Information Retrieval Journal
https://doi.org/10.1007/s10791-019-09356-x

Low-cost, bottom-up measures for evaluating search result diversification

Zhicheng Dou[1,2] · **Xue Yang**[1,2] · **Diya Li**[3] · **Ji-Rong Wen**[1,2] · **Tetsuya Sakai**[4]

Received: 13 October 2018 / Accepted: 21 March 2019

Abstract
Search result diversification aims at covering different user intents by returning a diversified document list. Most existing diversity measures require a predefined set of intents for a given query, where it is assumed that there is no relationship across these intents. However, studies have shown that modeling a hierarchy of intents has some benefits over the standard measure of using a flat list of intents. Intuitively, having more layers in the intent hierarchy seems to imply that we can consider more intricate relationships between intents and thereby identify subtle differences between documents that cover different intents. On the other hand, manually building a rich intent hierarchy imposes extra cost and is probably not very practical. In light of these considerations, we first propose a measure to build a hierarchy of intents from a given set of flat intents by clustering per-intent relevant documents and thereby identifying subintents. Furthermore, in our second measure, we consider a variant of our first measure that clusters per-topic relevance documents rather than per-intent ones, which is also intent-free. In addition, we propose our third measure, a simple, completely intent-free measure to search result diversity evaluation, which leverages document similarities. Our experiments based on TREC Web Track 2009–2013 test collections show that our proposed measures have advantages over existing diversity measures despite their low annotation costs.

Keywords Search result diversification · Evaluation measure · Hierarchical clustering

1 Introduction

A web search query is often ambiguous or broad (Dou et al. 2007, 2009; Song et al. 2010). The query may have several interpretations, also known as intents. For example, the query "defender" can represent land rover defender (a car model), defender game (an arcade game), or windows defender (an anti-spyware program). Search result diversification aims at covering different user intents by returning a diversified document list.

✉ Zhicheng Dou
dou@ruc.edu.cn

Extended author information available on the last page of the article

Most existing measures (Dang and Croft 2012, 2013; Radlinski and Dumais 2006; Dou et al. 2011; Santos et al. 2010a, b, 2011; Zhu et al. 2007) assume that the user's information need can be represented by a flat list of intents. The quality of a ranked list is evaluated by considering the number of intents covered by returned documents, and the relevance of these documents to the intents. Hence, relationships across different intents are not considered. However, in some circumstances, some of the intents for the same query are related to each other, while others are not. To take this into account in search result diversity evaluation, it may be worthwhile to consider a hierarchy of intents instead of a flat list of intents. Intuitively, having more layers in the intent hierarchy seems to imply that we can consider more intricate relationships between intents and thereby identify subtle differences between documents that cover different intents.

To introduce intent hierarchy in search result diversity evaluation, Wang et al. (2016) proposed a measure to build *superintents* over a given set of intents, to consider the fact that some of the given intents are more related to each other than others are. For example, the middle column in Fig. 1 shows the official intents for the query "defender" from the TREC Web Track 2009 (Clarke et al. 2009). As shown in the figure, the measure of Wang et al. can build a superintent "Windows Defender" (Wang et al. 2016) over the official TREC intents "Windows Defender Homepage" and "Windows Defender Reports." Wang et al. reported that their measures based on hierarchical intents outperform traditional diversity measures in terms of *discriminative power* (Sakai 2006b), i.e., the ability to detect many pairwise statistically significant differences.

While Wang et al. (2016) built superintents over the official TREC intents, they did not consider the possibility that the official intents could also have *subintents*, even though there is no guarantee at all that the official intents are atomic. For example, by manually examining the intent-level relevant documents for the aforementioned TREC Web Track topic "defender," we found that some of the documents judged relevant to the official intent "Defender Arcade Games" are related to "Defender games download", while others are about "Playing defender games online", as shown in the rightmost column in Fig. 1. Hence, to complement the measure of Wang et al., we first propose **a measure that automatically builds subintents under a given set of official intents**, by applying hierarchical clustering of intent-level relevant documents provided in a standard diversity test collection with a flat intent list. Our hypothesis was that it may be beneficial to consider the distinction between these subintents in search result diversity evaluation.

Given a diversity test collection with intent-level relevance assessments, our first measure (shown in Fig. 2) mentioned above does not require any additional manual effort, as it only involves automatic clustering of the intent-level relevant documents. However, assessing documents per intent is still more costly than assessing them per topic; hence, we also consider the problem of diversity evaluation without intent-level relevance assessments. More specifically, our second measure is **a variant of our first measure that clusters per-topic relevance documents rather than per-intent ones**. This variant is also intent-free, as shown in the left bottom part of Fig. 2. Furthermore, we try to abandon the intent hierarchy based on time saving and model simplification. Our third measure is to **evaluate search result diversity solely based on the similarity between relevant documents**, so that we can avoid rewarding systems that return near-duplicate documents and those that cover the same subintent. The third measure is an average of a traditional evaluation measure such as nDCG (normalized discounted cumulative gain) (Järvelin and Kekäläinen 2000) and a score that represents the overall redundancy of the search result.

We show the three measures we propose in this paper and their relationships in Fig. 2. In summary, the first and second measure are trying to extend the existing hierarchical

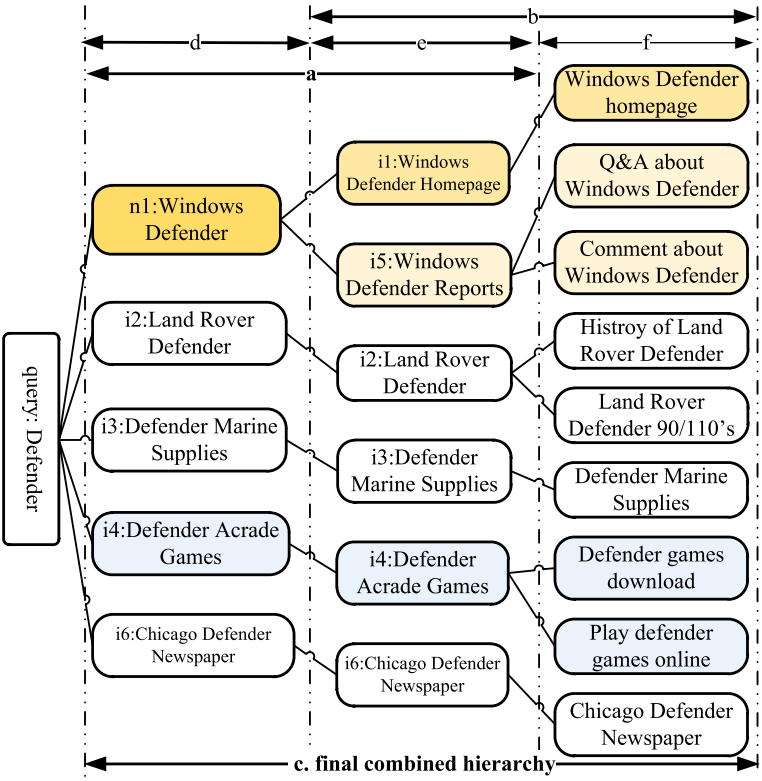

Fig. 1 Intent hierarchy of query "defender". Area 'd' refers to superintents, 'e' refers to the official intents (flat intents), 'f' refers to subintents. Area 'a' represents the intent hierarchy proposed by Wang et al., 'b' represents our intent hierarchy in this paper, 'c' represents the combination of two kinds of intent hierarchy

intents (Wang et al. 2016) by automatically building subintents, to improve the reliability and effectiveness of diversity evaluation. The second and the third measure are trying to reduce the annotation cost: they only require the topic-level relevance assessments. We evaluate these measures on the TREC Web Track 2009–2013 diversity test collections. The experimental results show that our measures that leverage intent hierarchies with subintents achieve higher discriminative power than existing flat-list measures, including I-recall (Sakai et al. 2010), α-nDCG (Clarke et al. 2008), IA-measures (Agrawal et al. 2009), and D♯-measures (Sakai and Song 2011). Moreover, the measures based on our intent hierarchies with subintents outperform those based on the superintent-based hierarchies of Wang et al. The highest discriminative power is achieved when these two measures are combined. Furthermore, we show that our first measure works well even when we start from the topic-level relevant documents instead of the intent-level ones. Our third measure based on document similarity also outperforms traditional measures in terms of discriminative power despite the fact that this measure does not require any explicit definitions of intents. Our proposed measures are also shown to be more consistent with the user's search result preferences than traditional measures. These results show that our low-cost, bottom-up measures to search result diversity evaluation are useful.

