An In-depth Study on Diversity Evaluation: the Importance of Intrinsic Diversity

Hai-Tao Yu[a,*], Adam Jatowt[b], Roi Blanco[c], Hideo Joho[d], Joemon M. Jose[e]

[a] *Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Japan*
[b] *Department of Social Informatics, Graduate School of Informatics, Kyoto University, Kyoto, Japan*
[c] *IRLab, Computer Science Department, University of A Coruña, Spain*
[d] *Research Center for Knowledge Communities, Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Japan*
[e] *School of Computing Science, University of Glasgow, Glasgow, UK*

Abstract

Diversified document ranking has been recognized as an effective strategy to tackle ambiguous and/or underspecified queries. In this paper, we conduct an in-depth study on diversity evaluation that provides insights for assessing the performance of a diversified retrieval system. By casting the widely used diversity metrics (e.g., ERR-IA, α-nDCG and D#-nDCG) into a unified framework based on *marginal utility*, we analyze how these metrics capture *extrinsic diversity* and *intrinsic diversity*. Our analyses show that the prior metrics (ERR-IA, α-nDCG and D#-nDCG) are not able to precisely measure intrinsic diversity if we merely feed a set of subtopics into them in a traditional manner (i.e., without fine-grained relevance knowledge per subtopic). As the redundancy of relevant documents with respect to each specific information need (i.e., subtopic) can not be then detected and solved, the overall diversity evaluation may not be reliable. Furthermore, a series of experiments are conducted on a gold standard collection (English and Chinese) and a set of submitted runs, where the *intent-square metrics* that extend the diversity metrics through incorporating hierarchical subtopics are used as references. The experimental results show that the intent-square metrics disagree with the diversity metrics (ERR-IA and α-nDCG) being used in a traditional way on top-ranked runs, and that the average precision correlation scores between intent-square metrics and the prior diversity metrics (ERR-IA and α-nDCG) are fairly low. These results justify our analyses, and uncover the previously-unknown importance of intrinsic diversity to the overall diversity evaluation.

Keywords: Extrinsic diversity, Intrinsic diversity, Marginal utility

*Corresponding author
Email addresses: yuhaitao@slis.tsukuba.ac.jp (Hai-Tao Yu), adam@dl.kuis.kyoto-u.ac.jp (Adam Jatowt), rblanco@udc.es (Roi Blanco), hideo@slis.tsukuba.ac.jp (Hideo Joho), joemon.jose@glasgow.ac.uk (Joemon M. Jose)

1. Introduction

Web search engines play an increasingly dominant role in our daily information access. However, generating a high-quality result list in which users can find their desired information from the top few slots is far from being resolved. For example, many users often submit short queries with little or no context, so it is hard to accurately capture their information needs. Thus, merely providing results that satisfy only the most likely information need, will result in dissatisfaction of users with rare information needs. To cope with the ambiguous and/or underspecified queries, the technique of *diversified document ranking* has been proposed and has attracted significant attention. In this context, a diversified retrieval system faces a trade-off between relevance and diversity. For a detailed review readers can refer to the works [1, 2, 3].

Effective diversity evaluation provides meaningful insights for assessing a diversified retrieval system, e.g., how well it meets the information needs of users, how to choose among different retrieval models, features, etc. A recent effort for diversity evaluation is the *subtopic based strategy*. The possible information needs underlying a query are represented by a set of subtopics. The number of subtopics that a result list covers and how well a specific subtopic is satisfied provide then the criteria for measuring the *overall diversity* (or *expected diversity*). Note that when using the terms overall diversity and expected diversity, we refer to the diversity that we expect a diversified retrieval system to achieve (sometimes they are used interchangeably). Based on the work by Radlinski et al. [4], the overall diversity essentially can be resolved into *extrinsic diversity* and *intrinsic diversity*. Extrinsic diversity corresponds to the problem of addressing the uncertainty about the information needs underlying a query. Intrinsic diversity corresponds to the problem of avoiding excessive redundancy of documents retrieved for a particular information need. The distinction between them is that: Enhancing extrinsic diversity helps to improve the effectiveness of a diversified retrieval system by covering different information needs. The value of enhancing intrinsic diversity helps to increase the satisfaction degree w.r.t. each specific information need. Therefore, extrinsic diversity and intrinsic diversity play different roles in the overall evaluation, both of them are very important.

Take the formal query 0083 *harry potter* from NTCIR-11 (cf. Section 5.1) for example, whose subtopic hierarchy is shown in Fig. 1. The *first-level subtopics*, e.g., *harry potter book* and *harry potter film*, represent different information needs or intents. This clearly reflects the necessity of taking into account the extrinsic diversity, since it is hard to know which first-level subtopic the user is interested in. Given a specific information need, say *harry potter film*, the *second-level subtopics*, e.g., *harry potter film cast* and *harry potter film music* indicate that we should take these different aspects into consideration instead of providing redundant information w.r.t. a single aspect, so as to fully satisfy this information need. This apparently shows the importance of enhancing intrin-

Figure 1: Query 0083 *harry potter*

sic diversity. In the following, the term *subtopic* refers to a first-level subtopic when it is solely used, both of them represent a possible information need or an intent. As a shorthand, we write first-level subtopic and second-level subtopic as *fls* and *sls* respectively.

For effective diversity evaluation, various measures (e.g., $ERR\text{-}IA$, $\alpha\text{-}nDCG$ and $D\#\text{-}nDCG$) have been proposed. Their properties are further studied and compared by the works [5, 6, 7, 8, 9, 10, 11, 12]. However, in most of the cases, the traditional diversity evaluation is conducted on a set of subtopics of the first level, and no fine-grained knowledge per subtopic (e.g., *sls*) is taken into consideration. In other words, only first-level subtopics are fed into the diversity metrics. Due to the unawareness of fine-grained knowledge per subtopic, the diversity metrics can not precisely capture intrinsic diversity. In result, the importance of intrinsic diversity to the overall diversity evaluation has not been well understood.

In this paper, we focus on investigating the impact of intrinsic diversity on the overall diversity evaluation. The major contributions of this work are as follows: (a) By casting the widely used diversity metrics into a unified framework based on marginal utility, we investigate how these metrics capture intrinsic diversity. (b) Through an extended usage of the existing metrics, a family of novel metrics (i.e., *intent-square metrics*) are proposed. They allow for a meaningful comparison against the traditional diversity evaluation. (c) A series of experiments are conducted on a crowdsourced collection (English and Chinese) and a set of submitted runs. We compare the evaluation results obtained via

a traditional usage of $ERR\text{-}IA$ and $\alpha\text{-}nDCG$ against those using the intent-square metrics. The evaluation differences are illustrated in different levels of granularity (ranking across runs, ranking across queries, specific queries). The experimental results clearly show what are the effects of better measuring intrinsic diversity by using fine-grained knowledge per subtopic, and uncover the previously-unknown importance of intrinsic diversity for the expected diversity evaluation.

The rest of this paper is organized as follows. In Section 2, we briefly survey the current state of the related work on diversity evaluation. In Section 3, a general framework based on marginal utility is formalized. In Section 4, we briefly introduce the widely used diversity metrics, and show how they measure intrinsic diversity under a unified perspective. In Section 5, a series of experiments are conducted and discussed based on a crowdsourced collection. We conclude our work in Section 6.

2. Related Work

In this section, we give a brief survey of the typical metrics for diversity evaluation. The methods [13, 14, 15, 16, 17, 18, 19, 20, 21] on how to perform search result diversification are not detailed. We refer the reader to the work [1] for an overview of search result diversification.

For evaluating the novelty and diversity, Zhai *et al.* [5] presented the problem of subtopic retrieval in the context of TREC interactive track, and proposed several metrics such as subtopic recall and subtopic precision. Clarke *et al.* [6] introduced $\alpha\text{-}nDCG$ that captures redundancy through the repetition of relevant nuggets by decomposing the information needs of a query and the information inside a document into information nuggets. Agrawal *et al.* [7] explored the diversity evaluation by applying a traditional metric to each subtopic independently, and combined the results based on the importance or probability of subtopics. Based on per-intent graded relevance assessments, Sakai *et al.* [8] combined subtopic recall and normalised Discounted Cumulative Gain ($nDCG$) into a single evaluation measure.

However, the above-mentioned metrics sometimes fail to work as expected. For $\alpha\text{-}nDCG$, Leelanupab *et al.* [22] found that a common setting of $\alpha = 0.5$ for α-nDCG tends to excessively penalize systems that cover many subtopics while rewarding those that redundantly cover only few subtopics. Sakai [9] argued that $\alpha\text{-}nDCG$ works well for navigational subtopics, but it discourages retrieval of multiple relevant documents for each subtopic. For $ERR\text{-}IA$, the experimental results by Leelanupab *et al.* [23] showed that ERR-IA [7] tend to neglect subtopic coverage by attributing excessive importance to redundant relevant documents. This happens for 134 out of 148 queries from TREC 2009-11 Diversity tasks. Different from prior studies, our explanation is that these diversity metrics can not precisely measure intrinsic diversity when only using first-level subtopics. Recently, machine learning methods (e.g., [24, 20]) that directly optimize diversity metrics have been proposed for diversified document

ranking. Our work also helps to improve these by providing a better understanding of diversity evaluation.

