Measures for Benchmarking Semantic Web Service Matchmaking Correctness

Ulrich Küster and Birgitta König-Ries
Friedrich Schiller University Jena
http://fusion.cs.uni-jena.de/ukuester
Ulrich.Kuester@uni-jena.de

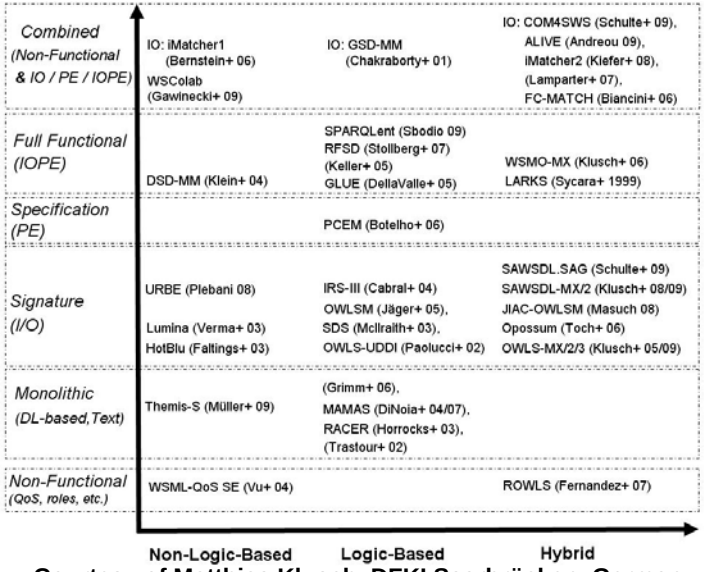
Agenda

1. Introduction and Motivation
2. Measures in Theory
3. Measures in Practice
4. Conclusions

Motivation: Service Discovery and Matchmaking

- Plethora of approaches
- Different techniques, formalisms, level of semantics
- Use cases from „manual repository search“ to „autonomous dynamic service binding“

How do they compare?
Where are their strengths?
How to improve them?



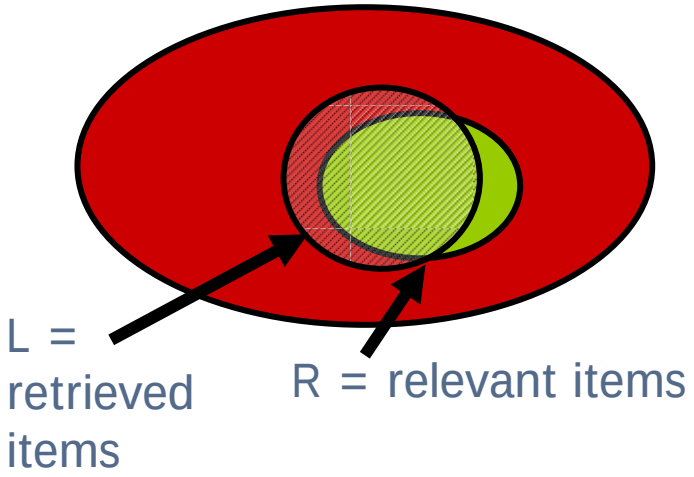
Courtesy of Matthias Klusch, DFKI Saarbrücken, Germany