The main contributions of the papers are:

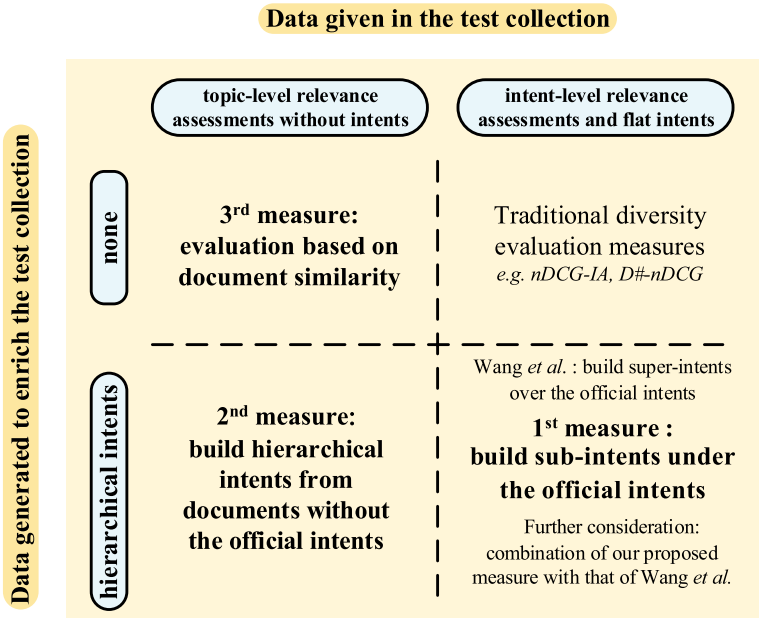

Fig. 2 The relationships of different measures. The horizontal axis indicates whether the measure requires intent level relevance assessments, and the vertical axis shows whether the measure requires hierarchical intents

- We propose three low-cost measures for evaluating search result diversification.
- We create a document clustering based method which could build intent hierarchy automatically. It can be performed either on topic-level (whole query level) or on intent-level (subtopic level).
- We make comparisons between our measures and existing measures. We find that our measures could achieve considerable results with lower cost.

The remainder of this paper is organized as follows. We briefly discuss related work in Sect. 2 including traditional diversity measures, and hierarchical diversity measures. We then propose our first measure and second measure for creating subintent hierarchies based on hierarchical clustering in Sect. 3. In Sect. 4, we introduce our third measure based on document similarity. We evaluate our measures on the TREC Web Track 2009–2013 diversity test collections, report and analyze experimental results in Sect. 5. We make a discussion about the drawback of our measures in Sect. 6. We finally discuss and conclude our work in Sect. 7.

2 Related work

2.1 Document relevance and redundancy in retrieval models

Relevance and redundancy have been widely discussed in the field of information retrieval. Many search result diversification models have been proposed. For example, Maximal

Marginal Relevance (MMR) (Carbonell and Goldstein 1998) generates a diversified ranking list by iteratively select the next best document which has the highest marginal relevance, which is a liner combination of relevance and redundancy. MMR is defined as:

$$MMR \stackrel{def}{=} \arg \max_{D_i \in R \setminus S} \left[\lambda \left(Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j) \right) \right]$$

where $Sim_1(D_i, Q)$ is the similarity between the candidate document D_i and query Q, $Sim_2(D_i, D_j)$ is the similarity between the candidate document and a selected document D_j. Compared to MMR which only considered the similarity and redundancy between documents, eXplicit Query Aspect Diversification (xQuAD) (Santos et al. 2010c) utilized more information from subtopics. The selected document needs to be relevant to the given query and at the same time it needs to cover more novel subtopics. More specifically, xQuAD is defined as:

$$r(d, q, Q(q)) \longleftarrow r(d, q) \times \left(\sum_{q_i \in Q(q)} i_X(q_i, q) r(d, q_i) / m(q_i) \right)^{\omega}$$

where $r(d, q)$ is the relevance score of d with respect to the query q, $i_X(q_i, q)$ is the relative importance of subtopic q_i in terms of query q, $r(d, q_i)$ is the relevance between document d and subtopic q_i, and $m(q_i)$ is the "mass" of information satisfying q_i that is already selected. $m(q_i)$ is updated to account for the selection of a document from all the subtopics it satisfies. TREC Novelty Track (Soboroff 2004) aimed to investigate systems' abilities to locate non-redundant information. Schiffman and McKeown (2004) used both relevant and novel sentences instead of relevant-only ones to minimize redundancy. Yu and Liu (2004) considered both feature relevance and feature redundancy to achieve efficient feature selection. The main focus of the paper is not retrieval models. We take document redundancy into consideration for search result diversity evaluation.

2.2 Diversity measures

To evaluate search result diversification algorithms, a wide range of diversity evaluation measures have been proposed (Clarke et al. 2008; Agrawal et al. 2009; Sakai and Song 2011; Dang and Croft 2012, 2013; Radlinski and Dumais 2006; Dou et al. 2011; Santos et al. 2010a, b, 2011; Zhu et al. 2007). Clarke et al. (2008) proposed α -nDCG. They assume that the number of intents covered by a document determines the graded relevance of that document. Agrawal et al. (2009) proposed Intent-Aware measures. The basic idea is to compute a traditional measure for each intent then sum them up based on the given probabilities of intents. Sakai and Song (2011) proposed *D-measures* which reward documents that are highly relevant to more popular intents. In addition, they proposed *D♯-measures* (Sakai and Song 2011) to visualize the trade-off between relevance and diversity. We briefly introduce the existing measures as follows.

Intent recall: Intent recall (I-rec) is the proportion of intents covered by a ranking list. Let d_r denote the document at rank r, and let $I(d_r)$ denote the set of intents in to which d_r is relevant. The intent recall (*I-rec*) is defined as:

$$I\text{-}rec@K = \frac{|\cup_{r=1}^{K} I(d_r)|}{|\{i\}|}$$

α-nDCG: In order to balance both relevance and diversity of ranked lists, α-nDCG is defined as:

$$\alpha\text{-}nDCG@K = \frac{\sum_{r=1}^{K} NG(r)/\log(r+1)}{\sum_{r=1}^{K} NG^*(r)/\log(r+1)}$$
$$NG(r) = \sum_{i\in\{i\}} J_i(r)(1-\alpha)^{C_i(r-1)}$$

where $NG^*(r)$ is $NG(r)$ in the ideal ranked list; $J_i(r)$ is 1 if the document at rank r is relevant to intent i, and 0 otherwise; $C_i(r) = \sum_{k=1}^{r} J_i(k)$ is the number of relevant documents to intent i within top r; and α is a parameter.

Intent-aware measures: Assuming that M is an ad-hoc retrieval evaluation measure, and $\sum_{i\in\{i\}} P_r(i|q) = 1$, intent-aware measures M-IA is defined as:

$$M\text{-}IA@K = \sum_{i\in\{i\}} P_r(i|q) M_i@K$$

where M_i is the per-intent version of measure M.

D♯-nDCG: Assume that $g_i(r)$ is the gain value of the document at rank r for intent i, and $g_i(r)$ is calculated based on per-intent relevance assessments. Then the global gain at rank r is defined as $GG(r) = \sum_{i\in\{i\}} P_r(i|q) g_i(r)$. Let $GG^*(r)$ denote the global gain at rank r in the ideal ranked list. The ideal list is obtained by listing up all relevant documents in descending order of global gains. D-nDCG is defined as:

$$D\text{-}nDCG@K = \frac{\sum_{r=1}^{K} GG(r)/\log(r+1)}{\sum_{r=1}^{K} GG^*(r)/\log(r+1)}$$

Then D♯-nDCG is defined by:

$$D\sharp\text{-}nDCG@K = \gamma I\text{-}rec@K + (1-\gamma) D\text{-}nDCG@K$$

where γ is a parameter controlling the diversity and relevance.

A common problem with these measures is that they assume that the user needs can be represented as a flat list of intents and that they ignore the relationships between intents. As we discussed in the previous section, this may be insufficient, because intents are not always independent and exclusive.

2.3 Hierarchical diversity measures

Wang et al. (2016) proposed to build superintents over a given set of intents and thereby evaluate search result diversity based on hierarchical intents. Their study showed that their measures are more discriminative and intuitive than traditional measures based on a flat list of intents. These measures are briefly described below.

2.3.1 Layer-aware measures

The key idea of *Layer-Aware Measures* is, for a given q and its intent hierarchy, to evaluate the ranked list based on each layer using existing measures and then combine all scores. Let H

denote the height of the intent hierarchy, and let $L = \{l_1, l_2, \ldots, l_H\}$ denote its first layer to the highest layer. *LA-measures* are defined as follows:

$$M\text{-}LA@K = \sum_{i=1}^{H} w_i * M_i@K \tag{1}$$

where w_i is the weight of layer l_i such that $\sum_{i=1}^{H} w_i = 1$. M_i is the evaluation score of measure M by using intents of layer l_i. For example, *D-nDCG−LA* is computed as follows: (1) Compute an *D-nDCG* score for each layer; (2) Compute a weighted average of the per-layer scores using (1). Therefore, *D-nDCG-LA* is defined as:

$$D\text{-}nDCG\text{-}LA@K = \sum_{i=1}^{H} w_i * D{-}nDCG_i@K \tag{2}$$

where $D\text{-}nDCG_i$ means only using the nodes of layer l_i.

2.3.2 Node recall, *LAD♯-measures*, and *HD♯-measure*

Given a query q, let V denote the nodes in its intent hierarchy except its root. Let d_r denote the document at rank r, and let $N(d_r)$ denote the set of nodes in V to which d_r is relevant. Similar to *I-rec* (Sakai et al. 2010; Zhai et al. 2003), the node recall (*N-rec*) is defined as:

$$N\text{-}rec@K = \frac{|\cup_{r=1}^{K} N(d_r)|}{|V|}$$

$N\text{-}rec@K$ is the proportion of nodes in the hierarchy covered by the top K documents. *N-rec* is a natural generalization of *I-rec* when using the hierarchical intent structures. *I-rec* is a binary-relevance (a document can either be relevant or irrelevant) measure for each intent, and it assumes that each intent is equally important. *N-rec* and *I-rec* are both rank-insensitive and cannot handle graded relevance assessments.

