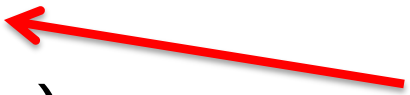
A Comparative Analysis of Cascade Measures for Novelty and Diversity

Charles Clarke, University of Waterloo
Nick Craswell, Microsoft
Ian Soboroff, NIST
Azin Ashkan, University of Waterloo

Background

- Measuring Web search effectiveness:
  - Editorial judgments (e.g., nDCG)
  - Implicit feedback (e.g., click curves)
  - Focused user studies



Limitations:

- we consider editorial measures only
- we assume ranked lists

Novelty and Diversity

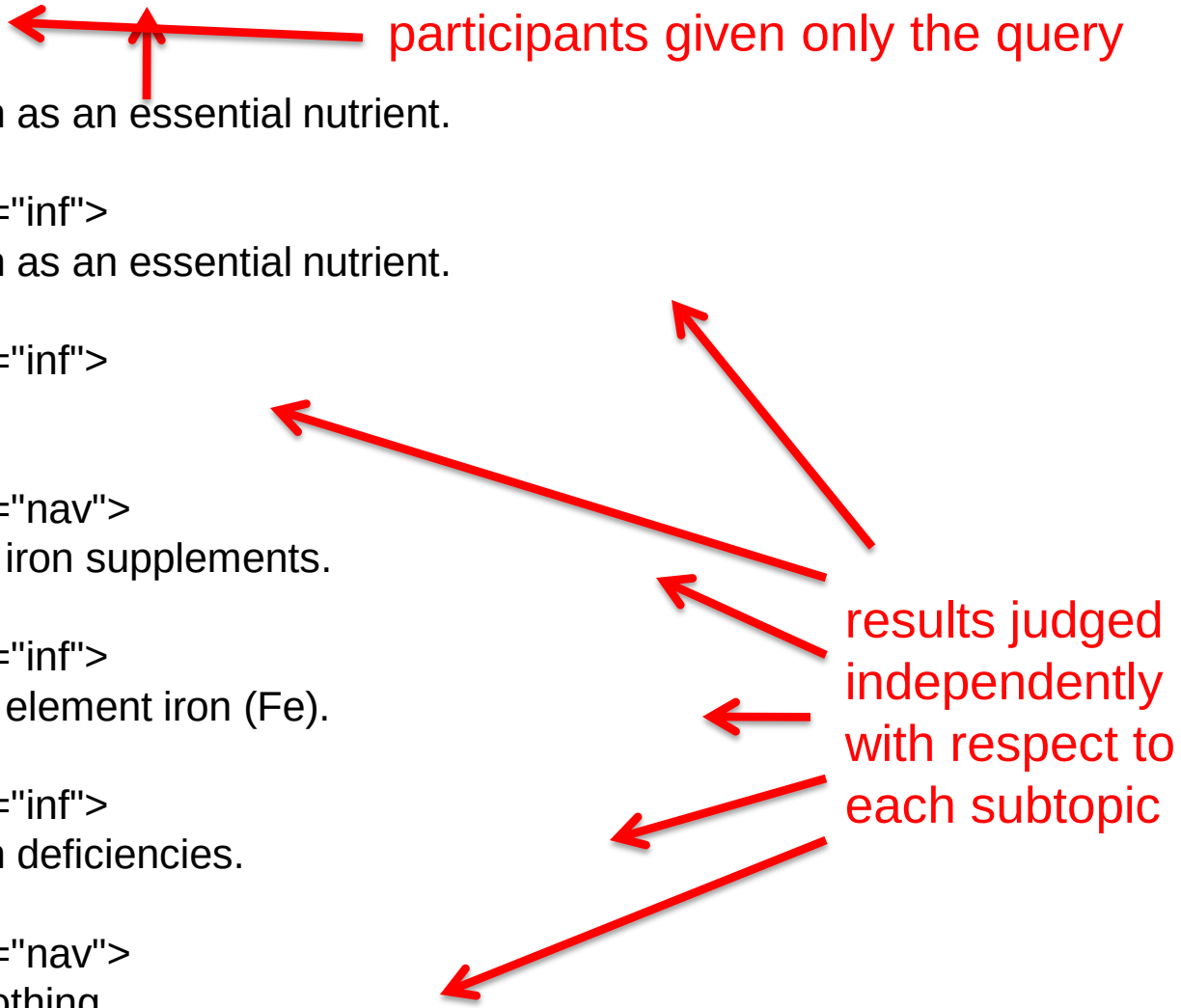
- Diversity: reflects the variety of user needs underlying a query
- Novelty: reflects the variety of information appearing in a search engine result page

Traditional measures ignore redundancy and inter-document relevance. Our goal is to validate tractable measures that address these issues.

TREC Web Track (2009-2011)

- Provides: A framework for exploring novelty and diversity in Web search (1TB ClueWeb09 document collection, 50 topics/queries).
- Goal: Return a ranked list of documents in order of decreasing probability of relevance, where relevance is considered in the context of higher-ranked documents.
- Evaluation: Assessed through binary judgments made with respect to explicit *sub-topics*.

```
<topic number="55" type="ambiguous">
  <query>iron</query>
  <description>
    Find information about iron as an essential nutrient.
  </description>
  <subtopic number="1" type="inf">
    Find information about iron as an essential nutrient.
  </subtopic>
  <subtopic number="2" type="inf">
    What foods contain iron?
  </subtopic>
  <subtopic number="3" type="nav">
    Find sites where I can buy iron supplements.
  </subtopic>
  <subtopic number="4" type="inf">
    Find information about the element iron (Fe).
  </subtopic>
  <subtopic number="5" type="inf">
    Find information about iron deficiencies.
  </subtopic>
  <subtopic number="6" type="nav">
    Find dealers in irons for clothing.
  </subtopic>
</topic>
```



```
<topic number="73" type="faceted">
  <query>neil young</query>
  <description>
    Find music, tour dates, and information about the
    musician Neil Young.
  </description>
  <subtopic number="1" type="nav">
    Find albums by Neil Young to buy.
  </subtopic>
  <subtopic number="2" type="inf">
    Find biographical information about Neil Young.
  </subtopic>
  <subtopic number="3" type="nav">
    Find lyrics or sheet music for Neil Young's songs.
  </subtopic>
  <subtopic number="4" type="nav">
    Find a list of Neil Young tour dates.
  </subtopic>
</topic>
```

