1. Purpose of the visit;

The past two years Julio Gonzalo and Enrique Amigò (at UNED), together with Stefano Mizzaro (at Udine University), have been working on the axiomatic definition of evaluation measures for Information Retrieval, i.e. a framework for defining, analysing and understanding the relationships among evaluation measures, in terms of both axiomatic properties and statistical relations. This has been one of the fundamental problems in the field of Information Retrieval; to this day, more than 100 evaluation measures exist leaving researchers and practitioners in the dark regarding the metric they should be using to optimize their retrieval system and algorithms.

The project, at its early stages, was funded by a Google Award; I was the sponsor and coordinator of the efforts of the two groups (Udine University and UNED). The two groups developed two independent approaches on the topic (and a number of papers was published around them): a set of formal properties metrics should hold by Amigò and Gonzalo and a measurement theory framework by Mizzaro. The two approaches are however compatible and can be fruitfully combined.

The main purpose of the meeting is to work on deep integration of the two approaches.
The outcome of this meeting is three-fold: (a) a research paper on the general framework that integrates both approaches, (b) a joint tutorial, already accepted at the ECIR 2014 conference, and (c) an already agreed monograph in the Morgan-Claypool Synthesis Series on Information Concepts, Retrieval, and Services edited by Gary Marchionini.

2. Description of the work carried out during the visit;

The work carried out during the visit covered all three points above (as it can be seen from the outcomes of this meeting). We further discussed and identified future open problems and common ground for further collaboration. Last, we designed a tutorial on the basis of our discussions and research and wrote and submitted a proposal to SIGIR 2015.

3. Description of the main results obtained;

The result of the meeting was tri-fold:

1. We concluded to a common terminology across the criterion-based and axiom-based approaches. On the basis of this common terminology we expressed axioms and theorems from one line of research in to the other, creating a common framework (which was one of the main goals of the visit). You can find this common framework of terms, axioms and theorems attached at the end of this report.
2. We identified and decided upon the structure of the monograph. You can also find the outline of it attached.
3. We wrote and submitted a SIGIR 2015 tutorial proposal (also attached at the end of this report).

4. Future collaboration with host institution (if applicable);

Together with Julio Gonzalo and Enrique Amigò (at the host institution) we are working towards a book that integrates the constraint-based meta-evaluation of effectiveness measures and the axiomatic definition of effectiveness measures.

Further we have submitted, as an outcome of this meeting, a joined Tutorial proposal at SIGIR 2015 that covers the exact same topic as the book.

Finally, we expect our collaboration to last; we have identified complex experimental setups (e.g. experiments on novelty and diversity of search results, experiments through user sessions) over which the proposed methods could be extended.

Axiometrics

February 26, 2015

1 Formal Constraints for Ordering Based Measures

These axioms work under the assumption that the user explores documents in a priority ordering given by the information retrieval system.

Axiom 1.1. *Swaping a relevant document with an irrelevant document with higher priority in the system output improves the system output quality.:*

$$\textit{If } \alpha(i) > \alpha(j) \textit{ and } \sigma(j) > \sigma(i) \textit{ then}$$

$$\mathcal{M}(\sigma) < \mathcal{M}(\sigma_{i\leftrightarrow j})$$

where

$$\sigma_{i\leftrightarrow j}(d) = \begin{cases} \sigma(i) & \text{if } d = j \\ \sigma(j) & \text{if } d = i \\ \sigma(d) & \text{if } i.o.c \end{cases}$$

According to every ordering based user models, a high priorized document in the system output has more probability to be explored by the user. Therefore, an error in this

Axiom 1.2. *Top Hevyness: A swaping in a high priority level affects the measure to a greater extent than a swap in a low priority level:*

$$\textit{If } Subseq(\sigma, i, j) \textit{ and } Subseq(\sigma, k, l) \textit{ and}$$

$$\alpha(i) = \alpha(k) > \alpha(j) = \alpha(l) \textit{ then}$$

$$\mathcal{M}(\sigma_{i\leftrightarrow j}) < \mathcal{M}(\sigma_{k\leftrightarrow l})$$

where

$$Subseq(\sigma, i, j) \equiv \sigma(i) > \sigma(j) \wedge \neg\exists k\,(\sigma(i) > \sigma(k) > \sigma(j))$$

The next axiom is grounded on the idea that the documents are showed to the user in an ordinal manner. Therefore there exists an area which is always explored and an area which is never explored. The following axiom states that a few relevant documents is better than only one at the top, but only one is better than a huge amount at the bottom of the ranking.