Let D-measure-LA denote the Layer-Aware version of D-measure (Sakai and Song 2011) (e.g., D-nDCG). Then, LAD♯-measure is defined as:

$$LAD\sharp\text{-measure}@K = \gamma N\text{-}rec@K + (1-\gamma) D\text{-measure-}LA@K \tag{3}$$

where γ is a parameter for balancing relevance and diversity, and *D-measure-LA* can be *HD-nDCG-LA*, which is defined in (2). Similarly, *HD♯-measure* is defined as:

$$HD\sharp\text{-}measure@K = \gamma N\text{-}rec@K + (1-\gamma) HD\text{-}measure@K \tag{4}$$

where *HD-measure* can be *HD-nDCG* or *HD-Q*. For example, *HD-nDCG* can be defined as:

$$HD\text{-}nDCG@K = \frac{\sum_{r=1}^{K}[\sum_{i=1}^{H} w_i * GG_i(r)]/\log_2(r+1)}{\sum_{r=1}^{K}[\sum_{i=1}^{H} w_i * GG_i^*(r)]/\log_2(r+1)}$$

where $GG_i(r)$ is the global gain for layer l_i at rank r.

The difference between the two measures is what to combine over layers: *HD-measures* combine the global gain for each layer while *D-measures-LA* combine *D-measures* for each layer.

In contrast to the above measure of Wang et al. that creates *superintents* on top of a given set of intents, our first measure in the present study creates *subintents* under the given intents, and hence removes the assumption that the given intents are atomic. As we shall demonstrate in our experiments, our measure is complementary to the work of Wang et al. and can indeed be combined effectively.

3 Intent hierarchy based evaluation

3.1 Overview of the framework

Assume that we already have a diversity test collection which is comprised of a set of queries and documents. Each query has a list of manually created official intents and each document is judged on whether it is relevant to each intent. This is the common format of diversity of test collections used in TREC Web Track 2009–2013 (Collins-Thompson ety al. 2013) and NTCIR Intent Mining tasks (INTENT and I-Mine) (Yamamoto et al. 2016).

Given a diversity test collection with flat intent lists, our first measure is to build *subintents* under the given intents without additional human efforts, so that we take take into account subtle differences and similarities across documents. Our measure is to automatically create subintents by clustering intent-level relevant documents in a bottom-up fashion.

Figure 3 shows the flow of our algorithm for building an intent hierarchy. Given a set of intents and intent-level relevance assessments for a particular topic, we first perform, for each of the given intents, hierarchical clustering with the relevant documents for that intent. Next, we prune branches, compress layers, and extend nodes in the hierarchy to ensure that it has a desired height. We then compute the importance of each node, and finally combine the trees built for each intent with the official list of intents to form a single intent hierarchy for the given topic.

Finally, we consider abandoning the official TREC intents altogether.The question addressed here is: can we apply our hierarchical subintent measure even in the absence of manually created official intents to start from? To this end, instead of using the intent-level relevance assessments from the diversity task, we started from the topic-level relevance assessments without user intents and built subintent hierarchies using the method discussed above. We regard the idea as our second measure in the paper.

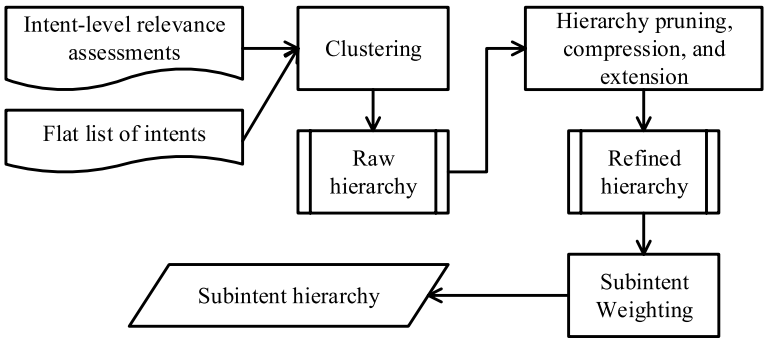

Fig. 3 Overview of our method to building an intent hierarchy with subintents

3.2 Building a raw intent hierarchy

We employ agglomerative clustering to cluster documents for a given *intent*: each document starts as a cluster on its own, and pairs of closest clusters are merged recursively to create the hierarchy. To measure the similarity of two clusters, we consider SimHash (Charikar 2002) and TF-IDF (Salton and McGill 1986). SimHash is an efficient algorithm suitable for handling massive webpage deduplication problems. It maps the original text to a short binary string (fingerprint) which can be computed offline. The similarity of two documents can then be efficiently measured by calculating the Hamming distance (Hamming 1950) of their corresponding binary strings. As for TF-IDF (Salton and McGill 1986), we create TF-IDF word vectors for each document or a document cluster, and employ the cosine similarity. For computing the IDF (inverse document frequency) of each word, we use the statistics from the ClueWeb09 (The clueweb09 dataset 2009) Category B document collection, which contains approximately 50 million web pages. TF-IDF vectors are expected to be more accurate than SimHash, but require more storage and computation costs. For both SimHash and TF-IDF, we use the complete-linkage (i.e., minimum similarity) as the linkage criterion for evaluating whether two clusters should be merged during clustering. While other methods to document clustering would certainly be possible, we leave this question to future work.

Figure 4a shows a raw intent hierarchy created from the fifth intent ("Windows Defender Reports") of Topic 20 ("defender") from the TREC Web Track 2009 diversity task. Here, each leaf node is a single document, as indicated by the circles; it can be observed that different leaf nodes are on different levels in this raw hierarchy. Whereas, internal nodes are shown with SimHash values: for example, the similarity between documents d_1 and d_2 is 0.33.

3.3 Pruning, compression, and extension

The raw hierarchy built for a particular intent as described above often have many layers, with different leaf nodes having different depths. This section describes how we transform the raw hierarchy into the final intent hierarchy that is suitable for diversity evaluation.

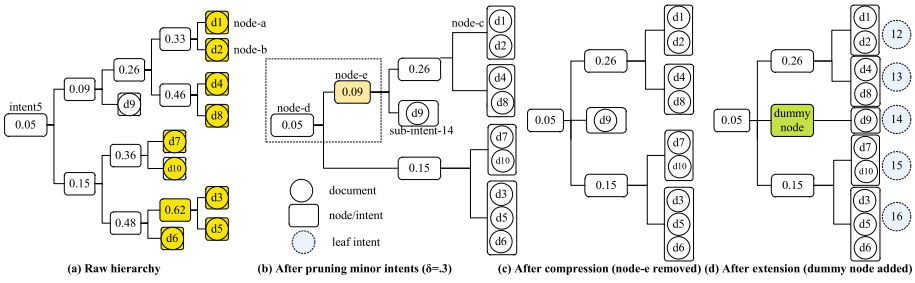

Fig. 4 Generating subintent hierarchy for intent-5 of topic 20 "defender", cutoff $\delta = .3$. The yellow nodes will be merged in the pruning procedure, the orange node will be removed in the compression procedure, and the green node is introduced in the extension procedure

3.3.1 Pruning

First, we perform pruning on the intent tree by removing nodes whose similarity values are larger than a threshold $\delta(0 \leq \delta \leq 1)$. For example, if $\delta = 0.3$, the two documents d_1 and d_2 in Fig. 4a are merged into a single node, as the similarity between them is 0.33. Both of these documents are about "Windows Defender Q&A" and therefore having them both in a search engine result page is in fact somewhat redundant. Figure 4b shows the tree after pruning.

The threshold δ controls the size and granularity of the hierarchy for each given intent. The smaller the δ is, the simpler the intent hierarchy is going to be. In particular, note that when $\delta = 0$, every subintent is merged into one, and therefore our measure reduces to the original flat list intents. We will discuss the effect of δ on our evaluation measures in Sect. 5.3.

3.3.2 Layer compression

We notice that, in many cases, the similarity range between child node and parent node is too small (< 0.1) that there may be unnecessary layers. To deal with this "layer redundancy problem", we compress the hierarchy by requiring that the similarity of a node must not be too similar to that of its parent node. More specifically, we partition the similarity range [0,1] into ten bins, [0, 0.1), [0.1, 0.2], ..., [0.9, 1], and remove the child node if its similarity value is in the same bin as that of the parent node. The parent node then inherits the subtree of the removed node. The above process is repeated for every parent-child pair in the subintent hierarchy until the aforementioned requirement is satisfied. For example, in Fig. 4b, the similarity values for node-e and node-d both lie in the same bin ([0, 0.1]), so node-e is removed, as shown in Fig. 4c. Note that, as a result, a parent node may have more than two children.

3.3.3 Extension

After pruning and layer compression, we perform extension on some of the leaf nodes to ensure that all leaf nodes are on the same level. For this purpose, we follow the measure of Wang et al. proposed in Wang et al. (2016), and introduce dummy internal nodes wherever necessary. For example, in Fig. 4c, the leaf node representing document d_9 is on level 2 while the other leaf nodes are on level 3; hence, we introduce a dummy node on level 2 for this leaf node. Figure 4d shows the result.