“iron”

RELATED SEARCHES
Iron **Man**
Iron **Maiden**
Iron **Search**
Iron **Deficiency**
Iron **Rich Foods**
Steam Irons
Iron **Supplements**
Iron **Mountain**

SEARCH HISTORY
Search more to see your history

See all
Clear all · Turn off

ALL RESULTS 1-11 of 234,000,000 results · Advanced

Sponsored sites

Driveway Gates · www.madeiniron.ca/
MADE IN **IRON** specializes in custom ornamental ironwork

Images of **iron**

Iron - Wikipedia, the free encyclopedia
Characteristics · Chemistry and compounds · History · Industrial production
Iron is a chemical element with the symbol Fe and atomic number 26. It is a metal in the first transition series. It is the most common element in the whole planet Earth ...
en.wikipedia.org/wiki/**Iron** · Cached page

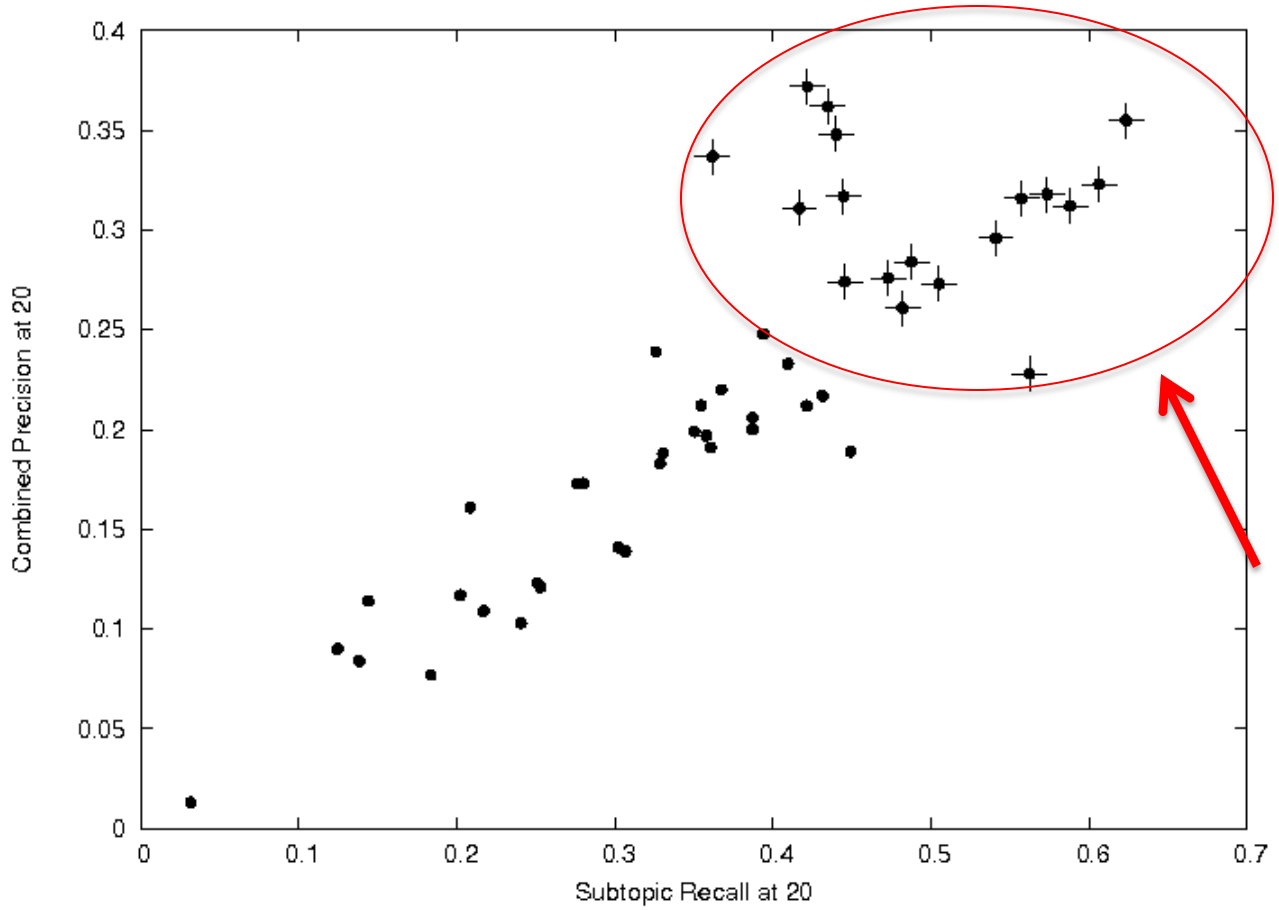
Dietary Supplement Fact Sheet: **Iron**
Iron: What is it? **Iron**, one of the most abundant metals on Earth, is essential to most life forms and to normal human physiology. **Iron** is an integral part of many proteins and ...
ods.od.nih.gov/factsheets/**iron** · Cached page

Irons | **Irons**.com
Welcome to **Irons**.com internet store - where you can find finest **Irons** to match your needs. Please browse our selection of electric **irons** and don't miss our bargains! **Irons** ...
www.**irons**.com · Cached page

Measuring Novelty

- Subtopic recall:
  - simple measure often used in early work
  - e.g, Zhai et al. (2003), Zhang et al. (2005)
- Intent aware versions of traditional measures:
  - Agrawal et al. (2009): nDCG-IA, MAP-IA
- Intent aware cascade measures:
  - Clarke et al. (2008): α-nDCG
  - Chapelle et al (2009): ERR
  - Clarke et al. (2009): NRBP
- Plus: Sakai et al. (2010), Rafiei et al. (2010), etc.