Axiom 1.3. *Deepnes Threshold: Being σ_n and σ'_n the two system outputs such that:*

$$(Rank(\sigma_n, i) = 1 \rightarrow \alpha(i) = 1) \wedge (Rank(\sigma_n, i) \in \{2..n\} \rightarrow \alpha(i) = 0)$$

$$(Rank(\sigma'_n, i) \in \{1..n/2\} \rightarrow \alpha(i) = 0) \wedge (Rank(\sigma_n, i) \in \{n/2..n\} \rightarrow \alpha(i) = 1)$$

then

$$\exists th | n > th \rightarrow \mathcal{M}(\sigma_n) > \mathcal{M}(\sigma'_n)$$

$$\exists th | 2 < n < th \rightarrow \mathcal{M}(\sigma_n) < \mathcal{M}(\sigma'_n)$$

In general, system results are returned as a list of items. This enforces a priority relationship between all documents in the system output. However, we can be more strict, saying that a false priority relationship must be penalyzed

Axiom 1.4. *Stating a priority difference between equally relevant documents decreases the system effectiveness. If $\alpha(i) = \alpha(j)$ and $\sigma(i) = \sigma(j)$ then:*

$$\mathcal{M}(\sigma_{i>j}) < \mathcal{M}(\sigma)$$

where

$$\sigma_{i>j}(i) > \sigma_{i>j}(j) \; and \, \forall (k,l) \neq (i,j) \, (\sigma(i) > \sigma(j) \leftrightarrow \sigma(i) > \sigma(j))$$

This constraint implies that, if a system return only irrelevant documents, the more the ranking is short, the more the system is better.

2 Formal Constraints for Measures in Absolute Values

In some cases, the information access task requires predicting the exact feature of information pieces. For instance, in polarity detection a user can be interested in negative or positive opinions about a certain entity, but not neutral. For these cases we have to consider the system output in an absolute scale. Some typical measures for this are accuracy ($P(\sigma(x) = \alpha(x)$)or average error ($Avg_d(\sigma(d) - \alpha(d))$). For this, we can define the following axiom:

Axiom 2.1. *Absolute Scale Monotonocity: Decreasing the difference in a certain direction between the system output and the gold increases the measure value (Mizzaro's axiom). Being $\forall d \neq i(\sigma'(d) = \sigma(d))$ if:*

$$\sigma(i) > \sigma'(i) > \alpha(i) \vee \sigma(i) < \sigma'(i) < \alpha(i)$$

then

$$\mathcal{M}(\sigma') > \mathcal{M}(\sigma)$$

This axiom prevents the cases in which there exists a trade-off between over and under estimated reliability.

In this scenario, assuming topheavyness does not have too much sense, given that the user does not explore the documents in an ordinal manner. Thus, any α range can be the user target, not necessarily the top values.

3 Formal Constraints for Measures in Interval Scales

In some cases, we are interested in the interval scale. This is the typical case of *predicting variables*. In this case, the absolute values are not predicted, the absolute ranges are unknown. Thus, the user has to access document by priority, but he knows when there exists a higher or lower difference between contiguous document (unlike in the ordering scale).

This match with the popular Pearson correlation coefficient. Instead of ordering or predicting the extact values, the ordering between differences must be consistant. For instance, in a perfect Pearson correlation (linear correlation), the absolute values do not necessarily matches, but there is a perfect correlation between σ and α differences:

$$\sigma(i) - \sigma(j) > \sigma(k) - \sigma(l) \leftrightarrow \alpha(i) - \alpha(j) > \alpha(k) - \alpha(l)$$

I did not work on it yet, but my intuition is that we could extrapolate the axioms for ordering scale to this case, just considering the order of differences instead of σ and α values.

References

[1] Enrique Amigó, Julio Gonzalo, and Felisa 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.

Contents

CHAPTER 1

The Big Bang: Introduction [k]

1.1 EVALUATION METHODOLOGIES

- ▷ User studies (Side-by-side, Eye-tracking, User models)
- ▷ Offline (Static collection benchmark)
- ▷ Online (A/B testing, Interleaving)

1.2 EVALUATION METRICS

- ▷ Tagcloud

1.3 AIMS AND METHODOLOGY

- ▷ What we want to do (understanding measures better)
- ▷ How do we do it: with axioms/constraints
- ▷ How we explain it: top-down (from general to specific)

CHAPTER 2

Looking at the Night Sky: Evaluation Metrics [k]

▷ Definitions of some (¡10) metrics used in IR and re-used later on

▷ Precision/Recall, MAP, NDCG, P@n, MRR, ERR, RBP, Q-m, TBG

▷ (ADM?)

▷ Enriques figure: metrics start being either precision or recall oriented; most recent metrics have a balance.

CHAPTER 3

Messages from Outer Space: Theoretical Background

3.1 METRIC SPACES [A]

- ▷ Evaluation metrics are not metrics in the mathematical sense because they are not symmetric, because of triangular inequality, maximality, etc.