3.4 Subintent weighting

The subintents obtained using our first proposed measure can be weighted for the purpose of computing diversity evaluation measures. Specifically, we consider two methods for weighting intents within the hierarchy, as described below.

3.4.1 Weighting by the number of leaf intents (WI)

In this weighting method, we assume that leaf intents are atomic and they are equally important. An intent can be weighted by the percentage of leaf intents it covers. Suppose that we totally have n leaf intents in an intent hierarchy. For an intermediate node (intent) i which has n_i descendent leaf intents (i.e., n_i is the number of leaf nodes within the subtree that has i as the root node), its weight can be calculated by: $\frac{n_i}{n}$. For example, in Fig. 5, there are 3 leaf intents in in the hierarchy in total. Node-a has two distinct leaf intents, and hence its weight is 2/3. Each leaf intent has a uniform weight, namely, 1 / 3. We call this weighting schema *WI*.

3.4.2 Weighting by document gains (WD)

In the above weighting method, we assume that each leaf intent is equally important regardless the number of relevant documents it contains and how relevant the documents are. Alternatively, we can assume that an intent is more important if it covers more relevant documents in the collection. Assume that g is the sum of the global gains (See Sect. 2.3) of all relevant documents within the hierarchy, and g_i is the sum of the global gains of all relevant documents covered by i. Then we can let the weight of intent i be $\frac{g_i}{g}$.

For example, in Fig. 5, the sum of global gains for node-a is 37 while that for the root node is 54, and hence the weight of node-a is $37/54 = 0.69$. We denote this method by *WD*.

3.5 Building the intent hierarchy for a query

After creating an intent hierarchy for each official intent, we merge them to form a single intent hierarchy for the entire query. Just as we introduced some dummy nodes within the

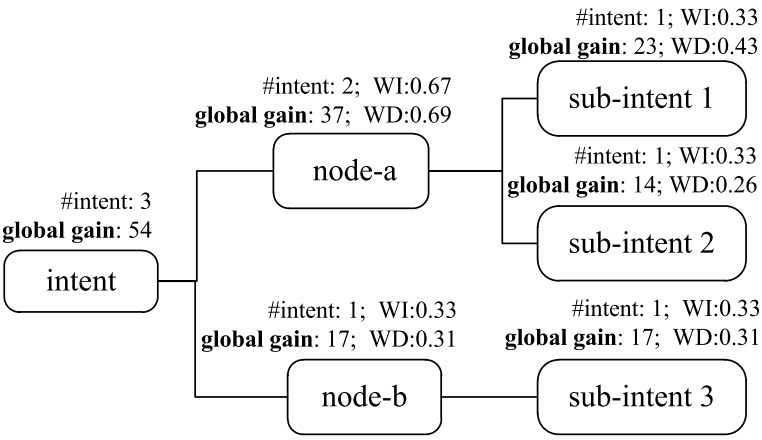

Fig. 5 Weighting subintents

hierarchy for each intent, here we add dummy nodes wherever necessary so that the leaf nodes of the final hierarchy all lie in the same level.

Figure 6a shows an example; because the depth of the hierarchy for intent-1 was three while that for intent-2 was two, a dummy node is introduced for the latter, as shown in Fig. 6b. As for the weights of nodes in Fig. 6b, since the TREC Web Track data does not provide intent probabilities, we assume uniform probabilities for the level-1 intents (intent-1 and -2), so each of the intents receive a 0.5. This weight is then passed on to the child nodes according to how many documents they cover, as shown in the figure.

In the above example, because the original tree depth for intent-2 was one (see Fig. 6a) and it has only one child (i.e., sub-intent 4), the weight assigned to sub-intent 4 in Fig. 6b is as high as 0.500. Hence we also tried an alternative weighting scheme shown in Fig. 6c. In this alternative scheme, we give $2/3 = 0.67$ to intent-1 as the tree depth for this intent is two, and give $1/3 = 0.33$ to intent-2 as the tree depth for this intent is one (see Fig. 6a). The probabilities are then distributed to the children, again according to the number of documents they cover.

3.6 Summary

In summary, our first measure is to build a subintent hierarchy under each official intent, where the complexity of the hierarchy can be controlled by the threshold δ. Given a diversity test collection with intent-level relevance assessments, our measure does not require any additional manual effort whatsoever, while freeing us from the assumption that the official intents are atomic.

As we have described earlier, we build subintents under the given official intents while the measure of Wang et al. proposed in Wang et al. (2016) builds superintents above the official intents, and hence the two are complementary. Hence, in our experiments, we consider combining these two measures. Going back to Fig. 1, given the middle layer (i.e., the official intents), our first measure creates the rightmost layer under; Wang et al. creates the leftmost layer; Fig. 1 in its entirety represents the combined measure.

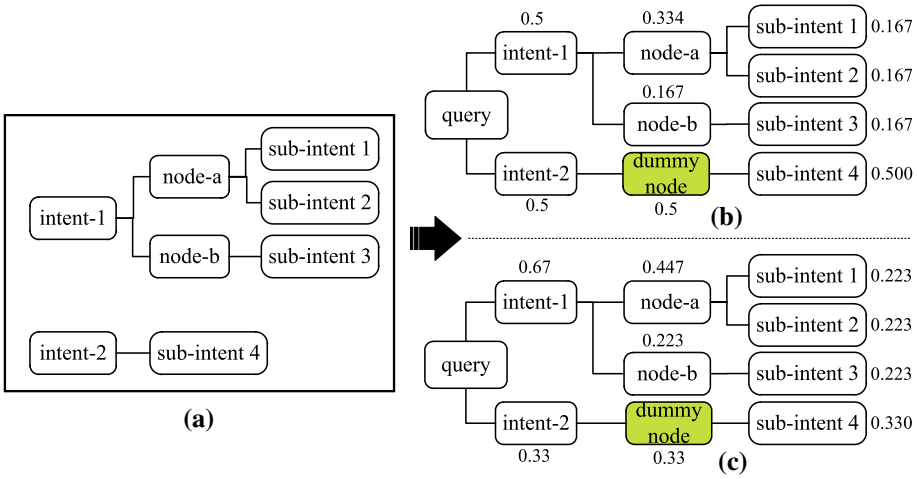

Fig. 6 Creating a query intent hierarchy

Furthermore, in our second measure, we consider a variant of our first measure that clusters per-topic relevance documents rather than per-intent ones. We use a subscript *TL* to identify the measures using topic-level relevance judgments, such as α-*nDCG-LA*$_{TL}$.

4 Document similarity based evaluation

Our third measure for search result diversity evaluation does not require any explicit identifications of intents for a given query. All we need is the set of topic-level relevance assessments for each topic. The assumption behind this new measure is that the overall similarity between relevant documents within the search engine result page directly governs the diversity of the page.

Given a ranked list of size K, we first define the following weighted sum of document similarities:

$$S@K = \frac{\sum_{i,j,i\neq j} w_{ij} \cdot I(i)I(j)sim(d_i, d_j)}{\sum_{i,j,i\neq j} w_{ij}} \tag{5}$$

where $sim(d_i, d_j)$ denotes the SimHash between documents ranked at i and j, w_{ij} is a weight applied to that particular similarity, and $I(i)$ is a flag which returns one if the document at i is relevant and zero otherwise. Our default weighting scheme is as follows:

$$w_{ij} = \frac{K - avg(i,j)}{K} \tag{6}$$

where $avg(i, j)$ is the average of ranks i and j. That is, the similarities for document pairs near the top of the ranking are considered important. Our final evaluation measure is given by:

$$D@K = \frac{1}{2}(M@K + (1 - S@K)) \tag{7}$$

where M is a traditional measure such as nDCG, Q, and ERR. Thus, $D@K$ is an average of a traditional measure and an overall dissimilarity measure.

The above default measure requires similarity computation for every document pair, and weights the similarities based on ranks. This strategy is referred to as *RA* (for Rank-weighted, All pairs). We also experiment with the following variants:

RP Rank-weighted, but consider only *adjacent* relevant document pairs in similarity computation, where the adjacency is defined by ignoring all nonrelevant documents in the top K results.

NA Non-weighted (i.e., $w_{ij} = 0$), consider all relevant document pairs.

NP Non-weighted, consider only adjacent relevant document pairs in similarity computation.

Table 1 Description of ClueWeb09 and ClueWeb12 document collections

	Category A	Category B
ClueWeb09	1 billion documents	50 million documents
ClueWeb12	733 million documents	50 million documents

Table 2 Assessment costs of the proposed evaluation measures

	Topic-level relevance assessment (98,840 labels)	Intent-level relevance assessment (173,069 labels)	Hierarchy structure
Intent hierarchy based evaluation (intent-level)	✓	✓	✓
Intent hierarchy based evaluation (topic-level)	✓		✓
Document similarity based evaluation	✓		

5 Experiments

In this section, we report on several experiments to demonstrate the advantages of our evaluation measures based on subintent hierarchies and document similarities over existing state-of-the-art measures. We describe our experimental setup including the data sets and evaluation metrics in Sect. 5.1. We then report and analyze overall results on rank correlation and discriminative power respectively in Sects. 5.2 and 5.3. We also design a user study on whether our proposed measure can generate more consistent preferences with users than existing measures. The results are reported in Sect. 5.4.