Besides the above studies, Sakai [9] argued that subtopics should be differentiated when performing diversity evaluation, into, e.g, informational subtopic and navigational subtopic. Chen et al. [25] used a decay function that incorporates the taxonomy information w.r.t. a subtopic to compute the gain value at each rank position. Wang et al. [26] explored how to take into account the intent hierarchy of a query when performing diversity evaluation. The proposed metrics build upon a particular subtopic hierarchy (i.e., extended intent hierarchy), for which five specific properties have to be satisfied. The user study by Xu and Yin [27] showed that *novelty seeking* is not equivalent to *diversity seeking*, and the novelty preferences of individual users are directed towards finding more information on specific subtopics. Instead of absolute relevance judgments, a series of studies [28, 29] explored how to perform diversity evaluation based on preference judgements. Furthermore, it has been shown that many other factors such as recency and length can also influence user preferences for one document over another in the context of novelty and diversity.

To clearly show the commonalities and differences among metrics, [11, 12, 30] comparatively studied the properties of traditional metrics (e.g., DCG [31] and precision) within a general framework. Chuklin et al. [30] showed that click-model based metrics are more strongly correlated with online experimental results. For diversity metrics, Clarke *et al.* [10] examined the properties of cascade diversity measures. Chapelle et al. [32] analyzed the properties of a number of diversity metrics using the notion of submodularity, e.g., ERR is a submodular metric. The marginal relevance by Carbonell and Goldstein [33] is defined at a query level, i.e., a document has high marginal relevance if it is both relevant to the query and contains minimal similarity to previously ranked documents. Indeed intrinsic diversity emphasizes high marginal relevance among documents w.r.t. a single information need. The coverage-based framework [18] being very related to our work is deployed to perform search result diversification rather than diversity evaluation. To the best of our knowledge, this is the first work to explore the importance of intrinsic diversity to the overall diversity evaluation.

3. A General Framework

In economics, utility is a quantifiable concept, and is defined as the gain (say satisfaction or pleasure) a user gets when purchasing a product or service. The marginal utility of a product is the gain/loss from an incremental/decremental consumption of that product or service [34]. In the context of document ranking, we use *utility* to refer to the satisfaction or contentment a user gets when provided with a single document. Assuming that users browse documents from top to down, *marginal utility* of a document refers to the additional satisfaction a user gets given the previously ranked documents. The distinction between utility and marginal utility is that: for a particular information need, the utility of a document d_k is usually definite, but its marginal utility depends on the previous k-1 documents. For clarification, suppose that two identical documents d_1

and d_2 are highly relevant to a subtopic, and d_1 is ranked ahead of d_2. Although they have the same utility, the marginal utility of d_2 appears to be very small, say zero, because d_2 merely provides redundant information.

Based on marginal utility, we propose a general framework to model the total effectiveness of a result list.

$$\mathcal{T}(L) = \frac{1}{\mathcal{N}} \sum_{k=1}^{n} w(k) \mathcal{M}(d_k | L^{k-1}) \tag{1}$$

where $\mathcal{T}(L)$ denotes the total effectiveness of the ranked list L. $L = \{d_1, ..., d_n\}$ represents a ranked list of documents and $k = 1, ..., n$ represents the ranked position. L^k is the sublist of *top-k* documents ($L^{k-1} \equiv \emptyset$ when $k = 1$). $\mathcal{N}$ is a normalizing factor, $w(k)$ is a function of rank positions. $\mathcal{M}(d_k|L^{k-1})$ represents the *query-level* marginal utility of d_k given the previous documents L^{k-1}. It equals to the utility of d_k (denoted as $\mathcal{U}(d_k)$) when $k = 1$, since the user would not have seen any other documents. Though the framework given by Equation 1 is simple, later we will see that the usage of marginal utility at different levels (i.e., query and subtopic) helps to understand how different metrics capture extrinsic diversity and intrinsic diversity.

4. Analyses of Diversity Metrics

In this section, we separate the components of different diversity metrics using the framework given by Equation 3, so as to understand how they measure intrinsic diversity.

In the context of diversified document ranking, the marginal utility of a document relies on its relevance to different subtopics as well as on the previously ranked documents. The commonly-used formulation of query-level marginal utility is:

$$\mathcal{M}(d_k \mid L^{k-1}) = \sum_{i=1}^{m} f(t_i) \mathcal{M}_i(d_k | L^{k-1}) \tag{2}$$

where t represents a first-level subtopic, $i = 1, ..., m$ denotes its index. $f(t_i)$ is a weight function of subtopics. With a subscript i, $\mathcal{M}_i(d_k|L^{k-1})$ denotes the marginal utility of d_k w.r.t. t_i. We will refer to it as *subtopic-specific marginal utility function*. In other words, the query-level marginal utility is expressed as the sum over the product of a weight function of subtopics and the subtopic-specific marginal utility. Analogously, we have $\mathcal{U}(d_k) = \sum_{i=1}^{m} f(t_i)\mathcal{U}_i(d_k)$. Chandar and Carterette [29] proposed another way to quantify query-level marginal utility based on preference judgements. Considering that they finally resorted to fine-grained relevance judgments to extract pair-wise preferences, we use the common way (i.e., Equation 2) instead, and leave the exploration of their method for future work. Substituting Equation 2 into Equation 1, it yields,

$$\mathcal{T}(L) = \frac{1}{\mathcal{N}} \sum_{k=1}^{n} w(k) \sum_{i=1}^{m} f(t_i) \mathcal{M}_i(d_k | L^{k-1}) \tag{3}$$

By swapping the order of summations, $\mathcal{T}(L)$ can be given as:

$$\mathcal{T}(L) = \frac{1}{\mathcal{N}} \sum_{i=1}^{m} f(t_i) \sum_{k=1}^{n} w(k) \mathcal{M}_i(d_k|L^{k-1}) \tag{4}$$

Equation 4 can be interpreted as: $\sum_{k=1}^{n} w(k)\mathcal{M}_i(d_k|L^{k-1})$ corresponds to the effectiveness of L w.r.t. the i-th information need, and it determines the intrinsic diversity through the subtopic-specific marginal utility function. The extrinsic diversity is captured through a weighted combination of the per-subtopic effectiveness. In fact, Equation 3 and Equation 4 formulate an equivalent framework. Now it is clear that how the total effectiveness of a ranked list is quantified through capturing extrinsic diversity and intrinsic diversity at the same time. Section 4.2 will show that most diversity metrics can be viewed as exemplars of this framework.

4.1. Widely Used Diversity Metrics

Before reviewing the widely used diversity metrics, the notations used throughout the paper are first explained.

e: a second-level subtopic, e_o^i represents the *o-th* second-level subtopic underlying the *i-th* first level subtopic, and $o = 1, ..., v$ indicates its index.

$g \in \{0, ..., Y\}$: graded relevance value, e.g., the ternary scale with $Y = 2$, contains nonrelevant ($g = 0$), relevant ($g = 1$) and highly relevant ($g = 2$). In particular, g_i^k denotes the relevance value of d_k w.r.t. the *i-th* first-level subtopic. $g_{e_o^i}^k$ denotes the relevance value of d_k w.r.t. e_o^i.

$I(g)$, $R(g)$ and $V(g)$ are three functions that map graded relevance value to numerical values or the probability of relevance. In particular, $I(g) = 1$ if $g > 0$, otherwise $I(g) = 0$, it is used by *AP-IA*, *α-nDCG* and their variants. $R(g) = \frac{2^g - 1}{2^Y}$, it is used by *ERR* and its variants. Finally, $V(g) = g$ is used by *D#-nDCG* and *DIN#-nDCG*.

Subtopic recall is defined as the number of unique subtopics retrieved up to a given rank n divided by the total number of subtopics [5]:

$$SRecall(L) = \frac{|\cup_{k=1}^{n} fls(d_k)|}{m} \tag{5}$$

where $fls(d_k)$ denotes the set of first-level subtopics covered by document d_k.

Intent-aware family is a family of metrics [7] that perform evaluation by applying a traditional metric to each subtopic independently and that combine the results based on the importance or probability of subtopics. For example, intent-aware **average precision** (AP) [35] is expressed as:

$$AP\text{-}IA(L) = \sum_{i=1}^{m} p(t_i|q) AP_i(L) \tag{6}$$

where $AP_i(L) = \frac{1}{n} \sum_{k=1}^{n} I(g_i^k) \frac{c_i^k}{k}$, $c_i^k = \sum_{j=1}^{k} I(g_i^j)$ is defined as the number of documents ranked up to position k that are judged relevant to subtopic i.