3.2 MEASUREMENT THEORY (SCALES)

- ▷ Evaluation metrics are not measure theory compliant

- ▷ Evaluation metrics can be classified on the basis of the scales they are based on (Stefanos figure, adapted)

3.3 SIMILARITY

- ▷ We talk about similarity between things with possibly different structure – similarity between a ranking and a binary classification – . So we mean compatibility or equivalence. It fits with Tverskys take on similarity.

CHAPTER 4

Our Telescope: A General Framework [s+k]

4.1 BASIC DEFINITIONS AND NOTATION [S]

- ▷ IR as measurement of relevance
- ▷ `\alpha, \sigma, \sim`,
- ▷ `\metric`? metric components?

4.2 ON THE GENERALITY OF THE NOTATION [S+K]

4.2.1 EVALUATION METHODOLOGIES [K]

- ▷ TREC-style, Side-by-side?, A/B testing?, e.g.
- ▷ `\alpha` defined on pairs of documents, and ¡d1, d2, d3¿, with d2 receiving a click. Then M(¡d1, d2, d3¿) ¡ M(¡d2, d1, d3¿)

4.2.2 EFFECTIVENESS METRICS [S]

- ▷ Stefanos table

CHAPTER 5

A Trip to the Stars: General Axioms on Information Retrieval [s]

5.1 AXIOMS

5.2 METRIC ANALYSIS (?)

CHAPTER 6

Landing on Earth: Axioms on Ad-Hoc Retrieval [j]

6.1 AXIOMS

6.2 METRIC ANALYSIS

CHAPTER 7

Exploring other Planets 1: Axioms on Clustering [j]

CHAPTER 8

Exploring other Planets 2: Axioms on Filtering [j]

CHAPTER 9

The Big Crunch: A General Document Organization Task [j]

9.1 DOCUMENT ORGANIZATION: MERGING RETRIEVAL, FILTERING AND CLUSTERING

- Motivation
- Definition

9.2 RELIABILITY AND SENSITIVITY

- Axioms
- Metric Analysis

CHAPTER 10

Parallel Universes: Related Work [k]

10.1 EMPIRICAL META-EVALUATION OF METRICS

- ▷ Informativeness - Learning to Rank (Yilmaz, et al.)
- ▷ Discriminative power (Sakai)
- ▷ Generalizability (Kanoulas & Aslam)

10.2 RELATED WORK ON AXIOMATIC DEFINITIONS

- ▷ Oldies
- ▷ Alistair, AIRS 2013
- ▷ Radlinski, SWDM 2013
- ▷ Hui Fang, ChengXiang Zhai
- ▷ Castells on diversity

CHAPTER 11

A very short summary of nearly everything, and more [k]

11.1 CONCLUSIONS

11.2 FUTURE WORK

Author’s Biography

FIRSTNAME LASTNAME

FirstName LastName ...

A General Account of Effectiveness Metrics for Information Tasks: Ranking, Filtering, and Clustering

[Tutorial Proposal for SIGIR 2015]

Enrique Amigó
nlp.uned.es
E.T.S.I. Informática, UNED
c/ Juan del Rosal, 16, 28040
Madrid, Spain
enrique@lsi.uned.es

Julio Gonzalo[*]
nlp.uned.es
E.T.S.I. Informática, UNED
c/ Juan del Rosal, 16, 28040
Madrid, Spain
julio@lsi.uned.es

Evangelos Kanoulas
Informatics Institute
University of Amsterdam
Science Park 904, 1098 XH
Amsterdam, The Netherlands
e.kanoulas@uva.nl

Stefano Mizzaro
Dept. of Math. & CompSci
Udine University
Via delle Scienze, 206, 33100
Udine, Italy
mizzaro@uniud.it

ABSTRACT

In this tutorial we will present, review, and compare the most popular evaluation metrics for some of the most salient information related tasks, covering: (i) Ranking, (ii) Clustering, and (iii) Filtering.

The tutorial will make a special emphasis on the specification of constraints for suitable metrics in each of the three tasks, and on the systematic comparison of metrics according to how they satisfy such constraints. This comparison provides criteria to select the most adequate metric or set of metrics for each specific information access task. The last part of the tutorial will investigate the challenge of combining and weighting metrics.

1. LENGTH AND INTENDED AUDIENCE

We propose a full-day (6 hours) tutorial.

The tutorial contains material suitable both for novices and experts, but it is probably better classified as "intermediate" if not "advanced". Familiarization with Ranking, Filtering, and Clustering is recommended, as well as a basic understanding of effectiveness evaluation methodologies.

2. INSTRUCTORS

The tutorial is given by four instructors. Their bios and publications show their expertise in the topics presented in the tutorial.