5.1 Experimental setup

Our experiments are all based on the TREC Web Track 2009–2013 diversity test collections (Clarke et al. 2009; Collins-Thompson ety al. 2013) with the ClueWeb09 and ClueWeb12 document collections (The clueweb09 dataset 2009; The clueweb12 dataset 2012). The description of the data sets is shown in Table 1. In this paper, we mainly use Category A in ClueWeb09 and ClueWeb-12. We use 250 topics and 12600 runs to conduct our experiments. The data set we used contains about 100,000 topic-level relevance assessment and 60,000 intent-level relevance assessment. Table 2 shows the assessment and structure costs of our proposed evaluation measures. To compare these evaluation measures, we use rank correlation and discriminative power, which are widely used methods for evaluating evaluation measures.

Rank correlation compares two system rankings. Although rank correlation is often measured by Kendall's τ (Kendall 1938), τ treats exchanges near the top of a ranked list and those near the bottom equally. τ is a monotonic function of the probability that a randomly chosen pair of ranked items is ordered identically in the two rankings; hence a swap near the top of a ranked list and that near the bottom of the same list has equal impact. τ_{ap} (Yilmaz et al. 2008) was proposed to solve this issue. τ_{ap} is "top-heavy", which means it is

a monotonic function of the probability that a randomly chosen item and one ranked above it are ordered identically in the two rankings. Since τ_{ap} is asymmetrical, we use the symmetric τ_{ap}, which can be computed as an average of two τ_{ap} values obtained by swapping the two ranked lists.

Discriminative power (Sakai 2012) represents the stability of measures. It obtains a p-value for every system pair and counts the number of statistically significant differences at a given significance level. It also discusses the Δ, which is an estimate of the minimum between-difference necessary to achieve statistical significance.

However, rank correlation only measures the similarity between measures; it does not show which measure is correct. Discriminative power identifies statistically stable measures, but statistically stable measures do not necessarily align with human perceptions about search results. Therefore, we also conducted user experiments on whether our proposed measures can generate more consistent preferences with users than existing measures.

Following previous work (Clarke et al. 2008; Agrawal et al. 2009; Sakai and Song 2011; Wang et al. 2016), we set use document cutoff at 20 for all intent hierarchy measures and $\gamma = .5$ in Eqs. 3 and 4. Unless stated otherwise, we use SimHash for computing document similarity (see Sect. 3.2). As for the cutoff threshold (See Sect. 3.3), we let $\delta = .3$ for the overall results reported in Sect. 5.3.

5.2 Rank correlation results

Table 3 shows the rank correlation among the measures considered in this study, in terms of τ_{ap}. Because correlation between WI-measures and WD-measures in τ_{ap} is over .900, we only discuss WI-measures here. The following observations can be made from the results.

1. The correlation among flat intent based measures is higher than the correlation between flat intent based measures and hierarchical measures using the subintent hierarchies. For example, the correlation between α-*nDCG* and *ERR-IA* is .870, while the correlation between α-*nDCG* and *HD♯-nDCG*$_{WI}$ is only .726. The correlation between *ERR-IA* and *HD♯-nDCG*$_{WI}$ is even lower (.675). This is reasonable because both α-*nDCG* and *ERR-IA* use flat intents while *HD♯-nDCG*$_{WI}$ uses subintent hierarchies. This means that *HD♯-nDCG*$_{WI}$ can provide evaluation viewpoints that existing measures α-*nDCG* and *ERR-IA* do not cover.

2. Using higher-level intent hierarchies (*SUP*) and using subintent hierarchies (*WI*) lead to different system rankings. When using the same higher-level intent hierarchies, α-*nDCG-LA*$_{SUP}$ and *ERR-IA-LA*$_{SUP}$ is highly correlated (.876) while the correlation between α-*nDCG-LA*$_{SUP}$ and *ERR-IA-LA*$_{WI}$ is relatively lower (.803). This is reasonable because the former hierarchy is based on human judgment while the latter is mostly based on document clustering.

3. The correlation among document similarity based measures is higher than the correlation between document similarity based measures and traditional measures. For example, the correlation between *n-DCG*$_{RA}$ and *n-DCG*$_{NP}$ is .891, while the correlation between *n-DCG*$_{RA}$ and *nDCG* is only .705. The correlation between *n-DCG*$_{NP}$ and *nDCG* is even lower (.660). This is reasonable because document similarity based measures take the similarity of returning documents into consideration and provide extra information, which is helpful to diversity evaluation.

Table 3 Correlation between measures in τ_{ap}

	ERR-IA	D♯-nDCG	HD♯-nDCG$_{WI}$	LAD♯-nDCG$_{WI}$
α-nDCG	.870	.796	.726	.724
ERR-IA	–	.699	.675	.677
D♯-nDCG	–	–	.760	.776
HD♯-nDCG$_{WI}$	–	–	–	.966
	ERR-IA-LA$_{SUP}$	α-nDCG-LA$_{WI}$	ERR-IA-LA$_{WI}$	D♯-nDCG-LA$_{WI}$
α-nDCG-LA$_{SUP}$	.876	.834	.803	.768
ERR-IA-LA$_{SUP}$	–	.851	.815	.702
α-nDCG-LA$_{WI}$	–	–	.796	.743
ERR-IA-LA$_{WI}$	–	–	–	.674
	n-DCG$_{RA}$	n-DCG$_{NP}$	HD♯-nDCG$_{SUP}$	HD♯-nDCG$_{WI}$
n-DCG	.705	.660	.599	.597
n-DCG$_{RA}$	–	.891	.583	.577
n-DCG$_{NP}$	–	–	.549	.534
HD♯-nDCG$_{SUP}$	–	–	–	.915
	α-nDCG	ERR-IA	Q-IA	D♯-nDCG
α-nDCG-LA$_{TL}$	.692	–	–	–
ERR-IA-LA$_{TL}$	–	.751	–	–
Q-IA-LA$_{TL}$	–	–	.718	–
D♯-nDCG-LA$_{TL}$	–	–	–	.700

WI, SUP: topic-level and intent-level relevance assessment, hierarchical intents

TL: topic-level relevance assessment, hierarchical intents

RA, NP: topic-level relevance assessment, no intents

No Subscript: intent-level relevance assessment, flat intents

4. Using intent hierarchies (*SUP* and *WI*) and documents similarities (*RA* and *NP*) lead to correlated but different system rankings. When using intent hierarchies, *HD♯-nDCG$_{SUP}$* and *HD♯-nDCG$_{WI}$* is highly correlated (.915) while the correlation between *HD♯-nDCG$_{WI}$* and *n-DCG$_{NP}$* is relatively lower (.534). It suggests that intent hierarchies and document similarities provide different types of information and reinforce diversity evaluation from different viewpoints.

5. The correlation between traditional measures using human created intents and their corresponding hierarchical intents created solely based on per-topic judgments is above 0.69, which indicates a relatively high correlation. This means that, with a reasonable accuracy, we can conduct diversity evaluation without using the official intents. We can just start from the per-topic relevance assessments and build hierarchical intents in a bottom up manner.

5.3 Discriminative power results

We measure discriminative power by conducting a statistical significance test for different pairs of runs, and counting the number of significantly different pairs. Following previous work (Sakai 2012, 2006a, b; Sakai and Robertson 2008), we adopt the paired bootstrap test to compute discriminative power. For significance testing, we use the two-tailed paired bootstrap test at the significance level of $\alpha = 0.05$ and set $B = 1000$ (B is the number of bootstrap samples).

Note that discriminative power is not about whether the measures are right or wrong; it is about how measures can be consistent across experiments and as a result how often differences between systems can be detected with high confidence. We regard high discriminative power as a necessary condition for a good evaluation measure, but not as a sufficient condition. The discriminative power method we adopted also provides a natural estimate of the performance difference (Δ) between two systems required to achieve statistical significance. This is done by recording, for every run pair, the Δ that corresponds to the borderline between significance and nonsignificance among the 1,000 trials, and then by selecting the largest value among all run pairs. We sample 20 submitted runs from every year, which produces $5 * 20 * (20 - 1)/2 = 950$ pairs of sampled runs in total. With the 950 pairs of sampled runs, we compute the discriminative power and performance Δ using all 250 queries in TREC 2009–2013 diversity test collections.

The discriminative power results are shown in Table 4. We experimented with the traditional measures using flat intents (such as *D♯-nDCG*), their corresponding hierarchical measures proposed by Wang et al. (2016) (introduced in Sect. 2.3, such as *D♯-nDCG-LA*, *HD♯-nDCG*, and *LAD♯-nDCG*) using the superintents (denoted with *SUP* in the column header), or using the subintents proposed in this paper (denoted with *WI* and *WD*, representing for different weighting methods described in Sect. 3.4). We further experimented with the combination of both types of hierarchical intents (denoted with *SUP* + *WI* and *SUP* + *WD*). Meanwhile, we also made experiments on topic-level intent hierarchy based measures (denoted with *TL*), such as α-*nDCG-LA*$_{TL}$. In addition, we examined the traditional *nDCG* without intents and our document similarity based measures (*RA*, *RP*, *NA* and *NP*). From the table we find that:

1. Hierarchical measures using subintent hierarchies (*WI*, *WD*) are at least as discriminative as the corresponding flat-list measures. For example, no matter which kind of weighting method is used (either *WI* or *WD*), hierarchical measures *ERR-IA-LA*$_{WI}$ (522) and *ERR-IA-LA*$_{WD}$ (522) outperform their corresponding measure *ERR-IA* (518). Similarly, both *HD♯-nDCG*$_{WI}$ (574) and *HD♯-nDCG*$_{WD}$ (573) outperform *D♯-nDCG* (557). Using subintents help describe minor differences between intents covered by documents, and hence is able to better identify diversity difference between ranking systems. This means that our methods for automatically creating subintents, which requires no extra human efforts, is useful in evaluating search result diversity.