Subtopic Recall vs. Precision (TREC 2009)



Top-performing groups at TREC 2009 exhibit a mixture of outcomes, with some groups excelling at novelty and others excelling at traditional relevance.

Intent Aware Measures

$$\sum_{i=1}^{M} p_i \, \mathcal{S}_i$$

measures applied independently to each of *M* subtopics

(e.g., nDCG, Map, etc.)

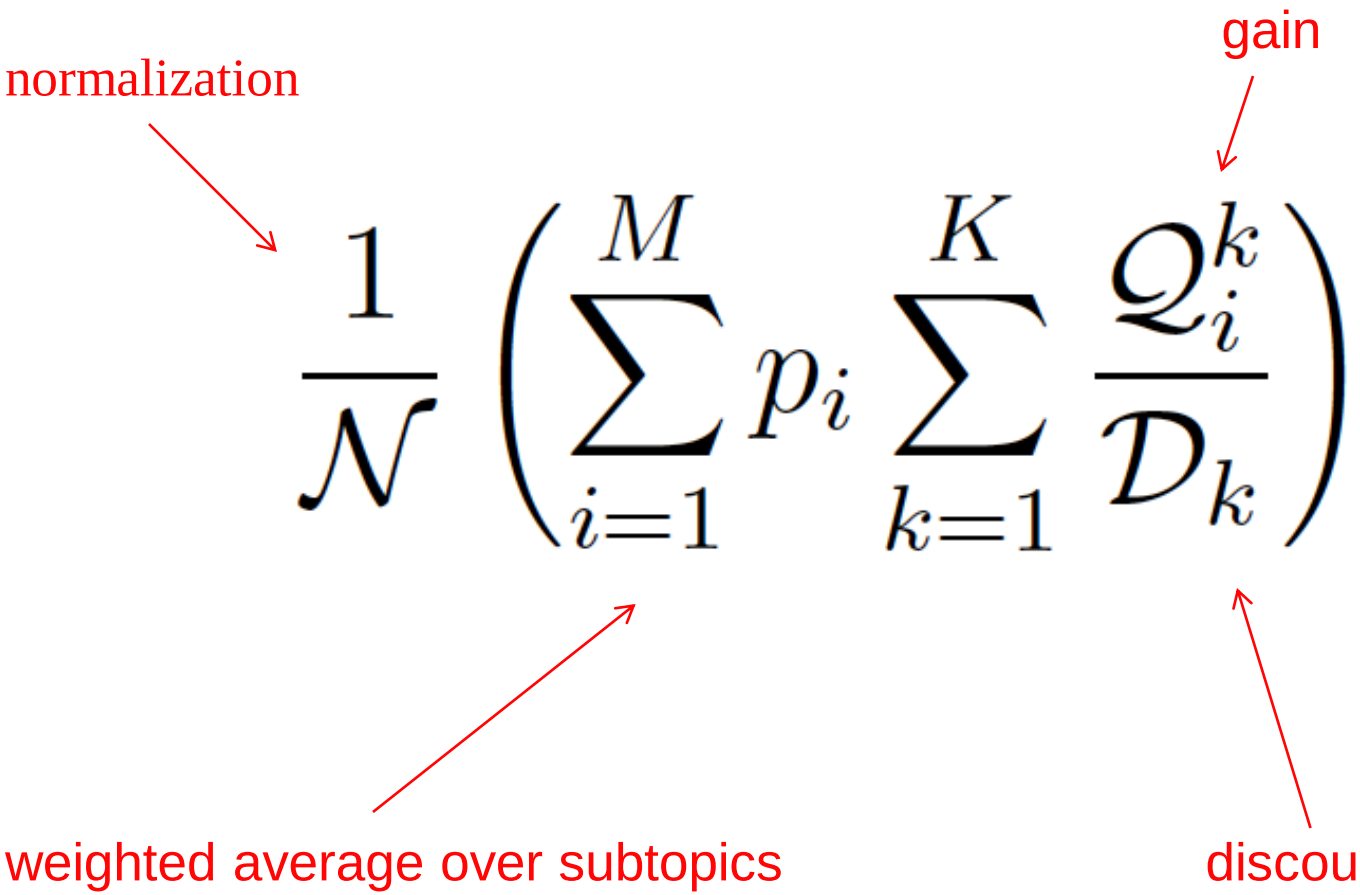
score for a topic is a weighted average over subtopics

(probabilities need not sum to 1)

Intent Aware Measures

- Intent aware versions of traditional measures (MAP, nDCG, etc.) do not penalize redundancy.
- To achieve best performance under these traditional measures, the search engine should focus on the most popular topic.
- In contrast, cascade measures explicitly penalize redundancy.