Expected Reciprocal Rank (ERR) by Chapelle *et al.* [36] can be regarded as the expectation of the reciprocal of the rank of a result at which the user stops. ERR interprets the relevance probability of the document at rank k as the probability that the user is satisfied. It is defined as:

$$ERR(L) = \sum_{k=1}^{n} \frac{1}{k} R(g^k) \prod_{j=1}^{k-1} (1 - R(g^j)) \tag{7}$$

The intent-aware version is expressed as:

$$ERR\text{-}IA(L) = \sum_{i=1}^{m} p(t_i|q) ERR_i(L) \tag{8}$$

where $ERR_i(L) = \sum_{k=1}^{n} \frac{1}{k} R(g_i^k) \prod_{j=1}^{k-1} (1 - R(g_i^j))$.

α-nDCG [6] extends the standard metric of *normalised Discounted Cumulative Gain* (nDCG) [31] by rewarding newly retrieved subtopics and penalizing redundant subtopics. Clarke *et al.* [6] assume binary relevance assessments, and use parameter α to reflect the possibility of assessor error. The gain value for document d_k is computed by summing over subtopics, i.e., $G[k] = \sum_{i=1}^{m} I(g_i^k)(1-\alpha)^{c_i^{k-1}}$, and $c_i^{k-1} = \sum_{j=1}^{k-1} I(g_i^k)$ ($c_i^{k-1} \equiv 0$ when $k = 1$). The discounted cumulative gain of a ranked list is $\alpha\text{-}DCG(L) = \sum_{k=1}^{n} \frac{G[k]}{log_2(k+1)}$. To compare the scores across various queries, $\alpha\text{-}DCG$ has to be normalized.

$$\alpha\text{-}nDCG(L) = \frac{\alpha\text{-}DCG(L)}{\alpha\text{-}DCG^*(L^*)} \tag{9}$$

where $\alpha\text{-}DCG^*(L^*)$ denotes the maximum $\alpha\text{-}DCG$ value attained by the ideal ranking L^*.

D#-nDCG is a linear combination of *S-recall* and $nDCG$. Sakai *et al.* [8] extend nDCG by incorporating per-intent graded relevance. The global gain value for document d_k is computed as $GG[k] = \sum_{i=1}^{m} p(t_i|q) V(g_i^k)$, the discounted cumulative gain of a ranked list is $D\text{-}DCG(L) = \sum_{k=1}^{n} \frac{GG[k]}{log_2(k+1)}$. By using a trade-off parameter γ, $D\#\text{-}nDCG$ is given as:

$$D\#\text{-}nDCG(L) = \gamma SRecall(L) + (1-\gamma) D\text{-}nDCG(L) \tag{10}$$

where $D\text{-}nDCG(L) = \frac{D\text{-}DCG(L)}{D\text{-}DCG^*(L^*)}$, $D\text{-}DCG^*(L^*)$ is the maximum $D\text{-}DCG$ value attained by the ideal ranking L^*.

DIN#-nDCG [9] is an improved version of $D\#\text{-}nDCG$. $DIN\#\text{-}nDCG$ differentiates subtopic set $\{t\}$ as informational subtopics $\{t^{inf}\}$ and navigational subtopics $\{t^{nav}\}$. When computing the global gain, a system does not receive any credit for returning multiple relevant documents for a navigational subtopic. For d_k, its global gain is $GG^{DIN}[k] = \sum_{t_i \in \{t^{inf}\}} p(t_i|q) V(g_i^k) + \sum_{t_j \in \{t^{nav}\}} isnew_j(k) p(t_j|q) V(g_j^k)$, where $isnew_j(k) = 1$ if no relevant document is observed between ranks 1 and

r-1, and $isnew_j(k) = 0$ otherwise. Then the discounted cumulative gain of a ranked list is $DIN\text{-}DCG(L) = \sum_{k=1}^{n} \frac{GG^{DIN}[k]}{log_2(k+1)}$. Finally, DIN#-nDCG(L) is given as

$$DIN\#\text{-}nDCG(L) = \eta SRecall + (1-\eta) DIN\text{-}nDCG \tag{11}$$

where $DIN\text{-}nDCG = \frac{DIN\text{-}DCG(L)}{D\text{-}DCG^*(L^*)}$, and $D\text{-}DCG^*(L^*)$ is the same as $D\text{-}nDCG$ (Equation 10).

Metric	$\boldsymbol{\mathcal{N}}$	$\boldsymbol{f(t_i)}$	$\boldsymbol{w(k)}$	$\boldsymbol{\mathcal{M}_i(d_k \vert L^{k-1})}$
$AP\text{-}IA$	n	$p(t_i\vert q)$	$\frac{1}{k}$	$I(g_i^k)c_i^k$
$ERR\text{-}IA$	1	$p(t_i\vert q)$	$\frac{1}{k}$	$R(g_i^k)\prod_{j=1}^{k-1}(1-R(g_i^j))$
$\alpha\text{-}nDCG$	$\alpha\text{-}DCG^*(L^*)$	1	$\frac{1}{log_2(k+1)}$	$I(g_i^k)(1-\alpha)^{c_i^{k-1}}$
$D\text{-}nDCG$ of $D\#\text{-}nDCG$	$D\text{-}DCG^*(L^*)$	$p(t_i\vert q)$	$\frac{1}{log_2(k+1)}$	$V(g_i^k)$
$DIN\text{-}nDCG$ of $DIN\#\text{-}nDCG$	$D\text{-}DCG^*(L^*)$	$p(t_i\vert q)$	$\frac{1}{log_2(k+1)}$	$\begin{cases} V(g_i^k) & if\ inf(i)=1 \\ isnew_i(k)V(g_i^k) & if\ nav(i)=1 \end{cases}$

Table 1: A unified view of metrics for diversity evaluation

4.2. Summary of Metrics

Except for the simple non-rank based metric, $SRecall$, which uses set-theoretic operations, the other metrics shown in Section 4.1 can be viewed as exemplars of the framework by Equation 3, and are summarized in Table 1. We can observe that: except for $\alpha\text{-}nDCG$, which assumes that subtopics are equally probable (i.e., a probability γ), other metrics capture the extrinsic diversity through a weighted combination of per-subtopic effectiveness. About how these metrics capture intrinsic diversity via the subtopic-specific marginal utility function, Table 1 shows that: For $D\text{-}nDCG$, the subtopic-specific marginal utility by $V(g_i^k)$ does not take into account the documents at earlier ranks, so does $D\#\text{-}nDCG$ because it is a linear combination of $SRecall$ and $D\text{-}nDCG$. For $AP\text{-}IA$, its subtopic-specific marginal utility depends on the count of relevant documents that appeared before. $ERR\text{-}IA$ and $\alpha\text{-}nDCG$ use a similar geometric discount w.r.t. the number of the previously retrieved relevant documents for a subtopic. A small difference is that $\alpha\text{-}nDCG$ takes into account the possibility of assessor error by the parameter α [6]. The component $DIN\text{-}nDCG$ in $DIN\#\text{-}nDCG$ takes into account differences among subtopics, and different marginal utility functions are used for informational subtopics and navigational subtopics.

When merely using first-level subtopics as in the traditional diversity evaluation, all these metrics can not precisely capture intrinsic diversity. This is because they are unaware of the fine-grained content differences among relevant documents per subtopic. For example, two identical documents w.r.t. an information need can not be differentiated. Therefore, it is still unclear to what extent each information need has been satisfied in the traditional diversity evaluation.

4.3. Incorporating Subtopic Hierarchy

To sufficiently consider both extrinsic diversity and intrinsic diversity simultaneously, a straightforward way is to make the diversity metrics aware of the fine-grained knowledge per information need. In particular, we propose to evaluate a ranked list with a diversity metric against each first-level subtopic based on the second-level subtopics. Thus the intrinsic diversity can be captured as successfully as the extrinsic diversity in the traditional diversity evaluation. Finally the per-fls intrinsic diversity results are combined, so as to capture the extrinsic diversity at the same time. An intuitive way of combination is the linear combination with the function $f(t_i)$. Let $\mathcal{T}_i(L)$ be the intrinsic diversity result w.r.t. the first-level subtopic t_i, the total effectiveness of a ranked list would be:

$$\mathcal{T}(L) = \sum_{i=1}^{m} f(t_i)\mathcal{T}_i(L) \tag{12}$$

This gives rise to another family of metrics, referred to as *intent-square family*. For example, if we set $f(t_i)$ as $p(t_i|q)$, the intent-square ERR would be:

$$ERR\text{-}IS(L) = \sum_{i=1}^{m} p(t_i|q)ERR\text{-}IA_i(L) \tag{13}$$

where $ERR\text{-}IA_i(L) = \sum_{k=1}^{n} \frac{1}{k} \sum_{o=1}^{v} p(e_o^i|t_i, q)R(g_{e_o^i}^k) \prod_{j=1}^{k-1}(1 - R(g_{e_o^i}^k))$ is the $ERR\text{-}IA$ score of L for the *i-th* first-level subtopic based on its child second-level subtopics (Equation 8).