2. Hierarchical measures using subintent hierarchies (*WI*, *WD*) are at least as discriminative as the corresponding measures using higher-level intents (*SUP*). For example, *LAD♯-nDCG*$_{WI}$ (573) outperforms *LAD♯-nDCG*$_{SUP}$ (560), which means building subintents under official intents can achieve a higher discriminative power than building higher-level intents. Note that creating superintents requires some extra human effort (Wang et al. 2016), while no additional human effort is required in our proposed measure. This suggests that when we want to apply hierarchical measures, we can first consider the use of the hierarchies proposed in this paper.

Table 4 Discriminative power of measures

Existing measures			*SUP*		*WI*		*WD*		*SUP* + *WI*		*SUP* + *WD*	
Measure	Disc.	*Δ*	Disc	*Δ*	Disc.	*Δ*	Disc.	*Δ*	Disc.	*Δ*	Disc.	*Δ*
LA-measures (*I-rec-LA*, *α-nDCG-LA*, etc)												
I-rec	499	0.13	500	0.13	551	0.12	551	0.12	552	0.12	551	0.12
α-nDCG	573	0.11	565	0.11	573	0.09	573	0.09	565	0.10	565	0.10
ERR-IA	518	0.12	521	0.12	522	0.08	522	0.08	524	0.08	524	0.08
Q-IA	459	0.07	464	0.07	473	0.04	465	0.05	483	0.05	480	0.05
D♯-nDCG	557	0.09	559	0.09	574	0.09	573	0.09	574	0.09	575	0.09
D♯-Q	546	0.09	553	0.09	567	0.09	568	0.09	565	0.09	568	0.09
HD-measures (*HD♯−nDCG* and *HD♯−Q*)												
D♯-nDCG	557	0.09	560	0.09	574	0.09	573	0.09	574	0.09	573	0.09
D♯-Q	546	0.09	554	0.09	565	0.09	565	0.09	566	0.09	566	0.09
LAD-measures (*LAD♯−nDCG* and *LAD♯−Q*)												
D♯-nDCG	557	0.09	560	0.09	573	0.09	573	0.09	574	0.09	574	0.09
D♯-Q	546	0.09	554	0.09	565	0.09	566	0.09	566	0.09	567	0.09

Existing measures			*TL*		
Measure	Disc.	*Δ*	Measure	Disc.	*Δ*
α-nDCG	573	0.11	α-nDCG-LA	547	0.10
ERR-IA	518	0.12	ERR-IA-LA	492	0.09
Q-IA	459	0.07	Q-IA-LA	460	0.11
D♯-nDCG	557	0.09	D♯-nDCG-LA	541	0.08

Table 4 (continued)

Existing measures			*RA*		*RP*		*NA*		*NP*	
Measure	Disc.	Δ	Disc.	Δ	Disc.	Δ	Disc.	Δ	Disc.	Δ
nDCG	515	0.12	520	0.09	551	0.09	515	0.09	555	0.10

The top part shows results of intent-level intent hierarchy based measures, the leftmost column shows existing measures' results; the next column shows hierarchical measures' results using *SUP*; the right columns show results using different subintent hierarchies (cutoff $\delta = .3$). The middle part shows results of topic-level intent hierarchy based measures. The bottom part shows results of document similarity based measures

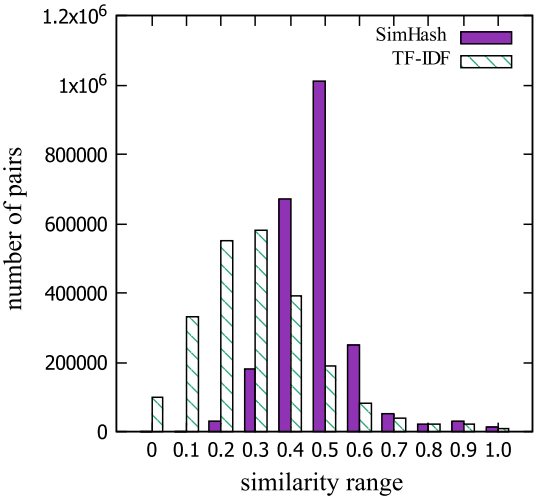

Fig. 7 Statistics about document similarity of ClueWeb09 (1 billion documents), ClueWeb12 (733 million documents) document collections

3. Combining subintents and superintents ($SUP + WI$ and $SUP + WD$) achieves the highest discriminative power for most measures. For example, hierarchical measures *Q-IA-LA*$_{SUP+WI}$ (483) and *Q-IA-LA*$_{SUP+WD}$ (480) outperform their corresponding flat measure *Q-IA* (459) with a more than 20 improvement in terms of discriminative power. This means that a combination of the subintents and superintents is beneficial. Creating higher level of intents can help identify the semantic relationship between human created intents, whereas subintents are useful to identifying subtle difference between rank lists.

4. The hierarchical measures (e.g., *ERR-IA-LA*$_{TL}$, 492) tend to be slightly less discriminative than the corresponding official measures (e.g., *ERR-IA*, 518). This suggests that the official intents created at TREC help traditional diversity evaluation measures achieve high discriminative power. This is probably because, while our measures based on the topic-level relevance assessments are based only on the documents contributed to the pools by the participating systems, the official intents may represent knowledge that goes beyond the pool of retrieved documents obtained for each query, namely, the human knowledge about the query itself. Moreover, our measures for computing the similarity between documents (i.e., SimHash and TF-IDF) are relatively crude: more sophisticated measures may help us identify subintents more accurately.

5. Document similarity based evaluation measures (*RA*, *RP*, *NA* and *NP*) are almost at least as discriminative as their traditional measures. For example, *RA* (520) and *NP* (555) outperform their corresponding measure *nDCG* (515). It indicates that using the similarities among returned documents can help detect the document redundancy problem and thereby help identifying the subtle difference between different systems.

6. Document similarity based evaluation measures (*RA*, *RP*, *NA* and *NP*) are almost as same discriminative as the corresponding measures using hierarchical intents. For instance, the difference in terms of the number of statistically significant differences between *HD*♯*−nDCG*$_{SUP}$ (560) and *NP* (555) is 5, which means that creating hierarchical intents and focusing on similarity between documents can provide us with fine-grained information in different viewpoints, both of them are helpful to reinforce diversity evaluation.

To examine the impact of cutoff threshold δ (Sect. 3.3) on discriminative power, we varied δ from 0 to 1.0. Figure 7 shows the distribution of the SimHash similarity for

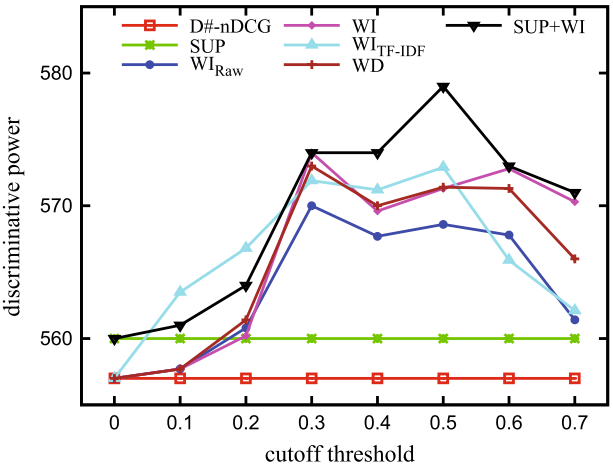

Fig. 8 Experiments with cutoff thresholds δ in *HD♯-nDCG*

every pair of relevant documents from the ClueWeb09, 12 document collections; the distribution for the TF-IDF similarity is also shown. It can be observed that most of the SimHash similarities lie in the 0.4-0.5 range and that only a small number of document pairs have similarities higher than 0.7. This means that a cutoff $\delta > .7$ will not effectively prune the raw sub-intent hierarchy and will only introduce noises to our results. For this reason, we only experiment with $\delta < .7$. The same goes for TF-IDF.

Take *D♯−nDCG* and its corresponding hierarchical measure *HD♯−nDCG* for example: their discriminative power results with different δ are shown in Fig. 7. Different curves represent different intent hierarchies as described earlier. WI_{Raw} denotes the raw subintent weighted hierarchy without layer compression and layer weight adjustment (see Figs. 4b, 6b). While we use SimHash for similarity calculation by default, the figure also shows TF-IDF result for *WI* (denoted as WI_{TF-IDF}).