Intent Aware Cascade Measures



Gain
Penalizes Redundancy

$$\mathcal{Q}_i^k = q_i^k \prod_{j=1}^{k-1} (1 - q_i^j)$$

simplified to

$$\mathcal{Q}_i^k = \alpha g_i^k (1 - \alpha)^{c_j^k}$$

α represents the user's tolerance for redundancy

Discount Penalizes Depth (*k*)

$$\mathcal{D}_k = \log(k+1)$$

α-nDCG (Clarke et al., 2008)

$$\mathcal{D}_k = k$$

ERR (Chapelle et al., 2009)

$$\mathcal{D}_k = (1/\beta)^{k-1}$$

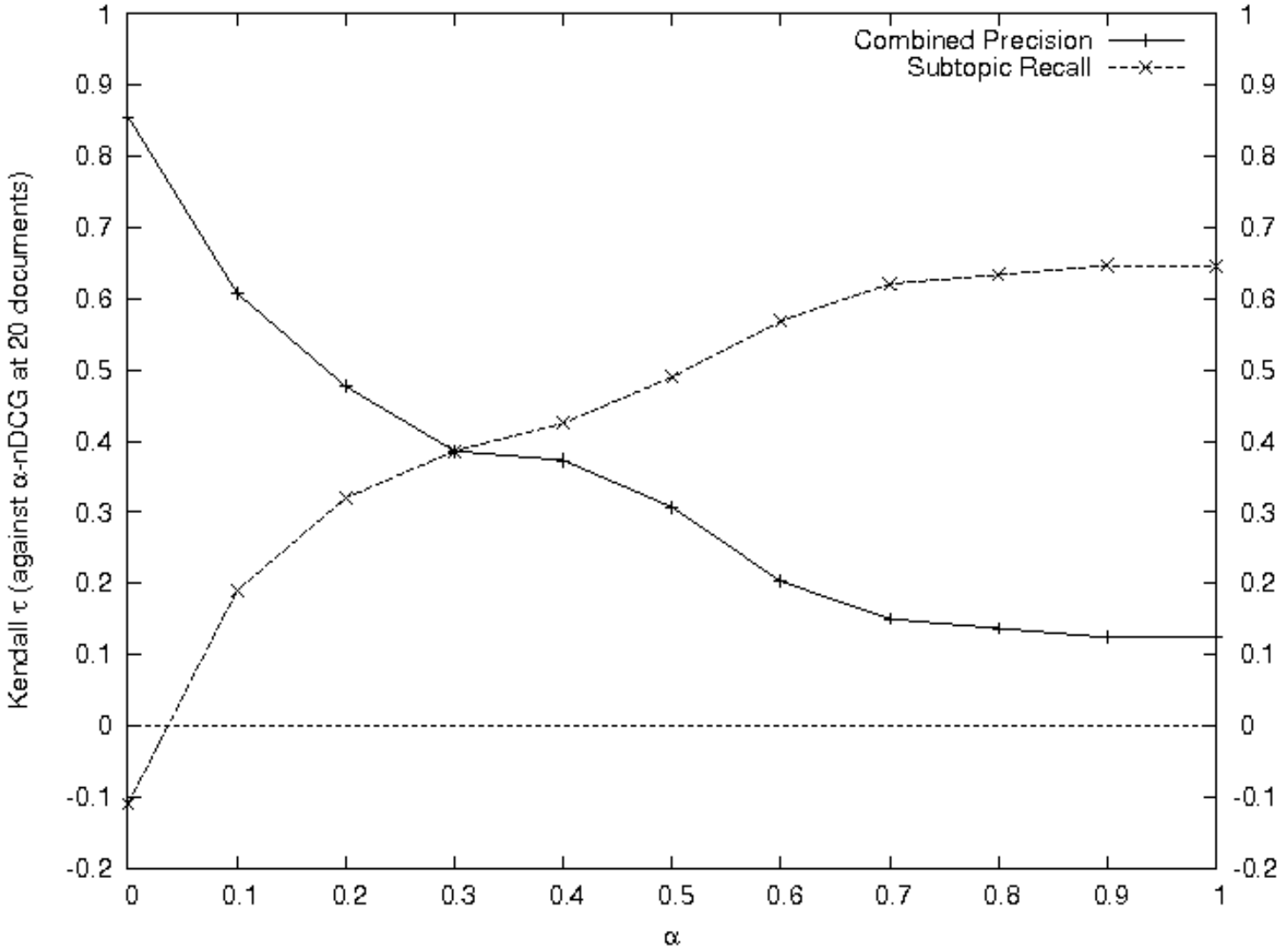
NRBP (Clarke et al., 2009)