2.1 Brief Biographies

Enrique Amigó (UNED, Madrid, Spain) is associate professor at UNED and member of the nlp.uned.es research group. He has published several papers (in venues such as SIGIR, ACL, EMNLP, Journal of Artificial Intelligence Research, Information Retrieval journal, etc.) on evaluation methodologies and metrics for Text Summarization, Machine Translation, Text Clustering, Document Filtering, etc. Publications: `http://scholar.google.com/scholar?hl=en&q=enrique+amigo`

Julio Gonzalo (UNED, Madrid, Spain) is head of nlp.uned.es, the UNED research group in Natural Language Processing and IR. He has recently been CLEF 2011 General Co-Chair, Area Chair for EACL 2012, ECIR 2012 and EMNLP 2010, and co-organizer of the RepLab 2012/2013 Evaluation Campaigns for Online Reputation Management Systems and the WePS evaluation campaign on Web People Search systems. His research interests include Cross-Language and Interactive IR, Search Results Organization, Entity-Oriented and Semantic Search, and Evaluation Methodologies and Metrics in Information Access. Publications: `http://scholar.google.com/citations?user=opFCmpYAAAAJ&hl=en`

Evangelos Kanoulas (U. Amsterdam, Netherlands) is an assistant professor at the University of Amsterdam. His expertise lies in the field of information retrieval, with specializations in experimental design, evaluation methodology, and statistical analysis. In 2010 he was awarded the Marie Curie Fellowship to explore the efficient and effective evaluation of information retrieval systems. Evangelos has extensively published his work on IR evaluation in top-tier conferences in the field, including SIGIR, CIKM, ECIR, and VLDB. Since 2007 together with others he has proposed and organized numerous search benchmark exercises under the umbrella of TREC (Million Query, Sessions, Tasks tracks). Since 2014 he is a member of the steering committee of CLEF. In the past, he gave two succesful full-day SIGIR tutorials, in 2010 and 2012, on Low-Cost Evaluation in IR and on Advances on the Development of Evalua-

[*]Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR 2015 Santiago de Chile, Chile
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

tion Measures, respectively. Further, in 2011 together with others he taught a full-week course on IR Evaluation, in RuSSIR/EDBT 2011 Summer School. Finally he has given numerous talks on the topic and he has taught the graduate IR course in two universities for two semesters. Publications: https://scholar.google.com/citations?user=0HybxV4AAAAJ

Stefano Mizzaro (U. Udine, Italy) is associate professor at Udine University since 2006, and in 2014 he spent his sabbatical at RMIT in Melbourne. Although, broadly speaking, his current research interests include IR, digital libraries, and mobile contextual information access, he has been focussing on IR evaluation for the last 20 years. He has been working on user evaluation, novel effectiveness metrics, and mining of test collection data. He has been university researcher (assistant professor) from 2000 to 2006. He published about 100 refereed papers, several as a single author, received two international awards, for two best papers, and authored two books on Java programming. He had an active role in several research projects at regional, national, and European level. He has been a lecturer at the ESSIR in 2005. In December 2013 he obtained the full professorship habilitation in Italy. Publications: http://scholar.google.com/citations?user=2wvJC6IAAAAJ&hl=en

2.2 Relevant Experience

The authors have published a number of papers related to Evaluation Metrics in IR and Natural Language Processing, and to evaluation more in general. Some instances are: a full paper in SIGIR about general constraints for IR metrics [6], a paper in IR Journal about clustering metrics [3], a paper in JAIR Journal about combining metrics [4]. Some papers specifically analyzed IR effectiveness metrics and proposed novel ones [18, 31, 17, 16, 32, 26, 38, 27, 13, 14], and some others concern the evaluation methodology more in general [33, 21, 8, 35, 12, 36, 42, 25, 39].

Recently, the second and fourth authors jointly received a Google Faculty Research Award, under the sponsorship of the third author, on the topic of “Axiometrics”, which is the unified view of evaluation metrics that will be used in the tutorial. The first results of the Axiometrics project have been published recently [6, 9, 29, 28].

Overall, the team of instructors has a long history of teaching undergraduate and graduate classes, tutorials, and courses at summer schools. Three of the four instructors presented the previous editions of this tutorial discussed below in Section 5.2; the other one has already given two SIGIR tutorials in the past.

3. MOTIVATIONS AND BACKGROUND

For most problems in Ranking, Clustering and Filtering, there are many competing evaluation metrics in the literature, and in general there is no clear procedure to choose the most adequate in a specific task/scenario. In practice, the tendency is often to choose the most popular metric (which has a snowball effect that tends to prefer the oldest metrics). In addition, there is often a lack of clear criteria to assign relative weights when combining metrics (e.g., precision and recall). In practice, the tendency is also to choose the most popular weighting scheme.

This tutorial relies on some recent results, that we have obtained applying measurement theory to derive properties, constraints, and axioms of effectiveness metrics and metric combinations for all the three above mentioned fields.