Standard Evaluation Approach

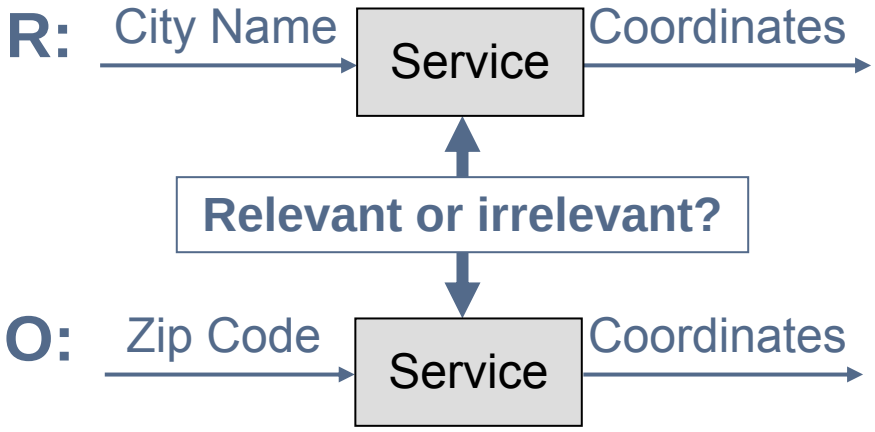
- Laboratory approach from IR (TREC, Cranfield)
  - Task: Order services ordered by relevance to query
  - Data: Test collection (services, queries, relevance sets)
  - Evaluation measure assigns a performance score to a sequence of services with respect to a query
  - Traditionally: Recall and Precision (and derived measures)

$$Recall = \frac{|L \bigcap R|}{|R|}$$

$$Precision = \frac{|L \bigcap R|}{|L|}$$



Binary versus Graded Relevance



Semantic match levels e.g.,:

Exact, Plugin, Subsumes, SubsumedBy, Intersection, None

- Binary relevance often too course grained (Tsetsos+ ECOWS06)
- Graded relevance employs multiple (or continuous) levels of relevance
- Requires adapted measures
- Some generalizations of recall and precision are available from IR

Measures from IR

- Binary
  - Recall, Precision
  - **AveP (average precision over relevant items)**
- Graded measures
  - Based on cumulated gain
    - **AWP (averaged weighted precision over relevant items)**
    - AWDP
    - DCG_l
    - GenAveP
    - Q-measure
  - Rank-correlation measures (Kendall's τ)

Binary Average Precision

- Precision averaged over (binary) relevant items
- E.g., three types of items, gold and silver relevant, bronze irrelevant

Relevant:	**yes**	**yes**	no	**yes**	no
$Precision_I$:	**1**	**1**	2/3	**3/4**	3/5

AveP: (1 + 1 + ¾) / 3 = 0,92

Graded Cumulated Gain and AWP

- Intuitively: cumulated gain of a user scanning items
- E.g., gold: gain 5, silver: gain 2, bronze: gain 0

Gain:	2	5	0	2	0	
CG:	**2**	**7**	7	**9**	9	(like recall)

ICG:	**5**	**7**	9	**9**	9	(CG of ideal ranking)
NCG:	**0,4**	**1**	0,78	**1**	1	(CG/ICG - like precision)