Figure 8 shows that:

1. When $\delta = 0$, the whole subintent hierarchy reduces to original flat intent list. Therefore *WD*, *WI*, WI_{Raw}, and $WI_{TF\text{-}IDF}$ all reduce to flat intent lists, while $SUP + WI$ reduces to *SUP*.
2. By comparing WI_{Raw} and *WI*, we find that our proposed solution for layer compression and layer weight adjustment improves discriminative power.
3. When $\delta = .3$, *WI* and *WD* perform well; when $\delta = .5$, $SUP + WI$ achieves the highest discriminative power. This further confirms that combining subintent and superintent hierarchies is beneficial.
4. *WI* and $WI_{TF\text{-}IDF}$ using different text similarity algorithms (SimHash and *TF-IDF*) have similar performance tendency. The latter reaches a peak when $\delta = .5$ which is different from the former, because their document similarity distributions are different.

Although not shown in the figure, similar observations apply to other diversity measures, not just *HD♯−nDCG*.

5.4 Agreement with user preferences

In addition to examining the measures in terms of discriminative power and rank correlation, we conduct a user preference test to investigate the agreement between the measures and human preferences given two ranked lists, since whether the measures are measuring what we want to measure is arguably the most important question.

Our user preference agreement experiments were conducted as follows. First, 50 queries were randomly chosen from the 250 TREC 2009–2013 Web Track topics. Then, for each query, we formed two separate sets of ranked list pairs using official TREC runs: the first set contains five randomly chosen system pairs, while the second contains five system pairs randomly chosen from those for which a traditional measure (*D♯-nDCG*) and our measure (*HD♯-nDCG*) disagreed. Hence, in total, we have 250 randomly chosen ranked list pairs plus 250 for which the two measures disagreed.

To collect user preferences for the above ranked list pairs, we designed a web interface that displays each pair side by side, and lets a participant choose from the Left, Equal, and Right buttons shown at the bottom. The top of the interface showed the description of the topic and an instruction saying that the search result that is more relevant and diverse should be chosen. We removed nonrelevant documents from the original ranked lists and then showed only the top 10 documents, so that the participant can focus on the question of diversity versus redundancy rather than the degree of relevance of each document. The interface allowed participants to click on a document to visit that page.

We hired eight participants who are non-native English speakers but are proficient in reading and understanding English. Each participant was assigned five sessions, where a session contains randomly 50 system pairs, and completed the work in about 250 min (i.e., about 1 min per system pair). Each of them was given two days to complete the work, and was required to take at least a 30-min break between sessions. We thereby collected $8 * 250 = 2000$ preference judgments, four for each system pair.

An evaluation measure and a participant independently say either "System1 > System2," "System1 < System2," or "System1 = System2." To quantify the agreement between the two, we also use Kendall's τ, by counting the number of agreements and disagreements instead of swaps in a ranking. The results are shown in Tables 5, 6 and 7.

Table 5 User preference agreement values in τ

Pool	Meas.	*HD♯*	User-1	User-2	User-3	User-4	avg
1–500 all pairs	*D♯*	− 0.100	.200	.184	.188	.228	.200
	HD♯	–	.632	.528	.528	.528	.554
1–250 random	*D♯*	0.800	.640	.560	.552	.584	.584
	HD♯	–	.752	.656	.632	.680	.680
251–500 disagreed	*D♯*	− 1.000	− .240	− .192	− .176	− .128	− .184
	HD♯	–	.512	.400	.424	.376	.428
Participants	User-1	–	–	.608	.692	.604	.651
	User-2	–	–	–	.668	.644	
	User-3	–	–	–	–	.692	

D♯ means D♯-nDCG; HD♯ means HD♯-nDCG$_{WI}$

Table 6 User preference agreement results of document similarity based measures

Pool	Meas.	User-1	User-2	User-3	User-4	Avg
1–500 all pairs	*nDCG*	− .096	− .088	− .064	− .108	− .089
	RA	.408	.452	.464	.464	.447
	RP	.432	.492	.468	.472	.466
	NA	.412	.456	.460	.460	.466
	NP	.428	.488	.472	.476	.466
1–250 random	*nDCG*	.352	.448	.440	.360	.400
	RA	.384	.456	.464	.448	.438
	RP	.416	.472	.480	.496	.466
	NA	.416	.472	.480	.464	.458
	NP	.440	.480	.488	.504	.478
251–500 disagreed	*nDCG*	− .544	− .624	− .568	− .576	− .578
	RA	.416	.496	.456	.448	.454
	RP	.408	.440	.440	.456	.436
	NA	.448	.512	.456	.448	.466
	NP	.432	.448	.464	.480	.456
Participants	User-1	–	.816	.780	.792	.808
	User-2	–	–	.808	.832	
	User-3	–	–	–	.820	

RA means nDCG_{RA}, and so on

Table 7 Agreement with user preference of measures with official intents and topic-level judgment based hierarchical intents

Correlation with user preference				
Measure	τ	Measure	τ	Pool
D♯-nDCG	.200	*D♯-nDCG-LA*$_{TL}$	.361	1–500
	.584		.616	1–250
	− .184		.106	251–500

First, it can be observed that the inter-participant agreement is reasonably high ($\tau > .6$), suggesting that our data is reliable. As for the agreement between a measure and a participant, we find that:

1. *HD♯-nDCG* consistently and substantially outperforms *D♯-nDCG* in terms of preference agreement. That is, regardless of who the participant is, *HD♯-nDCG*'s preference is more similar to him/her than that of *D♯-nDCG*. For example, *HD♯-nDCG*$_{WI}$ ($\tau = .554$) using subintents is more intuitive than *D♯-nDCG* ($\tau = .200$) using flat intents when considering all 500 system pairs.

2. The superiority of *HD♯-nDCG* over *D♯-nDCG* is striking especially for the second set of ranked list pairs, for which these two measures disagree. For *D♯-nDCG*, the agreement in terms of τ is actually negative, which means that there are more disagreements with the participants than there are agreements. In short, when the two measures disagree, the final verdict by the user is often "*HD♯-nDCG* is right."

3. Comparing to using flat human created intents, the hierarchical measure without the official intents (created solely based on topic-level judgments) is more highly

Table 8 User preference example

Runs	Rank-8	Rank-9	Rank-10	$D\sharp$	$HD\sharp$	User
Run-1	d_1 $L1\{i_1, i_2\}$ $L2\{i_{1a}, i_{2a}\}$	d_2 $L1\{i_2\}$ $L2\{i_{2b}\}$	d_3 $L1\{i_3\}$	=	⇑	⇑
Run-2		d_3 $L1\{i_3\}$	d_4 $L1\{i_1\}$ $L2\{i_{1a}\}$			

D♯ means D♯-nDCG; HD♯ means HD♯-nDCG$_{WI}$

correlated with user preference. This suggests that the results based on the official intents are by no means the gold standard of user satisfaction: indeed, it is known that replacing the intent sets for the same topic set may substantially affect the diversified system evaluation results (Sakai et al. 2013).

4. Document similarity based evaluation clearly outperforms traditional *nDCG* in terms of preference agreement. For traditional *nDCG*, the agreement in terms of τ is actually negative when considering all 500 system pairs or 250 disagreed pairs. It is reasonable because measuring the similarities among the returned documents can quantify document redundancy and add diversity information to traditional measures.

5. *D♯−nDCG* using hierarchical subintents outperforms document similarity based evaluation considering preference agreement. It shows that information from user intents better reflects participants' views than information from document similarity does. However, note that document similarity based evaluation can achieve relatively good results with low annotation cost.

6. Our different document similarity-based measures achieve similar results in terms of preference agreement. We find that rank weight and document pair selection have little impact on the agreement with users.

Table 8 shows an actual ranked list pair (Run-1 is UAmsAnc05LS and Run-2 is UAmsM705FLS) from our experiment, where *D♯−nDCG* and *HD♯−nDCG* disagreed, and *all of our four participants* agreed with *HD♯−nDCG*. The topic is "map of Brazil" (Topic 110 from the TREC 2011 Web Track), which has three official intents: i_1 ("What are the boundaries of the political jurisdictions in Brazil?"), i_2 ("I am looking for information about taking a vacation trip to Brazil"), i_3 ("I want to buy a road map of Brazil"). As the table indicates, the two runs have the same top eight results, with document d_1 at rank 8, but Run-1 returned d_2, d_3 at ranks 9, 10, while Run-2 returned d_3, d_4 at ranks 9, 10. Our subintents covered by these documents are shown as i_{1a}, i_{2a}, i_{2b}. In terms of the official flat-list intents, it can be observed that both runs cover i_1, i_2, i_3, and that the per-intent relevance level is $L1$ ("regular relevant") in every case. Hence, *D♯−nDCG* considers these two runs to be ties. Whereas, in terms of our subintents, Run-1 covers i_{1a}, i_{2a}, i_{2b}, while Run-2 covers only i_{1a}, i_{2a}. That is, at the subintent level, d_4 is redundant, and therefore *HD♯−nDCG* prefers Run-1 over Run-2, just like our four participants did.

6 Discussion

In this paper, we propose three low-cost evaluation measures for search result diversification. In order to observe subtle differences between the official intents, we create a method to generate minor intent hierarchy by clustering relevant documents. All the proposed measures are based on document similarity and avoid extra manual annotation cost.