Similarly, intent-square subtopic recall would be:

$$SRecall\text{-}IS(L) = \sum_{i=1}^{m} p(t_i|q)SRecall_i(L) \tag{14}$$

where $SRecall_i(L) = \frac{|\cup_{k=1}^{n} sls_i(d_k)|}{|\{e_o^i\}|}$ represents the subtopic recall corresponding to the *i-th* first-level subtopic t_i, i.e., the number of unique *sls*s under t_i retrieved up to the rank n divided by the total number of *sls*s underlying t_i. The intent-square $\alpha\text{-}nDCG$ would be:

$$\alpha\text{-}nDCG\text{-}IS(L) = \sum_{i=1}^{m} p(t_i|q)\alpha\text{-}nDCG_i(L) \tag{15}$$

where $\alpha\text{-}nDCG_i(L) = \frac{\sum_{k=1}^{n} \frac{1}{log_2(k+1)} \sum_{o=1}^{v} I(g_{e_o^i}^k)(1-\alpha)^{c_{e_o^i}^{k-1}}}{\alpha\text{-}DCG_i^*(L^*)}$ represents the $\alpha\text{-}nDCG$ score of L for the *i-th* first-level subtopic calculated with Equation 9 based on second-level subtopics. For the other diversity metrics, they are not further extended due to a space reason.

4.3.1. Relationships with Prior Metrics

To clearly show the relationships between the intent-square metrics and prior metrics, we take $ERR\text{-}IS$, $ERR\text{-}IA$ and ERR as an example. Specifically, by swapping the order of the summations w.r.t. i and k, we can find that $ERR\text{-}IS$ by Equation 13 can be equal to the case when we directly feed the entire set of second-level subtopics into $ERR\text{-}IA$, where the weight of each second-level subtopic should be $p(t_i|q)p(e_o^i|t_i, q)$. This reveals that the essence of incorporating subtopic hierarchy for diversity evaluation boils down to adjusting probabilities or weights of fine-grained subtopics based on the pre-defined subtopic hierarchy.

We note that the work by Chen et al. [25] using taxonomy-aware decay functions and the study by Wang et al. [26] building upon the extended intent hierarchy can also be viewed as exemplars of the framework by Equation 3. For example, the subtopic-specific marginal utility of the layer-aware metric $ERR\text{-}IA\text{-}LA$ in [26] is computed as the sum of each layer's marginal utility according to the extended intent hierarchy. However, the adopted intent hierarchies upon which new metrics are proposed, are different. For example, the measures proposed by Wang et al. [26] rely on the particular subtopic hierarchy (i.e., extended intent hierarchy). In order to ensure that all the leaf nodes of an extend intent hierarchy have the same height, some subtopics themselves are directly used as their children nodes. This operation which might result in a potential *redundancy issue* is not allowed in our work. Since we focus on investigating the impact of intrinsic diversity, a detailed comparison with [25, 26] and a further exploration on tuning $f(t_i)$ are left for future work.

5. Experiments

In this section, we conduct a series of experiments in order to clearly explore the impact of intrinsic diversity on novelty and diversity evaluation. We first detail the gold standard collection and the set of submitted runs used in the experiments. We then compare the aforementioned diversity metrics from the following aspects: system ranking, rank correlation and discriminative power.

5.1. Collection

As the basis for our experiments we adopt the standard test collections (Chinese and English) released in IMine *Diversified Ranking* task[1] of NTCIR-11. Each query set consists of 17 clear queries and 33 unclear queries (i.e., broad queries and ambiguous queries). In IMine the retrieval results for the unclear queries should be diversified. For each unclear query, a two-level hierarchy of subtopics is used to depict the underlying information needs and different aspects of each information need. An example query is shown in Fig. 1, where *poss* means the possibility of each subtopic (the examples of each subtopic are

[1] http://www.thuir.org/IMine/

not listed). Moreover, for each query, a pool of documents are annotated with graded relevance assessments, i.e., the first-level and second-level subtopics that a document is relevant to. These graded relevance values are used to compute either the probability of relevance or numerical scores (e.g., $I(g)$, $R(g)$ and $V(g)$ in section 4.1). As an example, Fig. 2 shows partial relevance annotations regarding query 0083 *harry potter* (Fig. 1).

```
<doc docid="clueweb12-1400tw-48-17405" queryid="0083" relevance="3" sls="harry potter book_reading"/>
<doc docid="clueweb12-1614wb-36-07554" queryid="0083" relevance="3" sls="harry potter games_quiz"/>
<doc docid="clueweb12-1513wb-01-01964" queryid="0083" relevance="2" sls="harry potter film_cast"/>
<doc docid="clueweb12-1613wb-56-03786" queryid="0083" relevance="3" sls="harry potter film_cast"/>
<doc docid="clueweb12-0800tw-67-17216" queryid="0083" relevance="3" sls="harry potter book_reading"/>
```

Figure 2: Partial document relevance annotations for query 0083 *harry potter*

Our query set consists of *30* Chinese queries and *32* English queries (queries numbered 0003, 0017, 0033 and 0070 are discarded due to the lack of the second-level subtopics for some of the first-level subtopics). For Chinese queries, the average number of first-level subtopics per query is 3.5, and the average number of second-level subtopics per subtopic is 5.64. For English queries, the numbers are 3.97 and 3.17, respectively. To look into performance of different metrics we use runs submitted to IMine (10 Chinese runs and 15 English runs).

We use $ERR\text{-}IA$ and $\alpha\text{-}nDCG$ to perform a traditional diversity evaluation, i.e., they are deployed merely using the first-level subtopics. The intent-square metrics $SRecall\text{-}IS$, $ERR\text{-}IS$ and $\alpha\text{-}nDCG\text{-}IS$ that use a two-level hierarchy of subtopics are then used to explore the impact of intrinsic diversity. The default cutoff value is 20. For $\alpha\text{-}nDCG$ and its intent-square variant, α is set as 0.5.

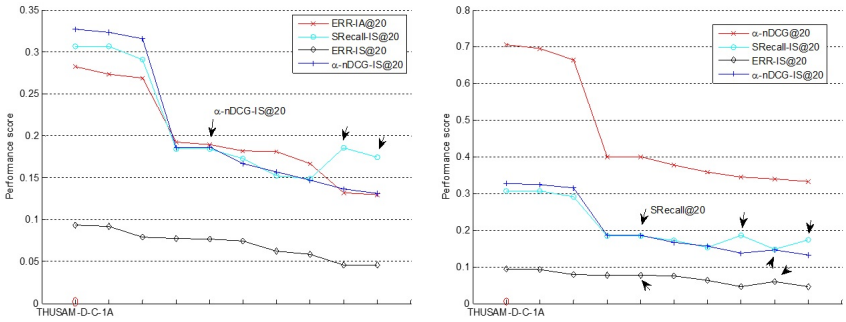
5.2. System Ranking

In this section, we compare $ERR\text{-}IA$, $\alpha\text{-}nDCG$ and the intent-square metrics by examining the differences when ordering a number of system results. Then we investigate the effect of query types during diversity evaluation. Furthermore, we conduct experiments at a fine-grained per-query level in order to well understand the agreements and disagreements among these metrics.

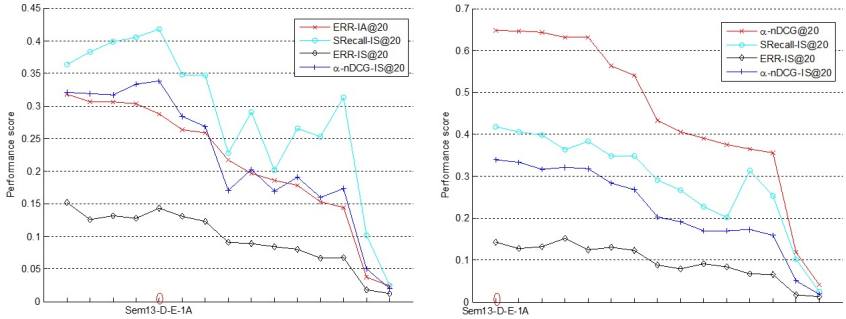
5.2.1. Overall Performance Order Comparison

In Fig. 3(a), Fig. 3(b), Fig. 3(c) and Fig. 3(d), we compare the run rankings (i.e., performance order) of IMine Chinese/English diversity runs evaluated with the intent-square metrics against $ERR\text{-}IA$ and $\alpha\text{-}nDCG$ respectively. Take Fig. 3(a) for example. Each point represents the performance score of a Chinese submitted run evaluated with a given metric. The Chinese runs are sorted on the x-axis in descending order of their scores given by $ERR\text{-}IA$@20. Each increase in the value of y-axis here along with the increase of x-axis indicates a disagreement with $ERR\text{-}IA$@20. A consistently decreasing trend indicates a consistent correlation. Fig. 3(b), Fig. 3(c) and Fig. 3(d) are obtained similarly.