AWP: (0,4 + 1 + 1) / 3 = 0,8

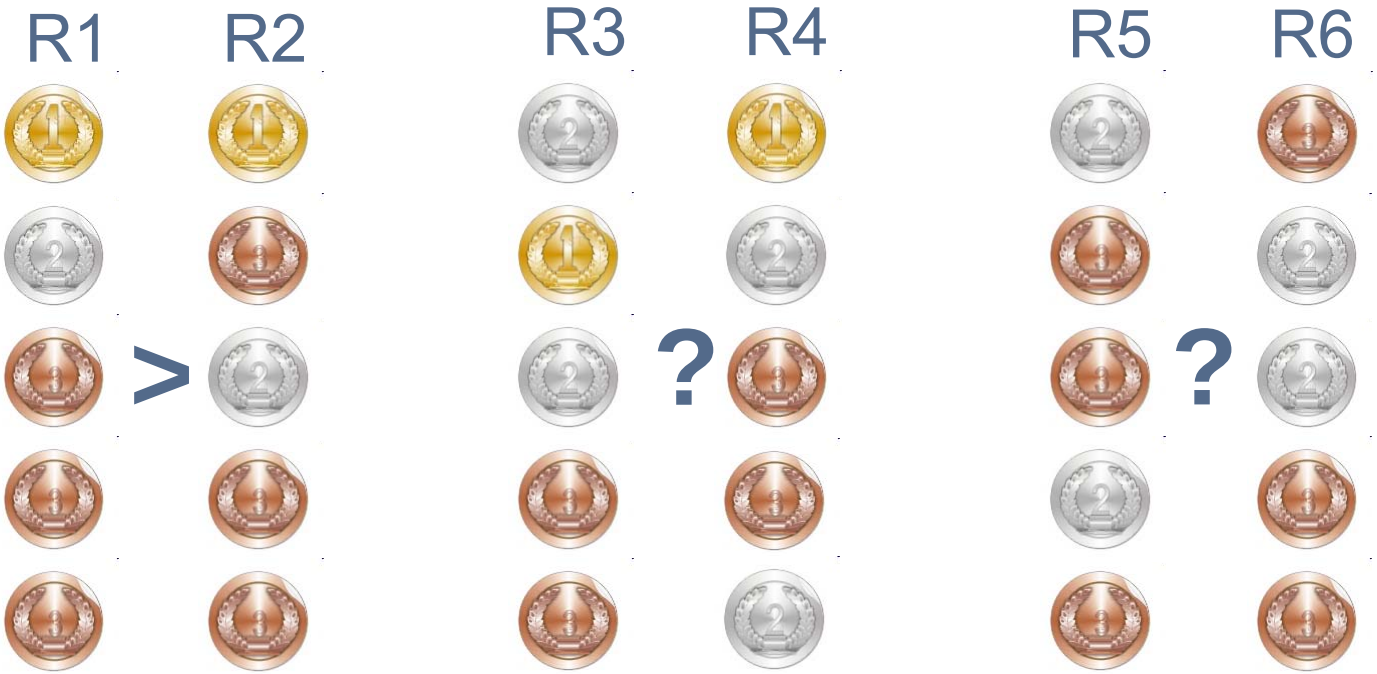
Research Questions and Goals

- Research Questions
  - What makes a good measure?
  - Which measure to use?
  - How to interpret results reliably?

- Research Goals
  - ➔**Quality criteria for measures**
  - ➔Discussion of existing measures in theory
  - ➔Investigation of measure behavior in practice

Measure Correctness

Three relevance levels (gold, silver, bronze)



➔Formalized intuitive notion of correctness (see paper)

➔Correctness alone does not suffice

Desireable Measure Characteristics

- Measures should be
  - Correct
  - Customizable wrt.
    - Emphasis of top rank over bottom rank performance
    - Preference for highly over marginally relevant items
  - Allow averaging results over multiple queries (ideal ranking scores 1)

Research Questions and Goals

- Research Questions
  - What makes a good measure?
  - Which measure to use?
  - How to interpret results reliably?

- Research Goals
  - ➔Objective quality criteria for measures
  - ➔**Discussion of existing measures in theory**
  - ➔Investigation of measure behavior in practice

Flaws of Existing Measures

- **AWP**
  - **Late retrieval defect**
  - **Order versus rank defect**
- AWDP (AWP + discounting)
  - Order versus rank defect
- DCG_I
  - Often inappropriate discount function used
- GenAveP
  - Order versus rank defect
- Q-measure
  - Order versus rank defect
- Kendall's τ
  - Does not emphasize top rank over bottom rank performance
  - Does not allow to configure degree of preference for highly relevant items

Average Weighted Precision (AWP)

- Late retrieval defect (known in the literature)

ICG: 5 7 9 9 9

NCG: **0,4** **1** **1** 1 1

AWP: (0,4 + 1 + 1) / 3 = 0,8

Late retrieval not penalized

NCG: **0,4** **1** 0,78 0,78 **1**

AWP: (0,4 + 1 + 1) / 3 = 0,8

Average Weighted Precision (AWP)

- Order versus rank defect (new)

ICG: 5 7 9 9 9

NCG: **0,4** **0,57** **1** 1 1

AWP: (0,4 + 0,57 + 1) / 3 = **0,66**

Order rewarded rather than rank

NCG: 0 0 **0,56** **0,78** **1**

AWP: (0,56 + 0,78 + 1) / 3 = **0,78**

Graded Measures from IR

- AWP (late retrieval + order versus rank defect)
- AWDP (order versus rank defect)
- DCG_I (inappropriate discount function)
- GenAveP (order versus rank defect)

