Search Result Diversity Evaluation based on Intent Hierarchies

Xiaojie Wang, Ji-Rong Wen, Zhicheng Dou, Tetsuya Sakai, and Rui Zhang.

Abstract—Search result diversification aims at returning diversified document lists to cover different user intents of a query. Existing diversity measures assume that the intents of a query are disjoint, and do not consider their relationships. In this paper, we introduce intent hierarchies to model the relationships between intents, and present four weighing schemes. Based on intent hierarchies, we propose several hierarchical measures that take into account the relationships between intents. We demonstrate the feasibility of hierarchical measures by using a new test collection based on TREC Web Track 2009-2013 diversity test collections and by using NTCIR-11 IMine test collection. Our main experimental findings are: (1) Hierarchical measures are more discriminative and intuitive than existing measures. In terms of intuitiveness, it is preferable for hierarchical measures to use the whole intent hierarchies than to use only the leaf nodes; (2) The types of intent hierarchies used affect the discriminative power and intuitiveness of hierarchical measures. We suggest the best type of intent hierarchies to be used according to whether the nonuniform weights are available; (3) To measure the benefits of the diversification algorithms which use automatically mined hierarchical intents, it is important to use hierarchical measures instead of existing measures.

Index Terms—Ambiguity, Diversity, Evaluation, Novelty, Hierarchy.

1 Introduction

Nowadays, People tend to meet their daily information needs by issuing keywords into search engines. However, these keywords, i.e. queries, are often ambiguous or broad [1], [2], [3], [4]. The queries usually have several interpretations or aspects, also known as subtopics or user intents. When users submit the same query to retrieval systems, they may want different information returned to fulfill their own information needs. This poses a challenge to search engines when the user intent cannot be known in advance.

To tackle this problem, a wide range of search result diversification algorithms ([5], [6], [7], [8], [9], [10], [11], [12], [13], [14]) have been proposed over the past years. They aim at returning a diversified ranked document list that covers different intents of a query. In the meantime, some researchers have introduced a variety of diversity measures, such as I-rec [15], α-nDCG [16], Intent-Aware measures [5], D♯-measures [17], etc. These measures evaluate ranked lists in terms of both diversity and relevance, and can be used to indicate which diversification algorithms are better. Existing diversity measures assume that the users' information needs could be represented by a single layer of intents and different types of intents are independent of each other. However, intents can be related to each other in reality, which is illustrated as follows.

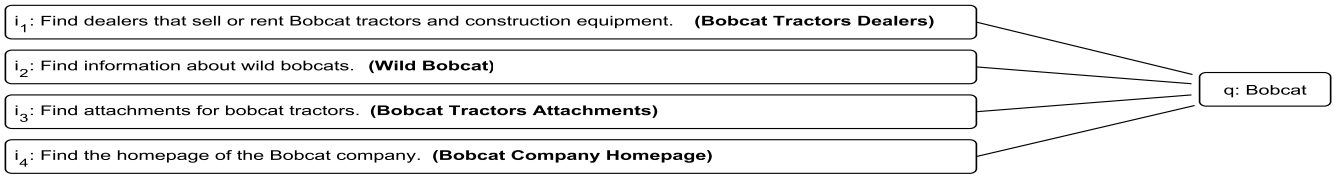
We use the query "bobcat", No. 77 topic in Text Retrieval Conference(TREC) 2010 Web Track [18], as an example. This query is ambiguous because of the polysemy of "bobcat": one interpretation is a company called "bobcat company" whose core business is about tractors; another interpretation is a kind of wild animals called "wild bobcat." We show its official intents, marked by i_1-i_4, in Figure 1(a). The figure shows that except intent i_2 which is about "wild bobcat," the remaining ones, i_1, i_3, and i_4, are all about "bobcat company." This indicates that i_1, i_3, and i_4 are more related to each other, but are less related to i_2. Even within the three intents about "bobcat company," i_1 and i_3 are closer because they are about the businesses involving tractors of the company, whereas i_4 is about homepage the company. We argue that this kind of relationships between intents should be modeled when evaluating search result diversity. However, none of existing measures considers this.