When looking at the results for Chinese runs shown in Fig. 3(a) and Fig. 3(b), we see that there are few points that exhibit an inconsistent correlation



(a) Chinese runs ordered by $ERR\text{-}IA$@20 (b) Chinese runs ordered by $\alpha\text{-}nDCG$@20

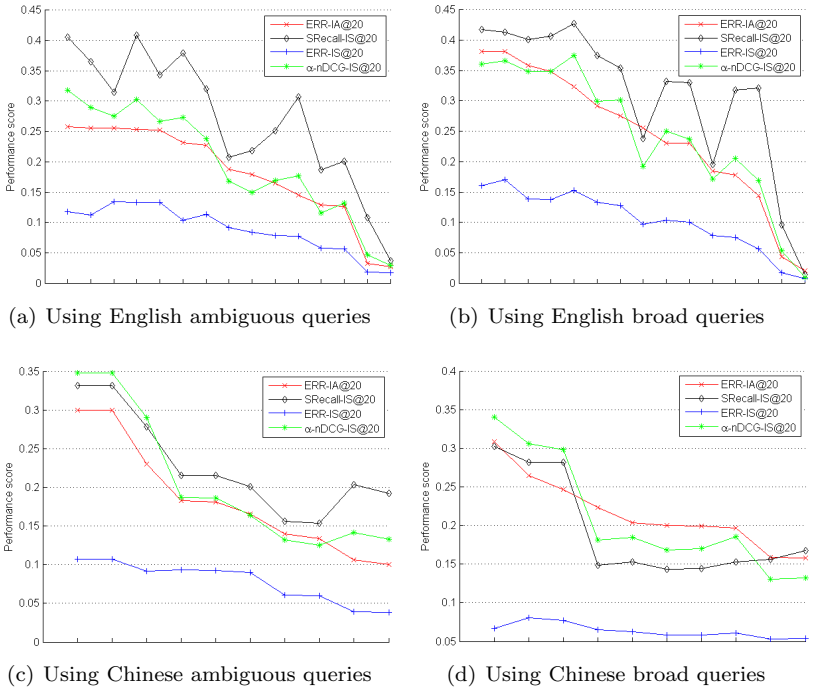


(c) English runs ordered by $ERR\text{-}IA$@20 (d) English runs ordered by $\alpha\text{-}nDCG$@20

Figure 3: Run ranking comparison among $ERR\text{-}IA$, $\alpha\text{-}nDCG$ and the intent-square metrics based on IMine Chinese/English diversity runs.

against $ERR\text{-}IA$@20 or $\alpha\text{-}nDCG$@20. For clarity, the points indicating significant inconsistence are marked with arrows. The matric name is labeled if the corresponding point overlaps with other metrics' values (e.g, $\alpha\text{-}nDCG\text{-}IS$@20 in Fig. 3(a)). This means that the intent-square metrics agree on the performance of many Chinese runs. On the other hand, for English runs shown in Fig. 3(c) and Fig. 3(d), we can find many points with the increases in the positions of x-axis. Take the points corresponding to the run *Sem13-D-E-1A* in Fig. 3(c) for example. $SRecall\text{-}IS$@20 and $\alpha\text{-}nDCG\text{-}IS$@20 achieve the maximum score respectively, $ERR\text{-}IS$@20 achieves a higher score than the left three runs. This reveals that the intent-square metrics disagree on the performance of Sem13-D-E-1A with $ERR\text{-}IA$@20.

5.2.2. Performance Order Comparison per Query Type



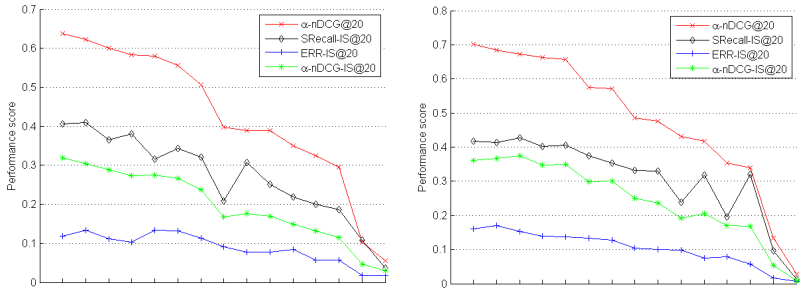
(a) Using English ambiguous queries (b) Using English broad queries

(c) Using Chinese ambiguous queries (d) Using Chinese broad queries

Figure 4: Run ranking comparison among $ERR\text{-}IA$ and the intent-square metrics based on Chinese/English ambiguous queries and Chinese/English broad queries, respectively.

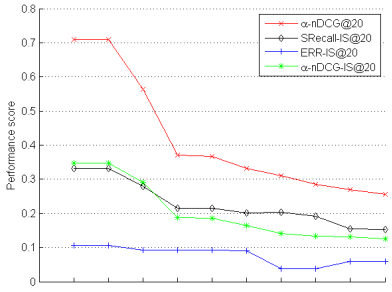
We now investigate the effect of query types when deploying specific metrics for diversity evaluation. Specifically, we categorize the adopted queries as either "ambiguous" or "broad" (using the labels provided by task organizers). Regarding the differences between ambiguous queries and broad queries, the TREC

assumption[2] goes like this: For an ambiguous query that has diverse interpretations, the users are assumed to be interested in only one of these interpretations. For a broad query (also called *faceted* query in TREC Web Track) that reflects an underspecified subtopic of interest, the users are assumed to interested in one subtopic, but may still be interested in others as well. In particular, for the English collection, there are 16 ambiguous queries and 16 broad queries. For the Chinese collection, they are 15 and 15, respectively. Fig. 4 and Fig. 5 show the run ranking comparison among $ERR\text{-}IA$, $\alpha\text{-}nDCG$ and the intent-square metrics based on Chinese/English ambiguous queries and Chinese/English broad queries, respectively. Take Fig. 4(a) for example. Each point represents the performance score of an English submitted run evaluated with a given metric. The English runs are sorted on the x-axis in descending order of their scores given by $ERR\text{-}IA$@20. Each increase in the value of y-axis here along with the increase of x-axis indicates a disagreement with $ERR\text{-}IA$@20. A consistently decreasing trend indicates a consistent correlation. Fig. 4(b), Fig. 4(c), Fig. 4(d), Fig. 5(a), Fig. 5(b), Fig. 5(c) and Fig. 5(d) are obtained similarly.

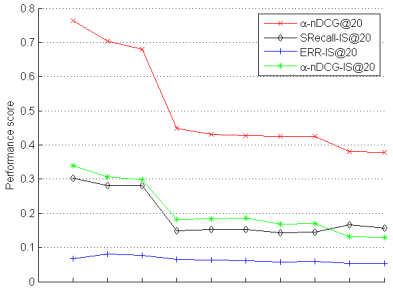


(a) Using English ambiguous queries

(b) Using English broad queries



(c) Using Chinese ambiguous queries



(d) Using Chinese broad queries

Figure 5: Run ranking comparison among $\alpha\text{-}nDCG$ and the intent-square metrics based on Chinese/English ambiguous queries and Chinese/English broad queries, respectively.

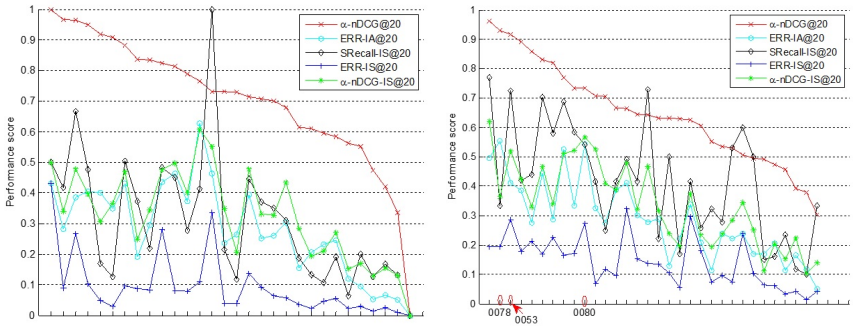
[2]http://plg.uwaterloo.ca/~trecweb/2010.html

At first glance, both Fig. 4 and Fig. 5 reveal that intent-square metrics exhibit inconsistent correlations against either $ERR\text{-}IA@20$ or $\alpha\text{-}nDCG@20$ based on either ambiguous queries or broad queries. Moreover, the inconsistences between intent-square metrics and $ERR\text{-}IA@20/\alpha\text{-}nDCG@20$ with respect to the ordered system rankings are more obvious based on English queries.