There is a remained problem that our proposed measures tend to favor those diversification models whose principles are similar to our evaluation measures. The human evaluation of search result diversification requires a large amount of annotation, including creating query intents, annotating relevance between documents and each intent. This is usually very costly, especially when the intent is a hierarchy. The motivation of the paper is to reduce the cost via some automatic methods or improve the evaluation quality by considering more information in addition to the human labels. This can at least be used as a preliminary analysis before a large amount of human annotation is created. In the future, we plan to improve the metric and make it more general.

7 Conclusions

Most of the existing diversity measures are based on a flat list of predefined intents for each topic. Inspired by the work of Wang et al. that creates superintents over the official intents, we propose a new diversity evaluation measure based on hierarchical intents, which creates *subintents* beneath the official intents. This measure applies hierarchical clustering to intent-level relevant documents provided in a standard diversity test collection with flat intent lists. While the above proposed measure relied on intent-level relevance assessments, we also propose a second measure that replaces the intent-level relevance assessments with the topic-level relevance assessments to completely automatically form an intent hierarchy for a given topic. Furthermore, our third measure solely relies on the similarity between *topic-level* relevant documents.

We evaluate our measures on the TREC Web Track 2009–2013 diversity test collections. The results show that our first measure achieves higher discriminative power than flat-intents measures and Wang et al.'s superintent-based hierarchies measures. Moreover, the combination of superintents and subintents achieves the highest discriminative power. Furthermore, our first measure performs well even when we abandon the per-intent relevance assessments and build hierarchical subintents from topic-level relevance documents. It confirms the finding of Wang et al. (2016) that hierarchical intents could improve the performance of diversity evaluation. Our third measure based on document similarity also outperforms traditional measures in terms of discriminative power, which confirms the finding of Carbonell and Goldstein (1998) and Santos et al. (2010c) that document relevance and redundancy can observe novel information between documents and are beneficial to diversification evaluation. More importantly, according to our user preference agreement evaluation, our measures outperform traditional measures.

The measures we proposed are all based on document similarity and avoid extra manual annotation cost. However, our evaluation measures will be biased to those diversification retrieval models which focus on document similarity and hierarchical intents. Our motivation is to improve the diversification evaluation quality with fewer human annotations by building richer structure automatically or getting more information from documents

directly. Our results suggest that it may indeed be possible to evaluate search result diversification without manually constructing intents and collecting intent-level relevance assessments. These measures are highly practical and deserve further studies, as they require no extra cost beyond what is already required in traditional ad-hoc information retrieval evaluation.

Acknowledgements Zhicheng Dou is the corresponding author. This work was funded by the National Natural Science Foundation of China under Grant No. 61872370, and National Key R&D Program of China No. 2018YFC0830703.

References

Agrawal, R., Gollapudi, S., Halverson, A., & Ieong, S. (2009). Diversifying search results. In *Proceedings of WSDM '09* (pp. 5–14).

Carbonell, J., & Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 335–336). ACM.

Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. In *STOC '02* (pp. 380–388). ACM.

Clarke, C. L., Craswell, N., & Soboroff, I. (2009). *Overview of the trec 2009 web track*. DTIC Document, Technical Report.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., & MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR '08* (pp. 659–666).

Collins-Thompson, K., Bennett, P., Diaz, F., Clarke, C., & Voorhees, E. (2013). Overview of the trec 2013 web track. In *TREC 2013*.

Dang, V., & Croft, B. W. (2013). Term level search result diversification. In *Proceedings of SIGIR '13* (pp. 603–612). ACM

Dang, V., & Croft, W. B. (2012). Diversity by proportionality: An election-based approach to search result diversification. In *Proceedings of SIGIR '12* (pp. 65–74).

Dou, Z., Hu, S., Chen, K., Song, R., & Wen, J. R. (2011). Multi-dimensional search result diversification. In *Proceedings of WSDM '11* (pp. 475–484).

Dou, Z., Song, R., & Wen, J. (2007). A large-scale evaluation and analysis of personalized search strategies. In C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, & P. J. Shenoy (Eds.) *Proceedings of the 16th international conference on world wide web, WWW 2007, Banff, Alberta, Canada, May 8–12, 2007* (pp. 581–590). ACM. https://doi.org/10.1145/1242572.1242651.

Dou, Z., Song, R., Wen, J., & Yuan, X. (2009). Evaluating the effectiveness of personalized web search. *IEEE Transactions on Knowledge and Data Engineering*, *21*(8), 1178–1190. https://doi.org/10.1109/TKDE.2008.172.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal*, *29*(2), 147–160. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x.

Järvelin, K., & Kekäläinen, J. (2000). Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of SIGIR '00* (pp. 41–48). ACM.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93.

Radlinski, F., & Dumais, S. (2006). Improving personalized web search using result diversification. In *Proceedings of SIGIR '06* (pp. 691–692).

Sakai, T. (2006a). Bootstrap-based comparisons of ir metrics for finding one relevant document. In *Proceedings of AIRS '06* (pp. 374–389). Springer.

Sakai, T. (2006b). Evaluating evaluation metrics based on the bootstrap. In *Proceedings of SIGIR '06* (pp. 525–532). ACM.

Sakai, T. (2012). Evaluation with informational and navigational intents. In *Proceedings of WWW '12* (pp. 499–508). ACM.

Sakai, T., Craswell, N., Song, R., Robertson, S., Dou, Z., & Lin, C. Y. (2010). Simple evaluation metrics for diversified search results. In *Proceedings of EVIA '10* (pp. 42–50).

Sakai, T., Dou, Z., & Clarke, C. L. (2013). The impact of intent selection on diversified search evaluation. In *Proceedings of SIGIR '13, SIGIR '13* (pp. 921–924). ACM, New York, NY, USA. https://doi.org/10.1145/2484028.2484105.

Sakai, T., & Robertson, S. (2008). Modelling a user population for designing information retrieval metrics. In *Proceedings of EVIA '08*.
Sakai, T., & Song, R. (2011). Evaluating diversified search results using per-intent graded relevance. In *Proceedings of SIGIR '11* (pp. 1043–1052).
Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*.
Santos, R. L., Macdonald, C., & Ounis, I. (2010a). Exploiting query reformulations for web search result diversification. In *Proceedings of WWW '10* (pp. 881–890).
Santos, R. L., Macdonald, C., & Ounis, I. (2010b). Selectively diversifying web search results. In *Proceedings of the 19th ACM international conference on information and knowledge management* (pp. 1179–1188). ACM.
Santos, R. L., Macdonald, C., & Ounis, I. (2011). Intent-aware search result diversification. In *Proceedings of SIGIR '11* (pp. 595–604). ACM.
Santos, R. L., Peng, J., Macdonald, C., & Ounis, I. (2010). Explicit search result diversification through sub-queries. In *European conference on information retrieval* (pp. 87–99). Springer.
Schiffman, B., & McKeown, K. (2004). Columbia university in the novelty track at trec 2004. In *TREC*.
Soboroff, I. (2004). Overview of the TREC 2004 novelty track. In E. M. Voorhees, & L. P. Buckland (Eds.) *Proceedings of the thirteenth text retrieval conference, TREC 2004, Gaithersburg, Maryland, USA, November 16–19, 2004, vol. Special Publication 500-261*. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec13/papers/NOVELTY.OVERVIEW.pdf.
Song, R., Dou, Z., Hon, H., & Yu, Y. (2010). Learning query ambiguity models by using search logs. *Journal of Computer Science and Technology*, *25*(4), 728–738. https://doi.org/10.1007/s11390-010-9360-y.
The clueweb09 dataset. (2009). http://boston.lti.cs.cmu.edu/Data/clueweb09/.
The clueweb12 dataset. (2012). http://boston.lti.cs.cmu.edu/Data/clueweb12/.
Wang, X., Dou, Z., Sakai, T., & Wen, J. R. (2016). Evaluating search result diversity using intent hierarchies. In *Proceedings of SIGIR '16*.
Yamamoto, T., Liu, Y., Zhang, M., Dou, Z., Zhou, K., Markov, I., et al. (2016). Overview of the ntcir-12 imine-2 task. *NTCIR*, *2016*, 94–123.
Yilmaz, E., Aslam, J. A., & Robertson, S. (2008). A new rank correlation coefficient for information retrieval. In *Proceedings of SIGIR '08* (pp. 587–594). ACM.
Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, *5*(Oct), 1205–1224.
Zhai, C. X., Cohen, W. W., & Lafferty, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of SIGIR '03* (pp. 10–17). ACM.
Zhu, X., Goldberg, A. B., Van Gael, J., & Andrzejewski, D. (2007). Improving diversity in ranking using absorbing random walks. In *Proceedings of HLT-NAACL '07* (pp. 97–104).

Affiliations

Zhicheng Dou[1,2] · Xue Yang[1,2] · Diya Li[3] · Ji-Rong Wen[1,2] · Tetsuya Sakai[4]

Xue Yang
ruc_yangx@ruc.edu.cn

Diya Li
lid18@rpi.edu

Ji-Rong Wen
jrwen@ruc.edu.cn

Tetsuya Sakai
tetsuyasakai@acm.org

[1] State Key Laboratory of Software Development Environment, School of Information, Renmin University of China, Beijing, People's Republic of China

[2] Beijing Key Laboratory of Big Data Management and Analysis Methods, National Engineering

Laboratory of Big Data System Software (Beijing Institute of Technology), Beijing, People's Republic of China

[3] Rensselaer Polytechnic Institute, New York, USA

[4] Waseda University, Tokyo, Japan