Fixed versions contributed

- Q-measure (order versus rank defect)
- Kendall's τ (inflexible - does not emphasize top rank over bottom rank performance, does not support customizing degree of preference for highly relevant items)

Cannot be easily fixed

Research Questions and Goals

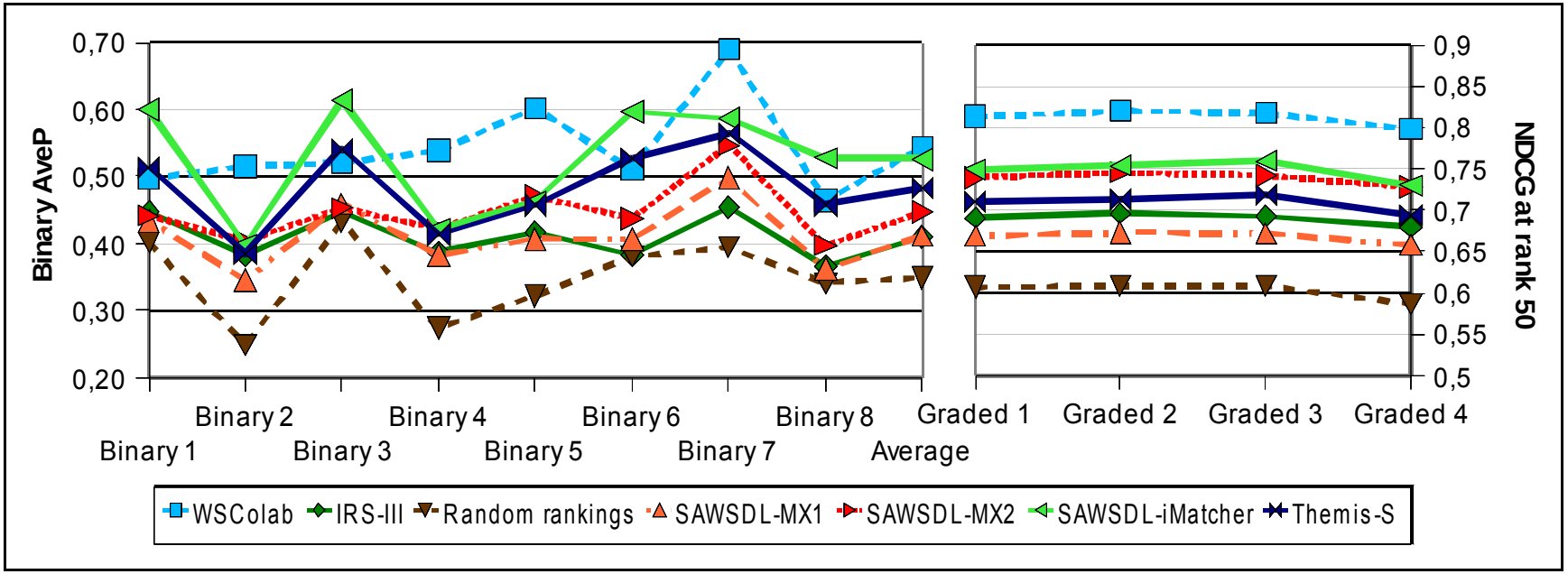
- Research Questions
  - What makes a good measure?
  - Which measure to use?
  - How to interpret results reliably?

- Research Goals
  - ➔Objective quality criteria for measures
  - ➔Discussion of existing measures in theory
  - ➔**Investigation of measure behavior in practice**