A closer comparison between results over English/Chinese ambiguous queries (i.e., Figs. 4(a), 4(c), 5(a), 5(c)) and results over English/Chinese broad queries (i.e., Figs. 4(b), 4(d), 5(b), 5(d)) reveal that: for both ambiguous queries and broad queries, the intent-square metrics disagree with either $ERR\text{-}IA@20$ and $\alpha\text{-}nDCG@20$. This is because intent-square metrics can measure the intrinsic diversity of the first-level subtopics of either an ambiguous query or a broad query.

5.2.3. Performance Order Comparison per Query

It should be noted that the agreements or disagreements shown in Fig. 6, Fig. 4 and Fig. 5 base on the averaged performance across a number of Chinese queries or a number of English queries.



(a) Per-query comparison of THUSAM-D-C-1A

(b) Per-query comparison of Sem13-D-E-1A

Figure 6: Per-query comparison based on two specific runs.

To get better understanding of the agreements and disagreements, Fig. 6(a) and Fig. 6(b) show a per-query comparison between the intent-square metrics and $\alpha\text{-}nDCG$ based on two top-ranked sample runs, i.e., the Chinese run *THUSAM-D-C-1A* shown in Fig. 3(b), and the English run Sem13-D-E-1A shown in Fig. 3(d). Let us look at Fig. 6(b) for example. Each point represents the performance score of the ranked results from Sem13-D-E-1A for a given query. All English queries are sorted in descending order on the x-axis by $\alpha\text{-}nDCG@20$. Each increase in the position of x's indicates disagreement on the performance of ranked results for a query with $\alpha\text{-}nDCG@20$. Although the intent-square metrics agree with $\alpha\text{-}nDCG$ on the top-one run in Fig. 3(b) (i.e, it is THUSAM-D-C-1), Fig. 6(a) illustrates that they actually disagree a lot on the per-query performance. Fig. 6(b) shows a similar phenomenon. For

ERR-IA, the per-query comparison is the same and not shown due to no space.

Query(0078)	$t_1(3)$,$t_2(3)$,$t_3(1)$
Query(0053)	$t_1(5)$,$t_2(2)$,$t_3(1)$,$t_4(2)$
Query(0080)	$t_1(4)$,$t_2(9)$

Table 2: Two-level subtopic information of queries 0078, 0053 and 0080.

Ranked list	Query (0078)	Query (0053)	Query (0080)
d_1	$t_1 : \{e_1^1\}$ $t_2 : \{e_1^2\}$	$t_4 : \{e_2^4\}$	$t_1 : \{e_1^1, e_2^1\}$
d_2	$\boldsymbol{t_1 : \{e_1^1\}}$	$t_3 : \{e_1^3\}$	$t_1 : \{e_3^1\}$
d_3	$t_1 : \{e_3^1\}$	$t_2 : \{e_1^2, e_2^2\}$	$\boldsymbol{t_1 : \{e_1^1, e_2^1\}}$
d_4	$\boldsymbol{t_1 : \{e_1^1\}}$ $\boldsymbol{t_2 : \{e_1^2\}}$	$\boldsymbol{t_2 : \{e_1^2, e_2^2\}}$	$\boldsymbol{t_1 : \{e_2^1\}}$ $t_2 : \{e_9^2\}$
d_5		$t_1 : \{e_2^1, e_5^1\}$	$t_2 : \{e_2^2, e_8^2\}$
d_6	$\boldsymbol{t_2 : \{e_1^2\}}$	$\boldsymbol{t_4 : \{e_2^4\}}$	$\boldsymbol{t_2 : \{e_9^2\}}$
d_7	$\boldsymbol{t_2 : \{e_1^2\}}$	$\boldsymbol{t_1 : \{e_2^1, e_5^1\}}$	$\boldsymbol{t_1 : \{e_3^1\}}$
d_8		$\boldsymbol{t_1 : \{e_5^1\}}$	$\boldsymbol{t_1 : \{e_3^1\}}$
d_9	$\boldsymbol{t_2 : \{e_1^2\}}$	$\boldsymbol{t_3 : \{e_1^3\}}$	$\boldsymbol{t_1 : \{e_2^1\}}$
d_{15}		$\boldsymbol{t_1 : \{e_2^1, e_5^1\}}$	

Table 3: Ranked results for queries 0078, 0053 and 0080 from the English run: Sem13-D-E-1A

To get a deeper understanding of the query-level disagreements, we select three example queries 0078, 0053 and 0080 (the 2nd, 3rd and 10th queries in Fig. 6(b)) for further analysis. Table 2 shows their subtopic information. For example, query 0078 has three first-level subtopics (i.e., t_1, t_2 and t_3). The numbers in brackets denote the numbers of underlying second-level subtopics, e.g., there are 3 second-level subtopics underlying t_1.

Table 3 illustrates the ranked list from Sem13-D-E-1A for each query respectively (a cutoff of 20). The first column shows the documents that are relevant to at least one query (nonrelevant documents are not shown). The 2nd, 3rd and 4th columns are the official relevance assessments of each document corresponding to each query. For example, for query 0053, the document d_3 is relevant to t_2, and e_1^2 and e_2^2 are covered. For clarity, the documents that contain redundant information given prior documents are marked in bold.

Table 4 shows the performance scores of the ranked lists w.r.t. queries 0078, 0053 and 0080 measured with *ERR-IA*, α-*nDCG* and the intent-square metrics, where @5 and @20 denote the cutoff values. For each metric, the maximum performance score by @5 among the three queries is underlined, and the maximum performance score by @20 is marked in bold.

SRecall-IS reveals the average extent to which the second-level subtopics underlying a first-level subtopic are covered. Low values of *SRecall-IS* serve as

Metric	Query (0078)		Query (0053)		Query (0080)	
	@5	@20	@5	@20	@5	@20
ERR-IA	0.5515	**0.5533**	0.393	0.4109	0.5312	0.5431
α-nDCG	0.8922	**0.9294**	0.992	0.9169	0.7101	0.7325
SRecall-IS	0.3333	0.3333	0.725	**0.725**	0.5417	0.5417
ERR-IS	0.1929	0.1934	0.2767	**0.2855**	0.2674	0.2729
α-nDCG-IS	0.4119	0.3643	0.4404	0.5198	0.5684	**0.5665**

Table 4: Performance scores with *ERR-IA*, *α-nDCG* and the intent-square metrics.

an indication of a poor average intrinsic diversity, i.e., a poorly diversified result list for each subtopic. For example, for query 0078, one aspect of the subtopic t_1 (i.e., e_2^1), two aspects of the subtopic t_2 (i.e., e_2^2 and e_3^2) and the subtopic t_3 are not satisfied. For query 0053, the subtopics t_2 and t_3 are well covered, only three aspects of the subtopic t_1 and three aspects of the subtopic t_4 are not covered, thus it is straightforward that query 0053 attains a higher *SRecall-IS* score (i.e., 0.725) than that of query 0078 (i.e., 0.3333). So does query 0080.

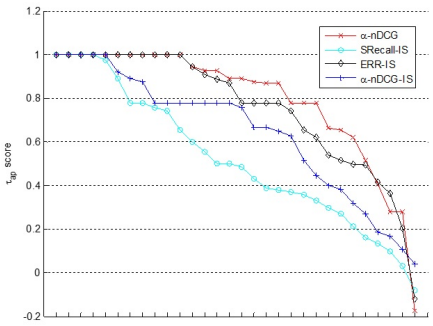
However, the high performance scores for query 0078 with *ERR-IA* (by @5 and @20) and *α-nDCG* (by @20) seem counterintuitive, because Table 3 clearly shows that the ranked list for query 0078 contains more redundant information compared with the result lists of the other two queries. This is not surprising, since as analyzed in Section 4.2, *ERR-IA* and *α-nDCG* are unaware of the fine-grained content differences underlying each subtopic. Thus they are not able to capture well the intrinsic diversity desired for each subtopic. A resulting effect is inability to precisely quantify the overall effectiveness of a ranked list. On the contrary, the intent-square metrics are based on a two-level subtopic hierarchy and have the advantage of better quantifying subtopic-specific marginal utility. Thus they can measure the overall effectiveness of a ranked list more precisely. Because *SRecall-IS* is a simple set-based metric, ties will occur when the number of covered unique *fls*s and *sls*s are the same (e.g., the *SRecall-IS* scores for queries 0053 and 0080 in Table 4). *ERR-IS* and *α-nDCG-IS* are more powerful than *SRecall-IS*, because the position and relevance grade of a specific relevant document are used when quantifying the effectiveness of a ranked list.