Specifically, we find two submitted runs for the query, cmuFuTop10D and THUIR10DvNov, in TREC Web Track 2010 diversity task. cmuFuTop10D covers i_1, i_3, and i_4, while THUIR10DvNov covers i_1, i_2, and i_4 in their top ten documents. Since i_1, i_3, and i_4 are all about "bobcat company," cmuFuTop10D misses another interpretation of bobcat, i.e. "wild bobcat," but THUIR10DvNov covers both interpretations. In this sense, the latter is more diversified but I-rec [15] treats them as equally good because they cover the same number of intents. Some other existing measures also have similar problems, which will be illustrated in Section 3.4.2. We think that this is due to their lack of recognition of the relationships among intents.

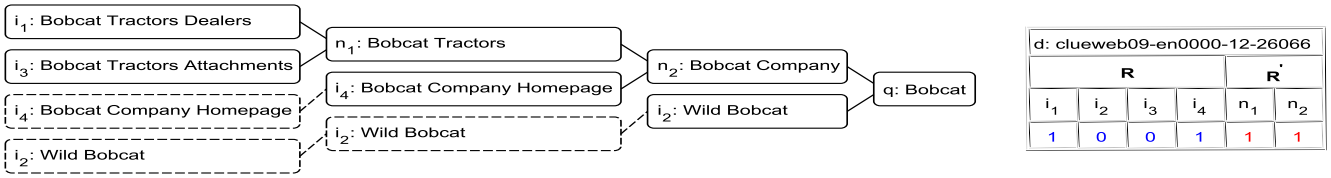
In light of the above observation, we introduce intent hierarchies to model the relationships among intents. We design hierarchical measures using the intent hierarchies

- *Xiaojie Wang is with the School of Information at Renmin University of China, and the School of Computing and Information Systems at The University of Melbourne. E-mail: xiaojiew1@student.unimelb.edu.au*
- *Zhicheng Dou and Ji-Rong Wen are with the School of Information, Beijing Key Laboratory of Big Data Management and Analysis Methods, and DEKE, Renmin University of China, Beijing 100872, P.R. China. E-mail: dou@ruc.edu.cn, jirong.wen@gmail.com*
- *Tetsuya Sakai is with Department of Computer Science and Engineering, Waseda University. E-mail: tetsuyasakai@acm.org*
- *Rui Zhang is with the School of Computing and Information Systems at The University of Melbourne. E-mail: rui.zhang@unimelb.edu.au*

Manuscript received August 19 2016; revised July 14, 2017.



(a) Official intents of the query "bobcat".



(b) LEFT: OIH is comprised of the solid boxes, whereas EIH includes both solid and dashed nodes. RIGHT: An example showing relevance assessments for the added nodes (under $\mathbf{R}'$ in red) derived from relevance assessments for the official intents (under **R** in blue).

Fig. 1. The official intents, original intent hierarchy (OIH), and extended intent hierarchy (EIH) of No. 77 query "bobcat" in TREC Web Track 2010.

to solve the problems mentioned above. This paper is the extended version of the SIGIR 2016 paper [19]. In the original work [19], we have found that hierarchical measures, especially those using the whole intent hierarchy, are better than existing measures, which uses an intent list, in terms of discriminative power and intuitiveness. The main extensions of this journal version are:

(1) We propose four weighting schemes that are used to model the node weights in an intent hierarchy. We examine the impact of the different types of intent hierarchies, including whether the leaf nodes have the same depth and which weighting scheme is used, for hierarchical measures in terms of discriminative power and intuitiveness. In the experiments, we find the best type of intent hierarchies when nonuniform weights are available and when only uniform weights are known respectively.

(2) We find that it is crucial for hierarchical diversification algorithms to be evaluated by hierarchical measures. The benefits in search result diversification by re-ranking the results to cover the automatically generated hierarchical intents as much as possible may be invisible to existing measures that measure the diversity using intent lists. Hierarchical diversification algorithms show more gains when evaluated by hierarchical measures than existing measures.