Experimental Setup

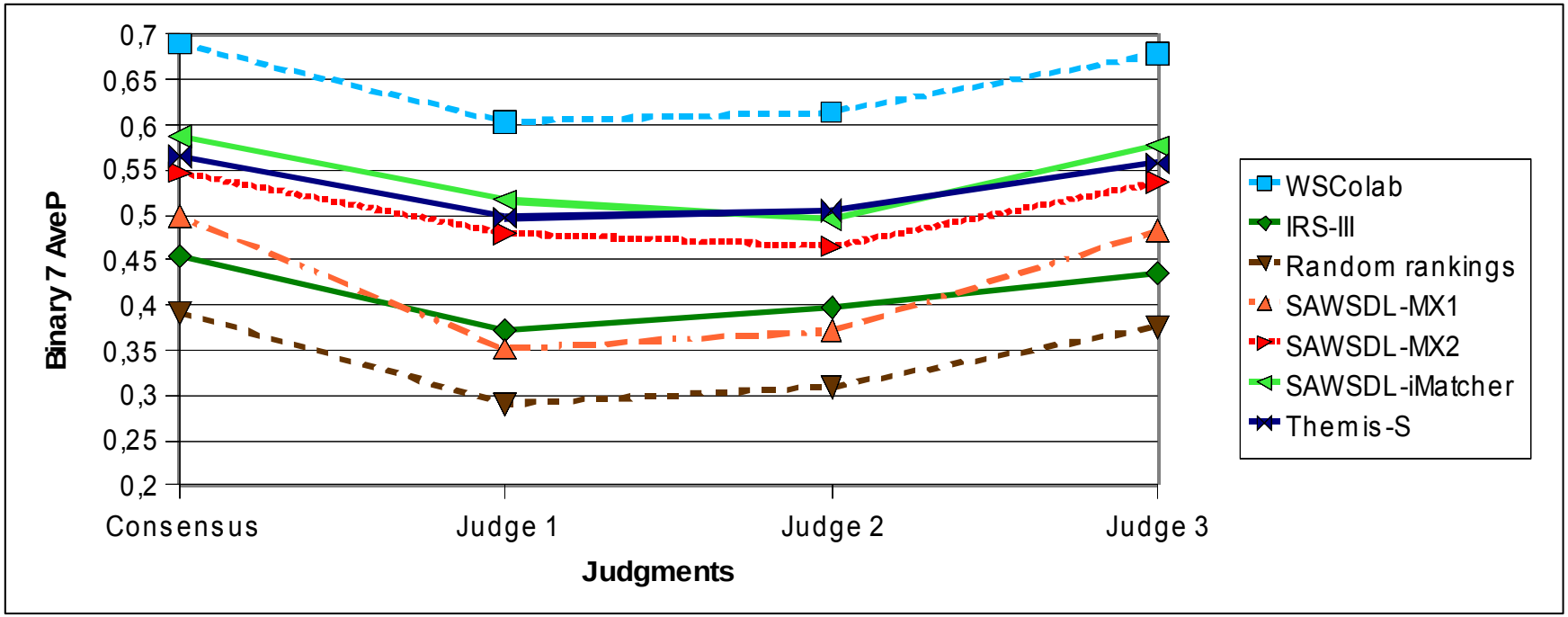
- Jena Geography Dataset
  http://fusion.cs.uni-jena.de/professur/jgd/
  - 200 real service operations (Geography domain)
  - Described with NL and WSDL
  - 10 sample requests
  - relevance judgments by 3 human experts (multi-dimensional, graded relevance)
- Organized as S3 Contest 2009 Cross Evaluation Track
  - 6 participants (NL, semantic tags, 3 * SAWSDL, WSML/OCML)
  - Participants provided semantic annotations (only for 50 services)