Normalization Controls for Number of Subtopics

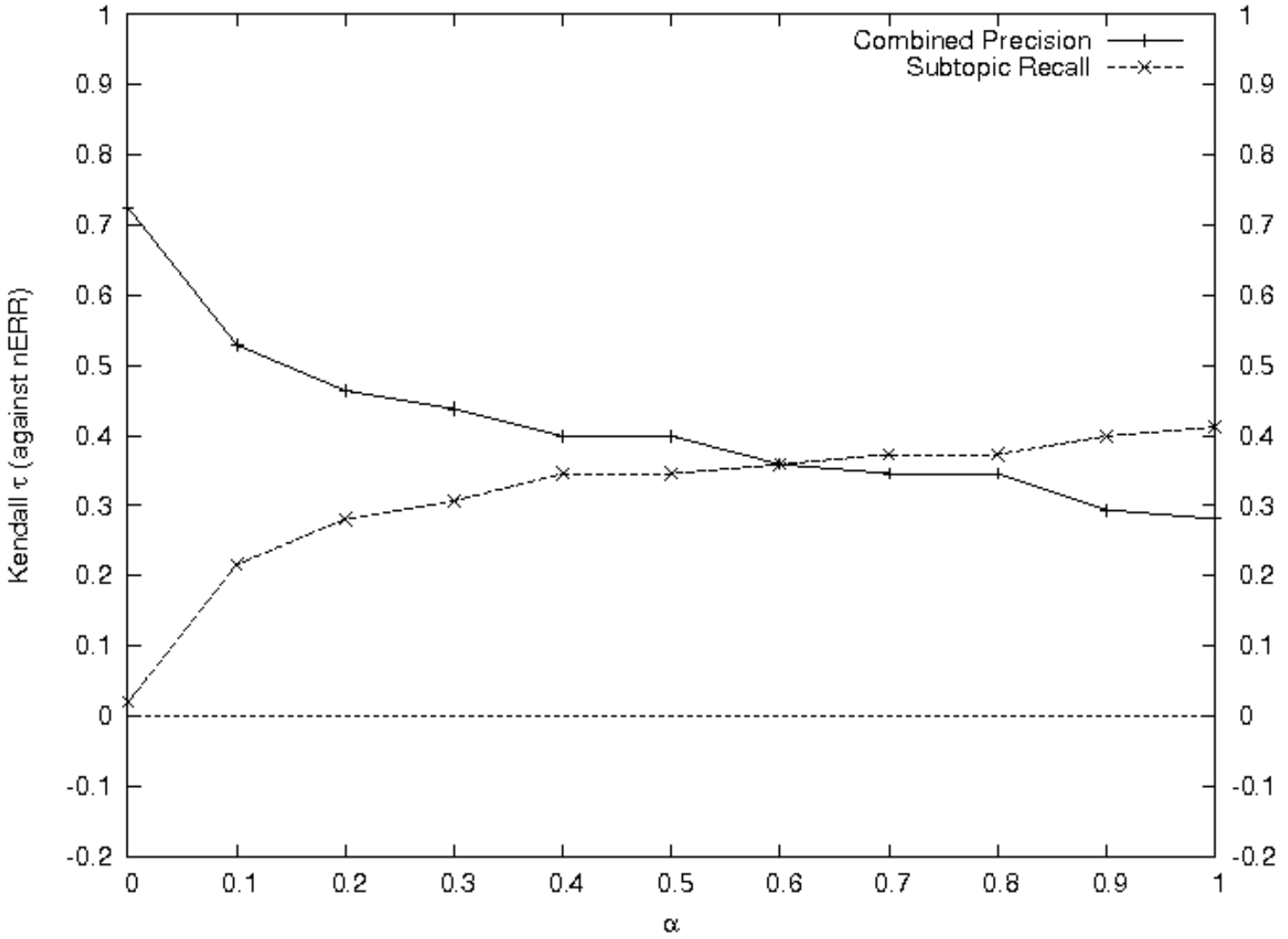
- Scores should be normalized into the range [0:1] for averaging.
- Value of highest possible score can grow with number of subtopics.
- Normalization either
  - collection independent
  - collection dependent

NP hard. See Carterette (2009) for discussion.