Based on the analyses shown above, it is reasonable to say that: *benefiting from the usage of fine-grained subtopics, the intent-square metrics can better capture intrinsic diversity.* On the contrary, *merely feeding first-level subtopics into the diversity metrics can not ensure the precise evaluation of intrinsic diversity. Namely, the satisfaction degree of each particular information need is unclear.* Since intrinsic diversity plays an important role in the overall diversity evaluation, the measures of the overall effectiveness of a diversified retrieval system and the system rankings would be greatly affected.

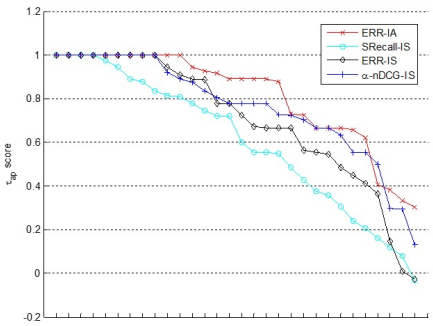
5.3. Rank Correlation

We further analyze the rank correlations among the cascade metrics (*ERR-IA* and *α-nDCG*) and the intent-square metrics. Kendall's τ is a widely used measure for comparing rank correlation [10, 37]. τ score ranges from -1 to $+1$, with 1 indicating the perfect agreement (i.e., two rankings are exactly the same), 0 indicating a random reordering and -1 indicating that the compared lists are reversed. Prior studies suggest that τ score of 0.9 or higher indicates high similarity between rankings while a score of 0.8 or lower indicates a significant difference [37]. In this paper, we use its variant version called *AP correlation* (τ_{ap}) [38], which is more sensitive to discrepancies among the top-ranked items.

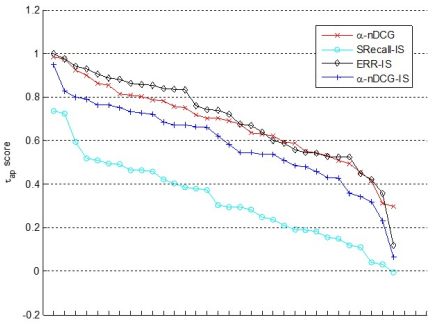
Figures 7(a), 7(b), 7(c) and 7(d) illustrate the rank correlation between *ERR-IA* (*α-nDCG*) and another four metrics (i.e., *α-nDCG* (*ERR-IA*) and the intent-square metrics) on a per-query basis.



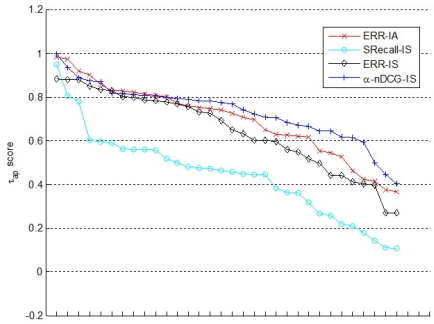
(a) Rank correlation against *ERR-IA* (Chinese)



(b) Rank correlation against *α-nDCG* (Chinese)



(c) Rank correlation against *ERR-IA* (English)



(d) Rank correlation against *α-nDCG* (English)

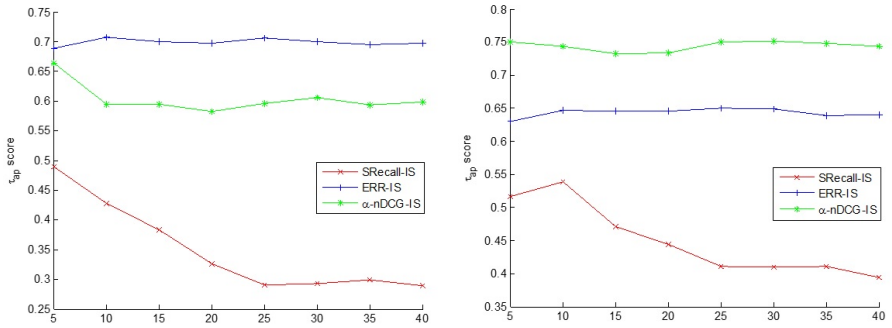
Figure 7: Rank correlation between per-query run rankings based on IMine Chinese/English diversity runs

Let's look at Fig. 7(a) for example. It depicts the rank correlation between *ERR-IA* and the other four metrics (*α-nDCG*, *Srecall-IS*, *ERR-IS*

and α-$nDCG$-IS) based on the Chinese runs. Each point represents the τ_{ap} score between two run rankings corresponding to a query. These per-query τ_{ap} scores between ERR-IA and another metric are sorted in descending order on the x-axis. Furthermore, Table 5 shows the average τ_{ap} score across all queries w.r.t. Figures 7(a), 7(b), 7(c) and 7(d).

	ERR-IA		α-$nDCG$	
	English	Chinese	English	Chinese
$SRecall$-IS	0.3266	0.5198	0.4438	0.6034
ERR-IS	0.697	0.7547	0.6452	0.7028
α-$nDCG$-IS	0.5824	0.6439	0.7342	0.763

Table 5: Average $\tau_{ap}score$.



(a) ERR-IA (b) α-$nDCG$

Figure 8: Variation of average $\tau_{ap}score$ w.r.t. different cutoff values.

From Figures 7(a) and 7(b), we find that for Chinese runs, ERR-IA and α-$nDCG$ are relatively highly correlated compared with their correlation with the intent-square metrics. The correlations between $Srecall$-IS and the cascade metrics (ERR-IA and α-$nDCG$) are especially low. Figures 7(c) and 7(d) exhibit a similar phenomenon over the English runs. Fig. 5 shows that the average correlation over the English runs is relatively lower than that over the Chinese runs. A probable reason is due to the differences between distributions of subtopics and relevance assessments of Chinese collection and English collection.

To explore the effect of different cutoff values, Figures 8(a) (ERR-IA) and 8(b) (α-$nDCG$) further show the variation of average τ_{ap} score across all queries (the y axis) w.r.t. different cutoff values from 5 to 40 (the x axis).

We can observe that: the average τ_{ap} correlation between the cascade metrics (ERR-IA and α-$nDCG$) and $SRecall$-IS decreases significantly with the increase of the cutoff value. Since ERR-IS and α-$nDCG$-IS take into account more factors (e.g., rank positions and relevance grades) than $SRecall$-IS, their

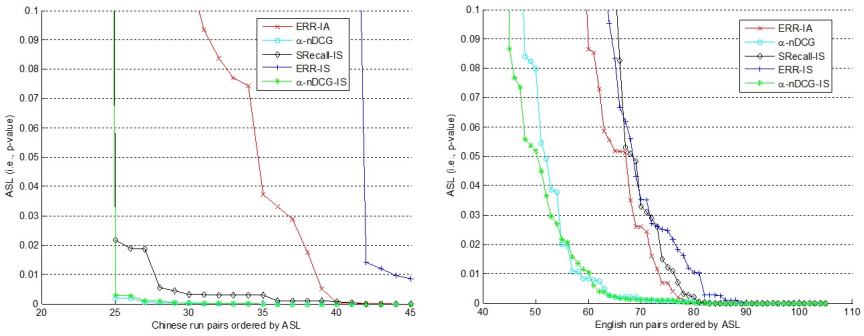
results should be more reliable. A further exploration of correlations among the intent-square metrics would be an interesting future work.

In summary, the above results show that *the AP correlation scores between the diversity metrics (ERR-IA and α-nDCG) that merely using first-level subtopics and the intent-square metrics are fairly low.* In other words, *feeding fine-grained subtopics into the diversity metrics substantially affects the system rankings.*

5.4. Discriminative Power

Given a test collection and a set of runs, *discriminative power* [39, 9] is frequently used for comparing metrics. Although high values of discriminative power do not ensure a good metric, extremely low values serve as an indication of a poor ability of distinguishing different rankings. In this paper, we use the randomised Tukey's Honestly Significant Differences (THSD) test [40] to conduct a statistical significance test. Our choice is motivated by the observation that it takes the entire set of runs into account when judging the significance of each run pair, and hence it is less likely to lead to significant differences that are not "real" (compared with the bootstrap test [39]).

Figures 9(a) and 9(b) illustrate the ASL (achieved significance level) curves of *ERR-IA*, *α-nDCG* and the intent-square metrics using the randomised THSD test (the trial number $B = 5000$) on the IMine diversity runs ($10 * (10 - 1)/2 = 45$ Chinese run pairs, $15 * (15 - 1)/2 = 105$ English run pairs). In ASL plots, the closer the metric's curve is to the origin, the higher the discriminative power this metric has. Namely, it can detect more significant differences.



(a) Chinese runs (b) English runs

Figure 9: Discriminative Power on Chinese/English runs.