4. OBJECTIVES

The overall tutorial aim is to describe effectiveness metrics with a general approach, to analyze their properties within a conceptual framework, and to provide tools to select the most appropriate metric when needed. More in detail, the specific goals of this tutorial are:

- To provide an overall introduction to retrieval, clustering, and filtering effectiveness metrics; we will discuss information retrieval tasks, user models that associate with these tasks, and effectiveness metrics defined over these user models.
- To seek generality by analyzing several metrics, and from three different fields (besides retrieval, also clustering and filtering). Of course the presentation will be IR-centric, but some properties and results will be better presented and understood by referring to clustering and filtering.
- To provide a general framework based on measurement theory to understand and define metrics and to state metric axioms.
- To discuss desirable basic constraints that should be satisfied by metrics. On the basis of these constraints, provide a taxonomy of metrics and discuss how different metric families satisfy different constraints.
- To provide the attendees the tools for selecting an appropriate metric for each user specific scenario.
- To understand the effect of weighting metrics arbitrarily; we will give tools based on measurement theory to check the robustness of evaluation results in this sense.
- To discuss empirical meta-evaluation methods to assess the discriminative power, the generalizability, and the informativeness of evaluation metrics.

5. RELEVANCE

5.1 The Importance of Evaluation

Effectiveness evaluation is of paramount importance in IR, which has been one of the most evaluation-oriented fields in computer science since the first IR systems were developed in the late 1950s. All IR conferences feature evaluation sessions; papers on evaluation are continuously being published in IR journals; a Dagsthul seminar http://www.dagstuhl.de/13441 was on IR evaluation.

Within any evaluation methodology, the metric being used is a fundamental parameter. Figure 1 shows a tagcloud of most IR effectiveness metrics. A survey in 2006 [18] counted more than 50 effectiveness metrics for IR, taking into account only the system oriented metrics. In an extended version of the survey [19], yet unpublished, about one hundred IR metrics are collected, let alone user-oriented ones or metrics for tasks somehow related to IR, like filtering, clustering, recommendation, summarization, etc.

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. However, some researchers simply do not investigate into

Average Distance Measure
PRecall adopted to XML
Generality
Average Gain Ratio
Normalized Recall
Brookes's S measure
Rpref
Ranked Half Life
Fallout/R curve
Weighted R-Precision
Estimated Ratio of Relevant
Relative Relevance
Recall
Recall Effort
Generalized Success@10
Passage bpref
Generalized AP
Binary Preference
Swets's E measure
MAP
Fallout
R-measure
Normalized Precision
Highlighting XML Evaluation
Logarithm of Found Position
Mean Reciprocal Rank
Inferred AP
Bpref-10
NDCG at rank r
Reciprocal Rank
Average Weighted Precision
Constant Penalty Natural Order Imposition
Egghe's measure
yaAP
Instance Precision
Average Precision
Interpolated MAP
GMAP
Passage Precision
F-measure
Polynomial Natural Order Imposition
Normalized xCG
Instance Recall
Success at rank r
area
Average Search Length
Utility
Q-measure
Precision at rank r
k-call
E-measure
Ranking Efficiency
Classification Accuracy
P/R curve
Expected Precision-Recall with User Modelling
R-Gain Ratio
Mean Average Effort-Precision
Normalized DCG
Rank-Biased Precision
Precision and Recall with overlap
Mean Average nxCG
Precision Averaged over Recall Points
R-Precision
Tolerance to Irrelevance
Average Search Length Transformed
Precision-Recall with User Modelling
Precision
Kendall, Spearman
DCG
Expected Search Length
Effort-Precision/Gain-Recall
Normalized Distance-Based Performance Measure
Passage Recall
Coverage Ratio
Normalized Search Length
Novelty Ratio
ESL Reduction Factor
Relative Recall
Sliding Ratio

Figure 1: IR effectiveness metrics (from [5])

the suitability of the metric for the problem itself and they seem to choose just the most popular metrics for their experiments. We cannot exclude the temptation for researchers to choose, among all available metrics, those that help corroborating their claims, or even to design a new metric to this aim. In addition, it is not clear what to do when two metrics disagree.

We firmly believe that a better understanding of metrics, and of their conceptual, foundational, and formal properties, would help to avoid wasting time in tuning retrieval systems according to effectiveness metrics inadequate to specific purposes, and it will also induce researchers to make explicit and clarify the assumptions behind metrics.

5.2 Tutorial History

Two of the four instructors have previously held an initial edition of this tutorial at SPIRE 2012 (Cartagena de Indias, Colombia), and as a part of a tutorial on Information Access metrics, and at the 2nd Open Interdisciplinary Mumia Conference (Cyprus 2013). Three of the four instructors have held a similar tutorial at SIGIR 2014 in Australia in July 2014 [5] and they will held another edition at ECIR 2015 in Vienna in early April 2015. The SIGIR 2014 tutorial was well received with 14 participants and good feedback.