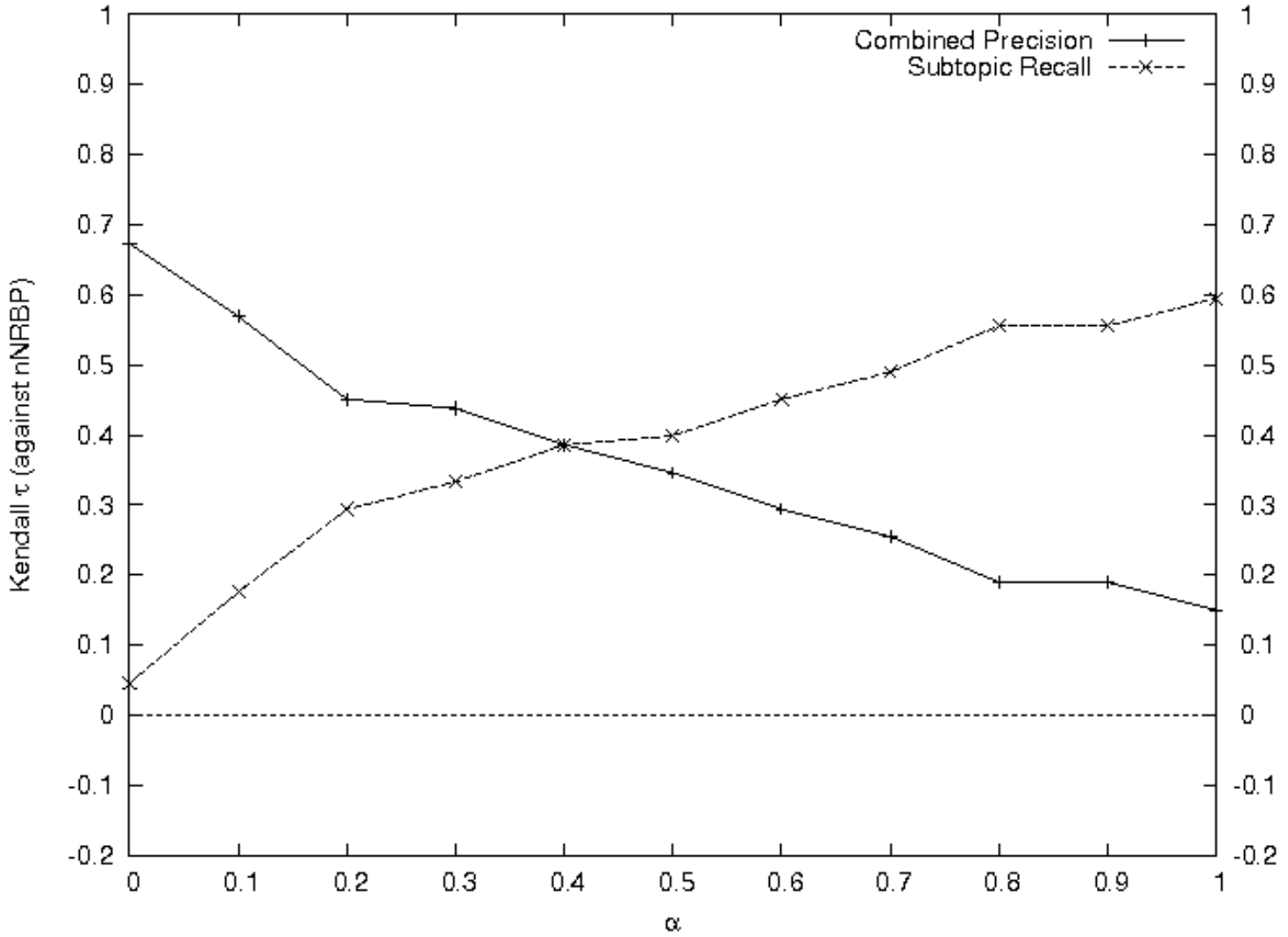
Impact of Varying α on α-nDCG



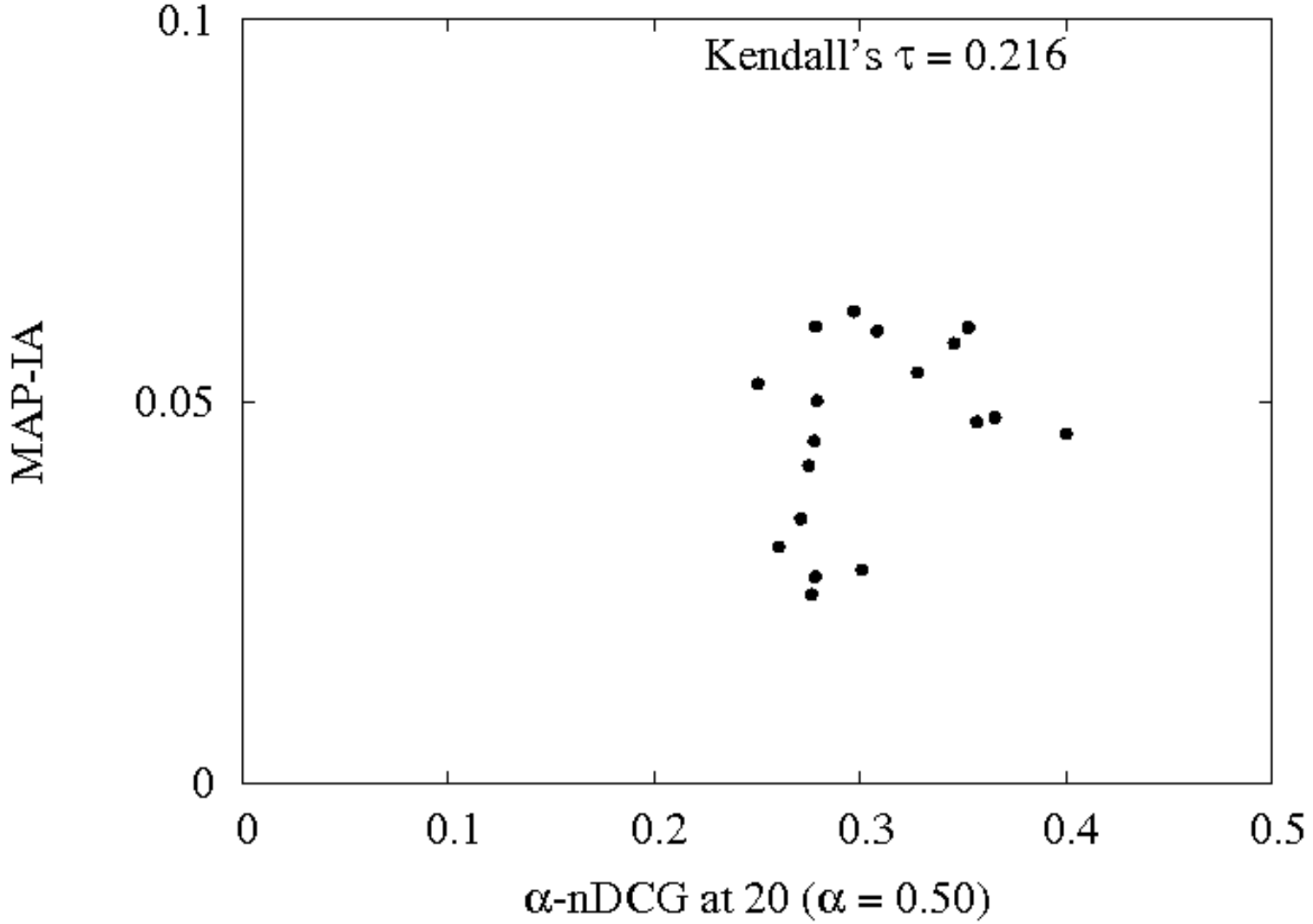
Impact of Varying α on nERR



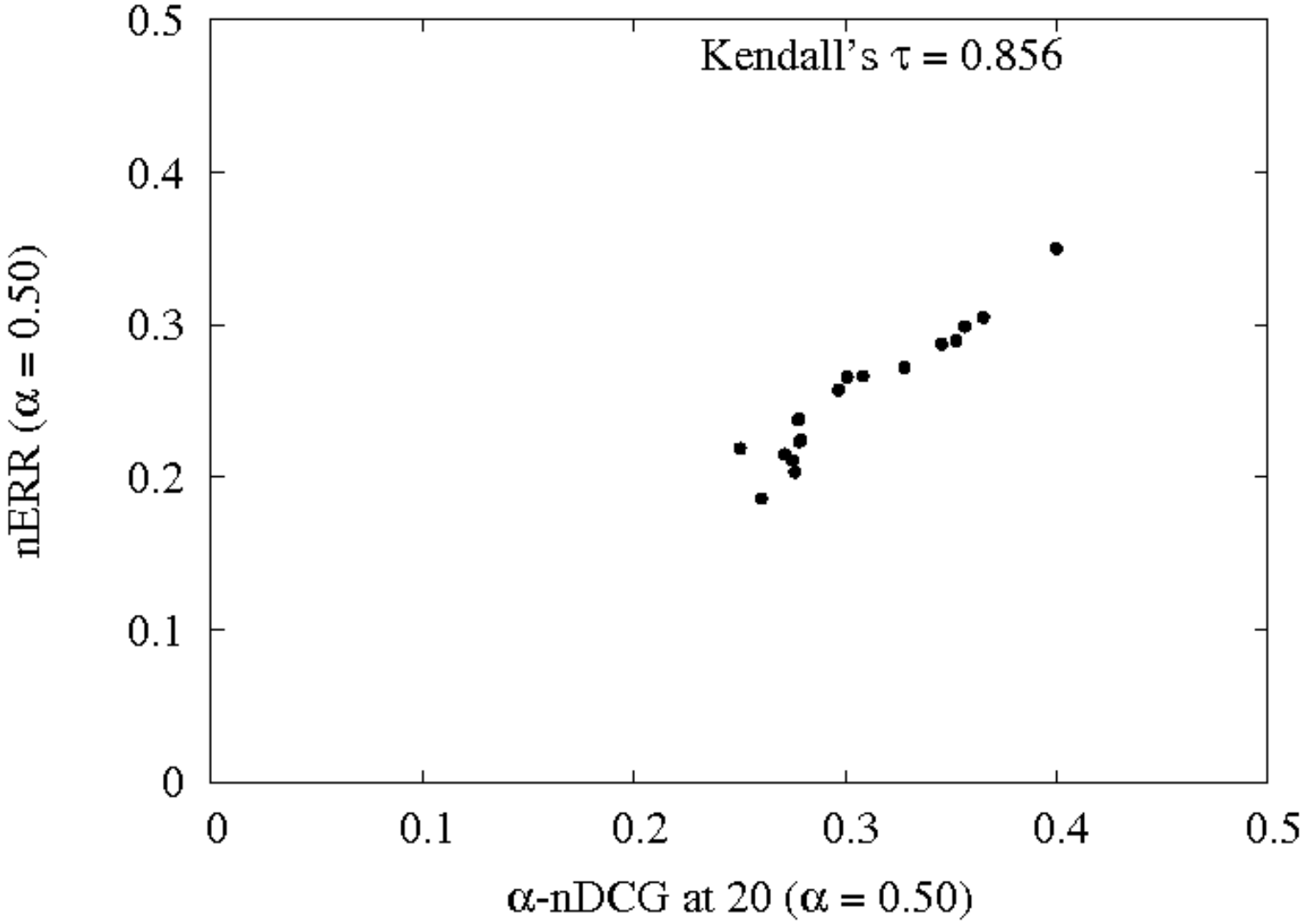
Impact of Varying α on nNRBP



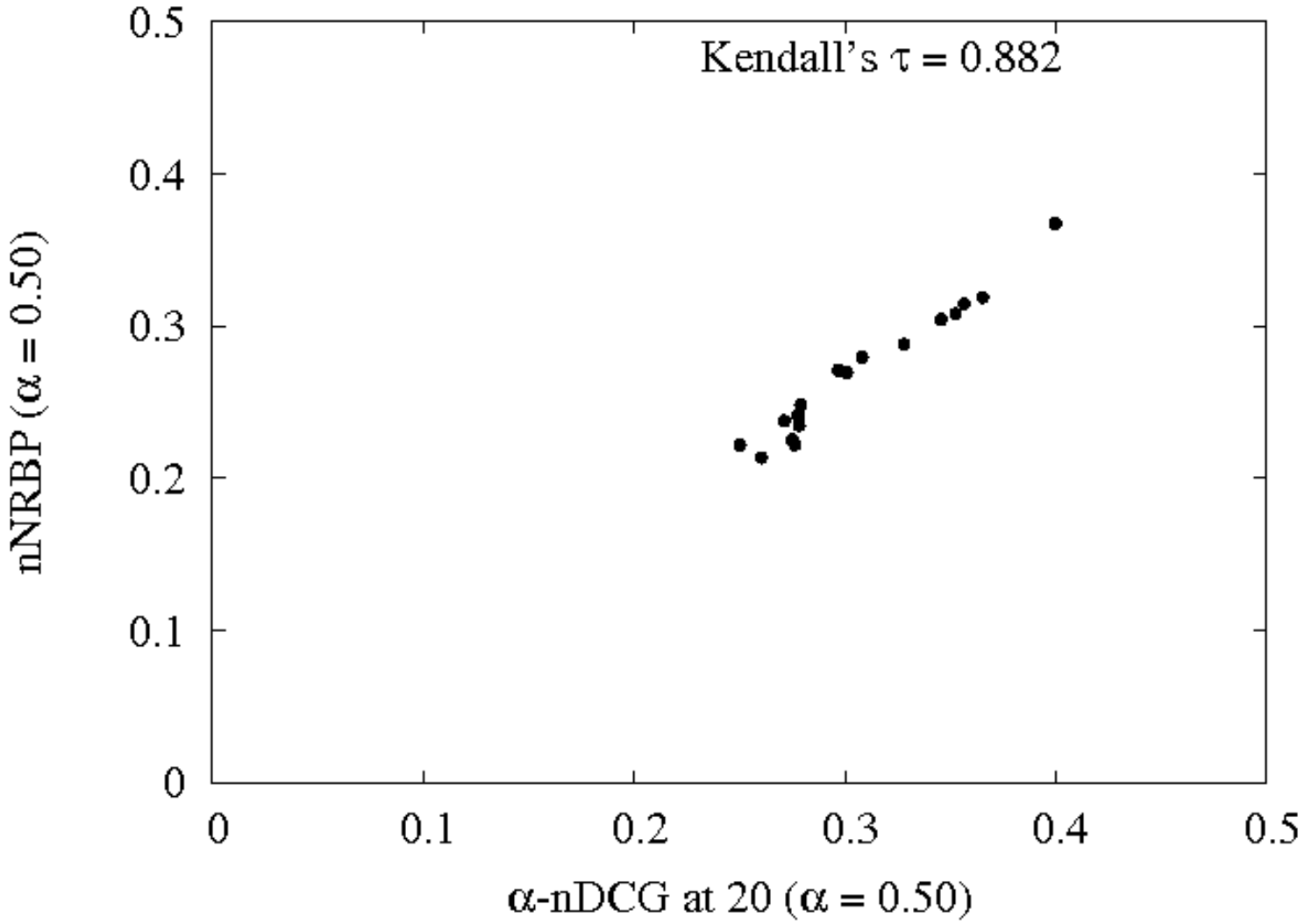
MAP-IA vs. α-nDCG



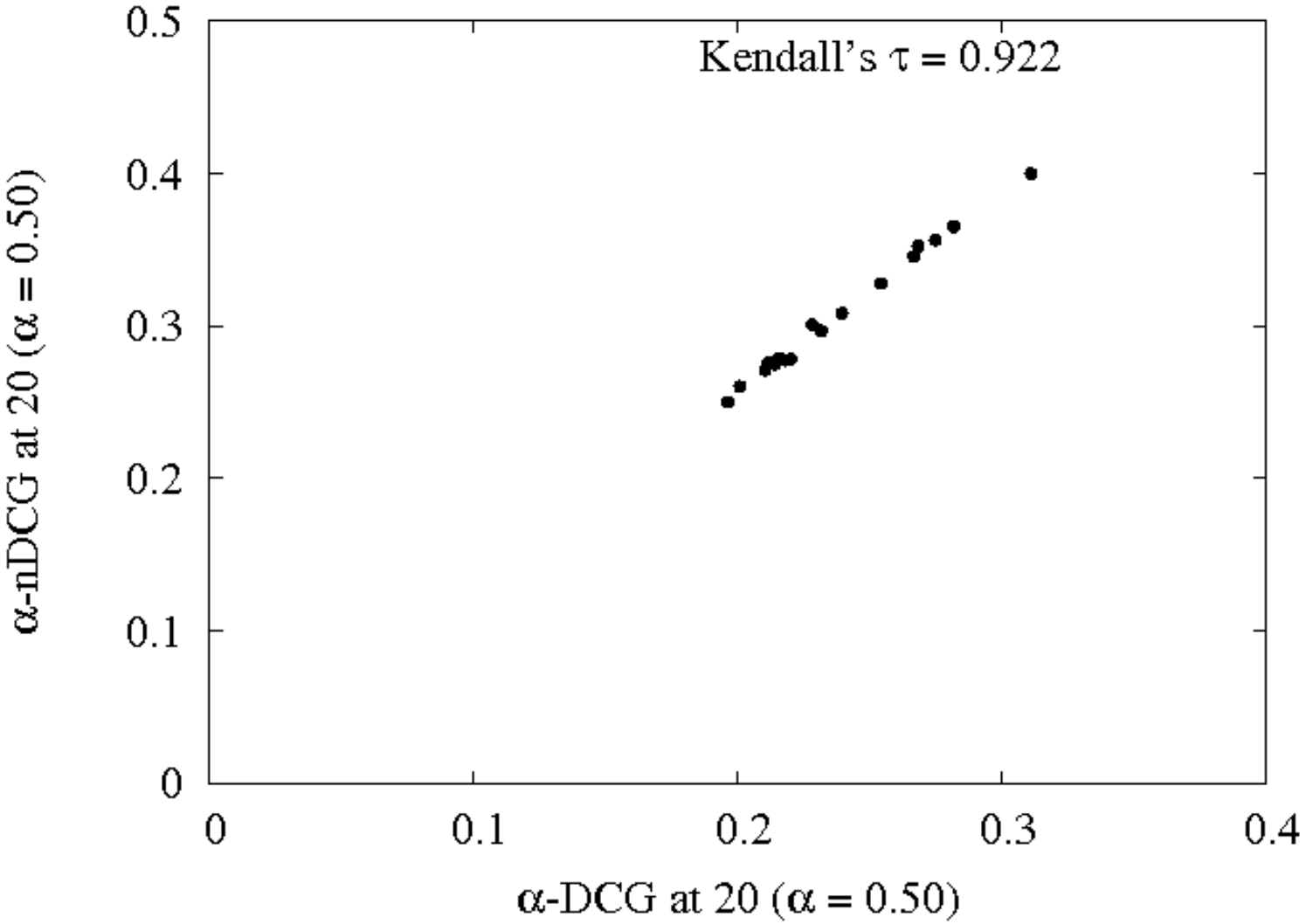
nERR vs. α-nDCG



nNRBP vs. α-nDCG



Impact of Normalization on α-(n)DCG



Discriminative Power

	bootstrap	paired t-test
Combined precision	20.9%	22.2%
Subtopic recall	44.4%	46.4%
α-DCG	30.7%	32.0%
α-nDCG	32.7%	34.0%
ERR	28.8%	29.4%
nERR	27.5%	28.8%
NRBP	28.7%	29.4%
nNRBP	30.1%	30.7%
MAP-IA	34.6%	37.9%
MAP	41.8%	49.0%

Figure 7: Discriminative power of measures under the two-tailed paired t-test and bootstrap tests with a significance level of 0.05.

Conclusions

- Unified framework for intent-aware measures, especially cascade measures.
- Intent aware measures must penalize redundancy.
- Measures provide a trade-off between traditional precision and subtopic recall.
- Collection dependent normalization may not be required.
- But discriminative power greatest with MAP.

Questions?