus on aspects such as: filtering, recommendation, reformulation, summarization, novelty, and difficulty of queries.

Acknowledgments

We thank Julio Gonzalo and Enrique Amigó for long and interesting discussions, Evangelos Kanoulas and Enrique Alfonseca for helping to frame the Axiometrics research project, Arjen de Vries for suggesting the name "Axiometrics", and organizers of (and participants to) SWIRL 2012. This work has been partially supported by a Google Research Award.

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 comparison of evaluation metrics for document filtering. In *CLEF*, volume 6941 of *LNCS*, pages 38–49. Springer, 2011.
3. P. Bollmann. Two axioms for evaluation measures in information retrieval. In *SIGIR '84*, pages 233–245, Swinton, UK, 1984. British Computer Society.
4. C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *SIGIR '00*, pages 33–40, New York, NY, USA, 2000. ACM.
5. L. Busin and S. Mizzaro. Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics. In *ICTIR 2013 — Proceedings of the 4th International Conference on the Theory of Information Retrieval*, pages 22–29, 2013.

6. G. Demartini and S. Mizzaro. A Classification of IR Effectiveness Metrics. In *ECIR 2006*, volume 3936 of *LNCS*, pages 488–491, 2006.
7. E. Maddalena and S. Mizzaro. Axiometrics: Axioms of information retrieval effectiveness metrics. In *Proceedings of the Second Australasian Web Conference*. Australian Computer Society, Inc., 2014, to appear.
8. S. Robertson. On GMAP: and other transformations. In *CIKM '06*, pages 78–83, New York, USA, 2006.
9. S. S. Stevens. On the theory of scales of measurement. *Science*, 103 (2684):677–80, 1946.
10. J. A. Swets. Information retrieval systems. *Science*, 141:245–250, 1963.
11. C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 2nd edition, 1979.
12. Wikipedia. Measurement — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Measurement`, 2012. [Last visit: October 2013].
13. Y. Y. Yao. Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information Science*, 46(2):133–145, 1995.