I: Influence of Relevance



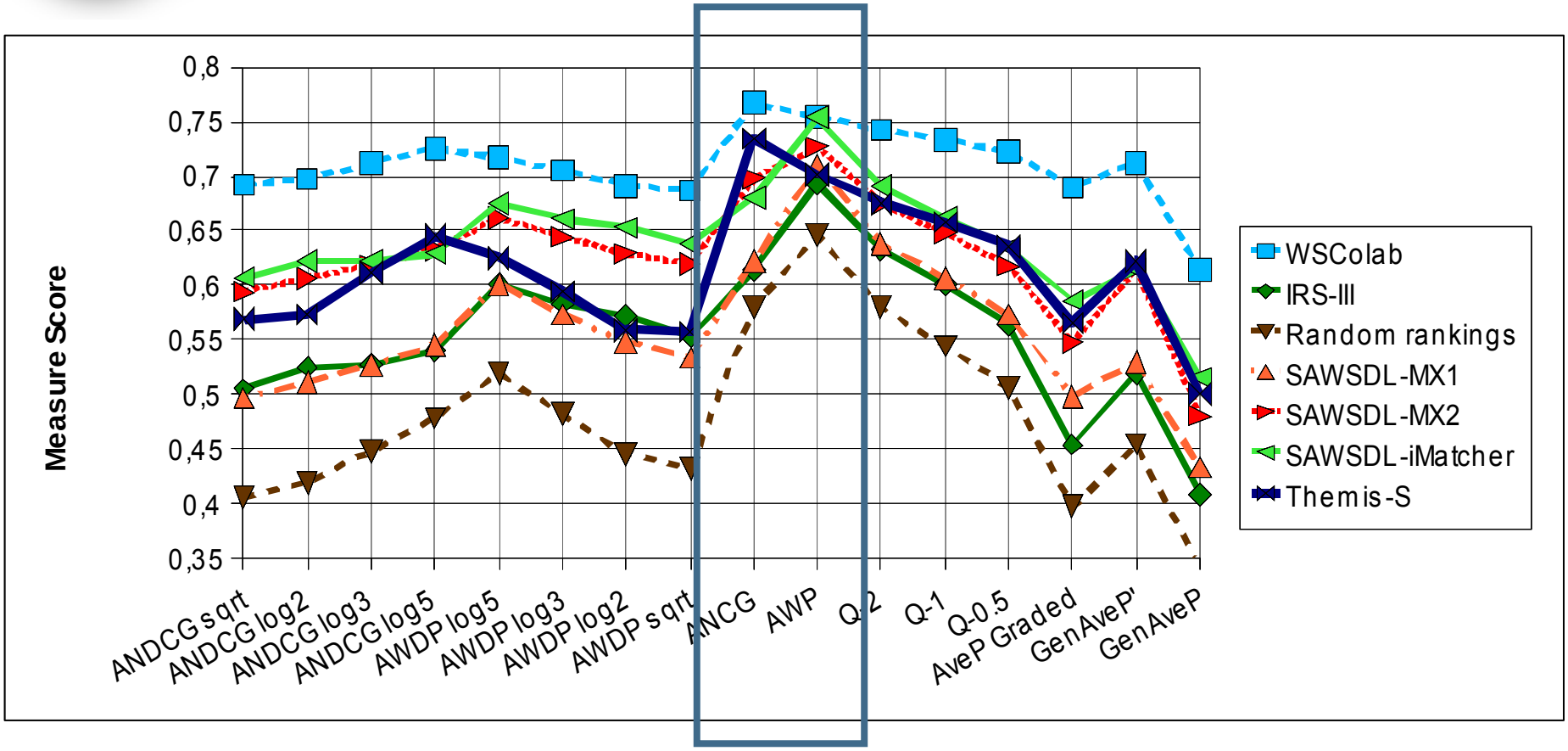
- Binary AveP sensitive to changes of relevance
- All graded measures stable against changes in relevance

II: Influence of Relevance Judge



- Reference judgments are subjective and inconsistent!
- All measures (binary + graded) relatively stable against subjective judgments

III: Influence of Measure



- Choice of evaluation performance measure matters!
- Themis-S illustrates order-versus-rank defect of AW(D)P

Summary

- Contributions
  - Investigation of properties of retrieval performance measures
  - Improvements w.r.t. identified defects suggested
  - Experimental investigation of measure behavior in practice
- Conclusions
  - AveP highly sensitive to definition of relevance
  - Subjective judgments have limited influence
  - Choice of evaluation measure matters
  - Some existing graded measures flawed; $NDCG_I$ and ANCG/ANDCG recommened

Thank You!

Ulrich Küster
Ulrich.Kuester@uni-jena.de
http://fusion.cs.uni-jena.de/ukuester