(3) Besides TREC Web Track 2009-2013 test collections, we also experiment with NTCIR-11 IMine [1] test collection that has official intent hierarchies with nonuniform weights. We find that the conclusions drawn from the two sources of test collections are consistent with each other.

(4) We reveal that Layer-Aware measures may be counterintuitive because of their preference of high relevance to popular nodes. The experiments confirm that they are less intuitive among the proposed hierarchical measures.

The remainder of this paper is organized as follows. Section 2 describes some existing diversity measures and the methods for testing them. In Section 3, we introduce intent hierarchies, and our method for creating a new test collection based on TREC Web Track test collections. We then propose several new diversity measures that can use the intent hierarchies. Section 4 describes the experimental results and analysis. We conclude our work in Section 5.

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

2 Related Work

Evaluation measures play an important role in the scientific research because they serve as the inexpensive methods for monitoring the technological progress. In the Information Retrieval experiments, evaluation measures use test collections to evaluate system performances. Depending on the task at hand, it is essential to analyze the properties of evaluation measures and use the appropriate ones. Search result diversification aims to cover different intents by a ranked list. Given a query q, most existing measures evaluate the diversified search results by modeling users' information needs as a flat list of intents fig. Some measures can handle intent probability $Pr(ijq)$ and graded relevance assessments but some cannot. In this section, we summarize the previous work on designing and evaluating diversity measures.

2.1 Diversity Measures

2.1.1 Intent Recall

Intent recall(I-rec) [15], also known as subtopic recall [20] 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 to which document d_r is relevant. Then, $I\text{-}rec$ for a certain cutoff K can be expressed as:

$$I\text{-}rec@K = \frac{j \bigcup_{r=1}^{K} I(d_r) j}{j f i g j} \quad (1)$$

The idea of I-rec is to credit minor intents. I-rec does not consider the positions of relevant documents, and cannot handle intent probability and graded relevance assessments.

2.1.2 α-nDCG

In order to balance both relevance and diversity of ranked lists, $\alpha\text{-}nDCG$ [16] is defined as:

$$\alpha\text{-}nDCG@K = \frac{\sum_{r=1}^{K} NG(r)/\log(r+1)}{\sum_{r=1}^{K} NG'(r)/\log(r+1)}$$
$$NG(r) = \sum_{i \in fig} J_i(r)(1-\alpha)^{C_i(r-1)} \quad (2)$$

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.

2.1.3 Intent-Aware measures

Intent-Aware measures ($IA\ measures$) [5] is a general framework to evaluate ranked document lists. Assuming that $\sum_{i \in \{i\}} Pr(i|q) = 1$, $M\text{-}IA$ can be computed as:

$$M\text{-}IA@K = \sum_{i \in \{i\}} Pr(i|q) M_i@K \quad (3)$$

where M_i is the per-intent version of measure M. Measure M can be nDCG [21], ERR [22], nERR [23], etc.

2.1.4 D-measures

Assume that $g_i(r)$ is the gain value of the document at rank r for intent i, and $g_i(r)$ is calculated using per-intent relevance assessments. Then the global gain at rank r is given by:

$$GG(r) = \sum_{i \in \{i\}} Pr(i|q) g_i(r) \quad (4)$$

Let $CGG(r) = \sum_{k=1}^{r} GG(k)$, i.e. the cumulative global gain at rank r. Let $GG\ (r)$ and $CGG\ (r)$ denote the global gain and the cumulative 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. Let $J(r) = 1$ if the document at rank r is relevant to any of the intents $\{i\}$, and $J(r) = 0$ otherwise. Let $C(r) = \sum_{k=1}^{r} J(k)$, which is the number of relevant documents within top r. $D\text{-}nDCG$ and $D\text{-}Q$ at document cutoff K are defined as:

$$D\text{-}nDCG@K = \frac{\sum_{r=1}^{K} GG(r)/\log(r+1)}{\sum_{r=1}^{K} GG\ (r)/\log(r+1)} \quad (5)$$

$$D\text{-}Q@K = \frac{1}{\min(K,R)} \sum_{r=1}^{K} J(r) \frac{C(r) + CGG(r)}{r + CGG\ (r)} \quad (6)$$

where R is the number of judged relevant documents.