Differently from early editions, in this tutorial we do not discuss text evaluation issues like Machine Translation or Summarization. Instead, we include aspects of measurement theory and more details about IR metrics which seem more adequate to SIGIR audience and did not appear previously. In particular, differently from the last two editions, we now plan to describe IR metrics systematically in terms of their implicit user models, and we include a discussion on empirical meta-evaluation criteria for evaluation metrics. This leads to a full day tutorial (6 hours) whereas the previous editions were shorter (3 hours). Finally, this would be the first edition to be held in the Americas.

6. FORMAT AND DETAILED SCHEDULE

The tutorial is divided into the following eight parts (for each part, the speaker and the duration is shown).

6.1 Introduction: [all, 15m]

After a brief presentation of the speakers, we will focus on the tutorial structure and on its relevance and importance.

6.2 Tasks, Models, and Metrics [EK, 60m]

in this first part we will present IR effectiveness metrics. We will start by discussing different retrieval tasks/scenarios and user models that associate with these scenarios. We will follow this discussion by describing models of user interactions with the search results, and effectiveness metrics defined on the basis of these models.

Clearly it is unfeasible to provide a complete coverage of all the above mentioned metrics in a tutorial, and probably it does not make much sense as well. What is probably more sensible, and what we will do, is to provide a general analysis and a classification based on [18, 19, 10] that should be useful to understand the IR metrics, as well as the definition of the most used and important ones.

6.3 Measurement Theory and Basic Axioms [SM, 90m]

Measurement theory (see, e.g., Measurement and Level_of_Measurement on Wikipedia) is a valuable tool to understand metrics. We will briefly introduce measurement theory and we will then show how it can be exploited to model IR metrics. Measurement theory will be shown to be useful both as a general framework where to define metrics and metric axioms, and as a practical tool to understand what is wrong about certain metrics.

6.4 Meta-evaluating Ranking Metrics with Formal Constraints [JG, 30m]

Metric meta-evaluation can be defined as the process of evaluating metrics themselves. In most of cases, metrics are meta-evaluated in terms of stability across data sets [11], discriminative power [40], or sensitivity in terms of statistical significant differences between systems [34]. However, these criteria do not necessarily reflect the suitability of metrics for evaluation purposes, that is, to understand to what extent a higher scored system is better than another one. Again, we will focus on basic properties that any metric should satisfy: we show how to meta-evaluate and categorize metrics in terms of a basic, intuitive set of formal constraints, and we will show that most of existing metrics fail on most of constraints, and how the most current metrics tend to satisfy most of them.

6.5 Other Tasks [JG+EA, 60m]

In an attempt to provide a general account, we do not restrict to IR metrics only and we discuss the metrics, and their properties, for two IR related tasks: clustering and filtering. This will allow to emphasize common properties, problems, and solutions.

6.5.1 Clustering Metrics [JG, 30m]

The evaluation of clustering tasks is non-trivial and still subject to discussion. We will start by reviewing some of the many effectiveness metrics that have been proposed for clustering, such as Purity and Inverse Purity (usually combined via Van Rijsbergen's F measure), Clusters and class entropy, VI measure, Q0, V-measure, Rand Statistic, Jaccard Coefficient, Mutual Information, etc.

Similarly to what we have done for IR metrics, we will then analyse current clustering metrics in terms of a few constraints. Although previous work on constraints for clustering metrics exists [30, 20], the constraints described in the tutorial have the following features: (i) they are intuitive and clarify the limitations of each metric; (ii) they discriminate metric families, grouped according to their mathematical foundations, pointing the limitations of each metric family rather than individual metric variants; (iii) they are strict enough to discard most of current metrics; (iv) they can be checked formally (some previously proposed constraints can only be checked empirically); and (v) they cover the basic intuitions of other constraint sets, like those in [30, 20].

Finally, we will show how these constraints can be exploited for meta-evaluating clustering metrics.

6.5.2 Filtering Metrics [EA, 30m]

We then turn to effectiveness metrics for information filtering. Filtering involves a wide set of IR tasks such as spam detection [15], IR over user profiles [23], or post retrieval for on-line reputation management [2].

Although document filtering is simple to define, there are a wide range of different evaluation metrics that have been proposed in the literature, all of which have been subject to criticism. Just as an illustration, TREC has organized at least three filtering tasks, all of them using different evaluation metrics [24, 15, 22, 1]. We briefly survey main filtering metrics and then we discuss how finding an optimal evaluation metric for filtering is, indeed, a challenging problem. Even metrics that satisfy the basic constraints, however, can say rather different things about comparative systems effectiveness. We also present some experimental results that show an extremely low correlation between metrics employed in different evaluation campaigns.