From Figures 9(a) and 9(b) we can observe that *α-nDCG* and *α-nDCG-IS* are the most discriminative metrics across Chinese and English runs, and *α-nDCG-IS* is more discriminative than *α-nDCG* over English runs since fine-grained subtopics are utilized. Due to a small number of run pairs, the ASL curves over the Chinese runs appear to be heavily skewed. Since there are more English run

pairs, the discriminative powers exhibited over the English runs would be more reliable. Somewhat surprisingly, $ERR\text{-}IA$ and $ERR\text{-}IS$ show weaker discriminative power than the set-based metric $Srecall\text{-}IS$, and $ERR\text{-}IS$ exhibits a weaker discriminative power than $ERR\text{-}IA$. We leave this as an interesting future work.

6. Conclusions And Future Work

By casting the existing diversity metrics (i.e., $AP\text{-}IA$, $ERR\text{-}IA$, $\alpha\text{-}nDCG$ $D\#\text{-}nDCG$ and $DIN\#\text{-}nDCG$) into a unified framework based on marginal utility, we show that their abilities of measuring intrinsic diversity rely on what kind of subtopic knowledge is provided. Furthermore, a series of experiments are conducted using a family of novel metrics (i.e., intent-square metrics) against the traditional way of diversity evaluation. The experimental results clearly uncover the previously-unknown importance of intrinsic diversity to the overall diversity evaluation. We believe our analyses based on the marginal utility framework and the experimental findings provide a novel view for better understanding the commonalities and differences among the aforementioned diversity metrics, which will be useful for exploring more effective diversity evaluation.

Although the intent-square metrics can better capture intrinsic diversity, it should be noted that the more elaborated (by human assessors) a collection is, the more susceptible the collection is to subjectiveness and annotation errors. In the future, besides a further study of the unsolved issues on the intent-square metrics (e.g., the impact of erroneous subtopic hierarchy), we plan to explore other possible ways to better capture the intrinsic diversity, as well as other methods to combine extrinsic diversity and intrinsic diversity.

References

[1] R. L. T. Santos, C. Macdonald, I. Ounis, Search result diversification, Foundations and Trends in Information Retrieval 9 (1) (2015) 1–90.

[2] M. Drosou, E. Pitoura, Search result diversification, SIGMOD Record 39 (1) (2010) 41–47.

[3] M. R. Vieira, H. L. Razente, M. C. N. Barioni, M. Hadjieleftheriou, D. Srivastava, C. Traina, V. J. Tsotras, On query result diversification, in: Proceedings of the 27th ICDE, 2011, pp. 1163–1174.

[4] F. Radlinski, P. N. Bennett, B. Carterette, T. Joachims, Redundancy, diversity and interdependent document relevance, in: SIGIR Forum, Vol. 43, 2009, pp. 46–52.

[5] C. X. Zhai, W. W. Cohen, J. Lafferty, Beyond independent relevance: methods and evaluation metrics for subtopic retrieval, in: Proceedings of the 26th SIGIR, 2003, pp. 10–17.

[6] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, I. MacKinnon, Novelty and diversity in information retrieval evaluation, in: Proceedings of the 31st SIGIR, 2008, pp. 659–666.

[7] R. Agrawal, S. Gollapudi, A. Halverson, S. Ieong, Diversifying search results, in: Proceedings of the 2nd WSDM, 2009, pp. 5–14.

[8] T. Sakai, R. Song, Evaluating diversified search results using per-intent graded relevance, in: Proceedings of the 34th SIGIR, 2011, pp. 1043–1052.

[9] T. Sakai, Evaluation with informational and navigational intents, in: Proceedings of the 21st WWW, 2012, pp. 499–508.

[10] C. L. Clarke, N. Craswell, I. Soboroff, A. Ashkan, A comparative analysis of cascade measures for novelty and diversity, in: Proceedings of the 4th WSDM, 2011, pp. 75–84.

[11] B. Carterette, System effectiveness, user models, and user utility: a conceptual framework for investigation, in: Proceedings of the 34th SIGIR, 2011, pp. 903–912.

[12] J. Wang, J. Zhu, On statistical analysis and optimization of information retrieval effectiveness metrics, in: Proceedings of the 33rd SIGIR, 2010, pp. 226–233.

[13] R. L. Santos, C. Macdonald, I. Ounis, Exploiting query reformulations for web search result diversification, in: Proceedings of the 19th WWW, 2010, pp. 881–890.

[14] V. Dang, W. B. Croft, Diversity by proportionality: an election-based approach to search result diversification, in: Proceedings of the 35th SIGIR, 2012, pp. 65–74.

[15] H. Yu, F. Ren, Search result diversification via filling up multiple knapsacks, in: Proceedings of the 23rd CIKM, 2014, pp. 609–618.

[16] W. Zheng, H. Fang, C. Yao, Exploiting concept hierarchy for result diversification, in: Proceedings of the 21st CIKM, 2012, pp. 1844–1848.

[17] H. Yu, J. Adam, R. Blanco, H. Joho, J. Jose, L. Chen, F. Yuan, A concise integer linear programming formulation for implicit search result diversification, in: Proceedings of the 10th WSDM, 2017, pp. 191–200.

[18] W. Zheng, X. Wang, H. Fang, H. Cheng, Coverage-based search result diversification, Journal of Information Retrieval 15 (5) (2012) 433–457.

[19] S. Hu, Z. Dou, X. Wang, T. Sakai, J. Wen, Search result diversification based on hierarchical intents, in: Proceedings of the 24th CIKM, 2015, pp. 63–72.

[20] L. Xia, J. Xu, Y. Lan, J. Guo, X. Cheng, Learning maximal marginal relevance model via directly optimizing diversity evaluation measures, in: Proceedings of the 38th SIGIR, 2015, pp. 113–122.

[21] Y. Yue, T. Joachims, Predicting diverse subsets using structural SVMs, in: Proceedings of the 25th ICML, 2008, pp. 1224–1231.

[22] T. Leelanupab, G. Zuccon, J. M. Jose, A comprehensive analysis of parameter settings for novelty-biased cumulative gain, in: Proceedings of the 21st CIKM, 2012, pp. 1950–1954.

[23] T. Leelanupab, G. Zuccon, J. M. Jose, Is intent-aware expected reciprocal rank sufficient to evaluate diversity, in: Proceedings of the 35th ECIR, 2013, pp. 738–742.

[24] Y. Zhu, Y. Lan, J. Guo, X. Cheng, S. Niu, Learning for search result diversification, in: Proceedings of the 37th SIGIR, 2014, pp. 293–302.

[25] F. Chen, Y. Liu, M. Zhang, S. Ma, L. Chen, A subtopic taxonomy-aware framework for diversity evaluation, in: Proceedings of EVIA 2013, 2013, pp. 9–16.

[26] X. Wang, Z. Dou, T. Sakai, J. Wen, Evaluating search result diversity using intent hierarchies, in: Proceedings of the 39th SIGIR, 2016, pp. 415–424.

[27] Y. Xu, H. Yin, Novelty and topicality in interactive information retrieval, Journal of the American Society for Information Science and Technology 59 (2) (2008) 201–215.

[28] B. Carterette, P. N. Bennett, Evaluation measures for preference judgments, in: Proceedings of the 31st SIGIR, 2008, pp. 685–686.

[29] P. Chandar, B. Carterette, Preference based evaluation measures for novelty and diversity, in: Proceedings of the 36th SIGIR, 2013, pp. 413–422.

[30] A. Chuklin, P. Serdyukov, M. De Rijke, Click model-based information retrieval metrics, in: Proceedings of the 36th SIGIR, 2013, pp. 493–502.

[31] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Transactions on Information Systems 20 (4) (2002) 422–446.

[32] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, S. Wu, Intent-based diversification of web search results: metrics and algorithms, Journal of Information Retrieval 14 (6) (2011) 572–592.

[33] J. Carbonell, J. Goldstein, The use of mmr, diversity-based reranking for reordering documents and producing summaries, in: Proceedings of the 21st SIGIR, 1998, pp. 335–336.

[34] M. Carl, Principles of Economics, Ludwig von Mises Institute, 518 West Magnolia Avenue, Auburn, Ala. 36832 U.S.A., 2007.

[35] S. Robertson, A new interpretation of average precision, in: Proceedings of the 31st SIGIR, 2008, pp. 689–690.

[36] O. Chapelle, D. Metlzer, Y. Zhang, P. Grinspan, Expected reciprocal rank for graded relevance, in: Proceedings of the 18th CIKM, 2009, pp. 621–630.

[37] C. Buckley, E. M. Voorhees, Retrieval evaluation with incomplete information, in: Proceedings of the 27th SIGIR, 2004, pp. 25–32.

[38] E. Yilmaz, J. A. Aslam, S. Robertson, A new rank correlation coefficient for information retrieval, in: Proceedings of the 31st SIGIR, 2008, pp. 587–594.

[39] T. Sakai, Evaluating evaluation metrics based on the bootstrap, in: Proceedings of the 29th SIGIR, 2006, pp. 525–532.

[40] B. A. Carterette, Multiple testing in statistical analysis of systems-based information retrieval experiments, ACM Transactions on Information Systems 30 (1) (2012) 4:1–4:34.