We then turn to understanding the aspects that determine which is the most appropriate filtering metric for a given scenario. We analyze three mutually exclusive features (that can be expressed also as formal constraints) that help classifying evaluation metrics, meta-evaluating them, and selecting the most appropriate in a given application scenario.

6.6 Combining Metrics [EA, 45m]

Quite often, an information related tasks cannot be evaluated with a single quality criterion, and some sort of weighted combination is needed to provide system rankings. A well known example is the F measure [41] that combines precision and recall by computing their harmonic mean. A problem of weighted combination metrics is that relative weights are established intuitively for a given task, but at the same time a slight change in the relative weights may produce substantial changes in the system rankings. An overall improvement in the combined metric is often caused by an improvement in one of the individual metrics at the expense of a decrease in the other. Indeed, in this last (and more advanced) part of the tutorial we analyze empirical results showing that an important amount of research results are actually sensitive to the particular metric weighting scheme in the combination.

We will then analyze the theoretical basis supported by the measurement theory that limit the conditions in which an evaluation result is independent from arbitrary metric weighting. In addition, we show techniques that allow to quantify to what extent an evaluation result is robust under changes in metric weighting.

6.7 Empirical Meta-evaluation Methods [EK, 45m]

All previous sections discuss formal methods from measurement theory that allows the meta-evaluation of effectiveness metrics on the basis of axions, and constraints. However, there is also a number of empirical methods that evaluate different aspects of effectiveness metrics: (a) how discriminative a metric is, i.e. whether it can easily say apart, in a statistically significant manner, two comparing systems [37], (b) how informative a metric is, i.e. given the value of a metric how well can one predict the ranked list of relevant and non-relevant documents [7], and (c) how generalizable the metric is, i.e. whether comparison conclusions drawn by the use of a metric over a sample of queries can be generalized to the population [26]. In this section we will discuss the three methods and how they can prove useful in choosing the right IR metric.

6.8 Summary and Wrap-up [all, 15m]

We summarize the main topics and results and we hint at some future work. We will refer again to our broad approach, by recalling again the analyzed commonalities and variabilities across IR, clustering, and filtering, and discuss the generality of our approach. We will briefly sketch future developments of this research area.

7. SUPPORT MATERIALS

We will provide a copy of all the slides, plus an annotated bibliography of the relevant papers (see the citations above). We also plan to distribute a draft of the metrics survey paper [19] and a draft of an in-progress book that we are currently working on (both of them should be finished by July 2015).

8. REFERENCES

[1] E. Amigó, J. Artiles, J. Gonzalo, D. Spina, B. Liu, and A. Corujo. WePS3 Evaluation Campaign: Overview of the On-line Reputation Management Task. In *2nd Web People Search Evaluation Workshop (WePS 2010), CLEF 2010 Conference, Padova Italy*, 2010.

[2] E. Amigó, A. Corujo, J. Gonzalo, E. Meij, and M. de Rijke. Overview of replab 2012: Evaluating online reputation management systems. In P. Forner, J. Karlgren, and C. Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

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

[4] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. Combining evaluation metrics via the unanimous improvement ratio and its application to clustering tasks. *J. Artif. Int. Res.*, 42(1):689–718, Sept. 2011.

[5] E. Amigó, J. Gonzalo, and S. Mizzaro. A general account of effectiveness metrics for information tasks: retrieval, filtering, and clustering. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1289–1289. ACM, 2014.

[6] 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, 2013.

[7] J. A. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analyzing retrieval measures. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in*

Information Retrieval, SIGIR '05, pages 27–34, New York, NY, USA, 2005. ACM.

[8] A. Berto, S. Mizzaro, and S. Robertson. On using fewer topics in information retrieval evaluations. In *Proceedings of the ICTIR 2013 Conference on the Theory of Information Retrieval*, pages 30–37, New York – USA, Oct. 2013. ACM.

[9] 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.

[10] B. Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 903–912, New York, NY, USA, 2011. ACM.

[11] B. Carterette. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 903–912, New York, NY, USA, 2011. ACM.

[12] B. Carterette, E. Kanoulas, V. Pavlu, and H. Fang. Reusable test collections through experimental design. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 547–554, New York, NY, USA, 2010. ACM.

[13] B. Carterette, E. Kanoulas, and E. Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 611–620, New York, NY, USA, 2011. ACM.

[14] B. Carterette, E. Kanoulas, and E. Yilmaz. Incorporating variability in user behavior into systems based evaluation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 135–144, New York, NY, USA, 2012. ACM.

[15] G. Cormack and T. Lynam. Trec 2005 spam track overview. In *Proceedings of the fourteenth Text Retrieval Conference 8TREC 2005)*, 2005.

[16] V. Della Mea, G. Demartini, L. Di Gaspero, and S. Mizzaro. Experiments on average distance measure. In M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors, *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006*, volume 3936 of *Lecture Notes in Computer Science*, pages 492–495, Londra, GB, Apr. 2006. Springer.

[17] V. Della Mea and S. Mizzaro. Measuring retrieval effectiveness: A new proposal and a first experimental validation. *Journal of the American Society for Information Science and Technology*, 55(6):530–543, 2004.

[18] G. Demartini and S. Mizzaro. A classification of IR effectiveness metrics. In M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors, *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006*, volume 3936 of *Lecture Notes in Computer Science*, pages 488–491, Londra, GB, Apr. 2006. Springer.

[19] G. Demartini, S. Mizzaro, and F. Scholer. A survey and classification of information retrieval effectiveness metrics. Draft.

[20] B. E. Dom and B. E. Dom. An information-theoretic external cluster-validity measure. Technical report, Research Report RJ 10219, IBM, 2001.

[21] J. Guiver, S. Mizzaro, and S. Robertson. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Transactions on Information Systems*, 27(4):1–26, 2009.

[22] B. Hedin, S. Tomlinson, J. R. Baron, and D. W. Oard. Overview of the trec 2009 legal track, 2009.

[23] K. Hoashi, K. Matsumoto, N. Inoue, and K. Hashimoto. Document filtering method using non-relevant information profile. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 176–183, New York, NY, USA, 2000. ACM.

[24] D. A. Hull. The TREC-7 filtering track: description and analysis. In *Proceedings of TREC-7, 7th Text Retrieval Conference*, pages 33–56, Gaithersburg, US, 1998.

[25] T. Jones, A. Turpin, S. Mizzaro, F. Scholer, and M. Sanderson. Size and source matter: Understanding inconsistencies in test collection-based evaluation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1843–1846. ACM, 2014.

[26] E. Kanoulas and J. A. Aslam. Empirical justification of the gain and discount function for ndcg. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 611–620, New York, NY, USA, 2009. ACM.

[27] E. Kanoulas, B. Carterette, P. D. Clough, and M. Sanderson. Evaluating multi-query sessions. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1053–1062, New York, NY, USA, 2011. ACM.

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

[29] E. Maddalena and S. Mizzaro. The Axiometrics Project. In R. Basili, F. Crestani, and M. Pennacchiotti, editors, *Proceedings of the 5th Italian Information Retrieval Workshop, Roma, Italy, January 20-21, 2014.*, volume 1127 of *CEUR Workshop Proceedings*, pages 11–15. CEUR-WS.org, 2014.

[30] M. Meila. Comparing clusterings. In *Proc. of COLT 03*, 2003.

[31] S. Mizzaro. A new measure of retrieval effectiveness (Or: What's wrong with precision and recall). In T. Ojala, editor, *International Workshop on*

Information Retrieval (IR'2001), pages 43–52, Oulu, Finland, Sept. 2001. Infotech Oulu. ISBN: 951-42-6489-4.

[32] S. Mizzaro. The Good, the Bad, the Difficult, and the Easy: Something Wrong with Information Retrieval Evaluation? In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, editors, *Advances in Information Retrieval, 30th European Conference on IR Research, ECIR 2008*, volume 4956 of *Lecture Notes in Computer Science*, pages 642–646, Glasgow, GB, 2008. Springer.

[33] S. Mizzaro and S. Robertson. HITS hits TREC: exploring IR evaluation results with network analysis. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, editors, *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 479–486, Amsterdam, Olanda, 2007. ACM.

[34] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, Dec. 2008.

[35] A. Omar and S. Mizzaro. Using crowdsourcing for TREC relevance assessment. *Information Processing and Management*, 48:1053–1066, 2012.

[36] S. E. Robertson and E. Kanoulas. On per-topic variance in ir evaluation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 891–900, New York, NY, USA, 2012. ACM.

[37] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 525–532, New York, NY, USA, 2006. ACM.

[38] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 555–562, New York, NY, USA, 2010. ACM.

[39] F. Scholer, E. Maddalena, A. Turpin, and S. Mizzaro. Magnitudes of relevance: Relevance judgements, magnitude estimation, and crowdsourcing. In *Proceedings of the Sixth International Workshop on Evaluating Information Access (EVIA 2014)*, pages 9–16, Tokyo, Japan, Dec. 9 2014. National Institute of Informatics. ISBN: 978-4-86049-066-9.

[40] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 95–104, New York, NY, USA, 2012. ACM.

[41] C. J. van Rijsbergen. Foundation of evaluation. *Journal of Documentation*, 30(4):365–373, 1974.

[42] E. Yilmaz, E. Kanoulas, and N. Craswell. Effect of intent descriptions on retrieval evaluation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 599–608, New York, NY, USA, 2014. ACM.