2.1.5 D♯-measures

D♯-measures [17] aim to boost intent recall, and to reward documents that are highly relevant to popular intents. $D\sharp\text{-}measure$ is defined as:

$$D\sharp\text{-}measure@K = \gamma\ I\text{-}rec@K + (1-\gamma)\ D\text{-}measure@K \quad (7)$$

where $D\text{-}measure$ can be D-nDCG or D-Q. γ is a parameter between 0 and 1. D♯-measures are free of the under-normalization problem of α-nDCG and IA measures.

The measures mentioned above are widely used in several tasks of TREC Web Track [2] or NII Testbeds and Community for Information access Research (NTCIR) [3], *but they do not take the relationships between intents into consideration*, which is what we aim to deal with in this paper.

2. http://plg.uwaterloo.ca/ trecweb/
3. http://research.nii.ac.jp/ntcir/index-en.html

2.2 Evaluation of Diversity Measures

Given a significance level, *discriminative power* measures the stability of measures across queries and experiments based on significance tests, e.g. paired bootstrap test [24], Tukey's Honestly Significant Differences(HSD) [25] test, etc.

Concordance test [26] is proposed to quantify the intuitiveness of diversity measures. In concordance test, one or more gold standard measures are chosen and assumed to truly represent intuitiveness. Given two diversity measures M_1 and M_2, the relative intuitiveness of M_1 (or M_2) is measured in terms of preference agreement with the gold standard measures. The preference agreement is that M_1 (or M_2) agrees with the gold standard measure(s) about which one of two ranked lists should be preferred.

Kendall's τ [27] is a statistic to measure the *rank correlation* of two rankings. However, τ lacks the property of top heaviness. In the context of IR evaluation, the swaps near the top is generally more important than those near the bottom. τ_{ap} [28] is proposed to deal with the problem. Note that τ is symmetric but τ_{ap} is not. However, a symmetric τ_{ap} can be obtained by averaging two τ_{ap} values when each list is treated as the former one. Both τ and τ_{ap} range from -1, which implies two ranked lists perfectly disagree, to 1, which implies two ranked lists are identical.

3 Proposed Methods

In this section, we define two kinds of intent hierarchies with four weighting schemes to represent the relationships between user intents. We then introduce our method for creating such intent hierarchies and obtaining the relevance assessments based on TREC Web Track 2009-2013 diversity test collections. Last, we propose several diversity measures based on intent hierarchies, and show that the new measures outperform the corresponding existing measures.

3.1 Intent Hierarchies

Given a query q, the users' information needs are represented as a set of intents $\{i\}$. We assume these intents cannot be further subdivided, and refer to them as *atomic intents*. Based on the semantic relatedness of the intents, we build an intent hierarchy possessing the following properties:

Property 1. The intent hierarchy is in a tree structure, where every child has only one parent.

Property 2. The root of intent hierarchy is denoted by q itself, which stands for the users' information needs as a whole. The root is a dummy node for the completeness of the tree, and is not considered in our measures.

Property 3. When q is broad, the intent hierarchy is built in such a way that a parent node refers to a more general concept than its children, and a child node refers to one aspect of its parent. When q is ambiguous, each child node of the root denotes one interpretation of q, and each of its subtrees is built in the same way as a broad query.

Property 4. These atomic intents, i.e. $\{i\}$, correspond one to one with leaves of the intent hierarchy. This means the number of leaves in the intent hierarchy is the same as the number of the atomic intents.

An intent hierarchy that satisfies the above four properties is called an *original intent hierarchy* (OIH). The leaf nodes of an OIH may not have the same depth. We extend such an OIH by recursively adding one child node to the leaves which do not have the highest depth until it satisfies:

Property 5. All leaf nodes of the intent hierarchy, i.e. the atomic intents, have the same depth.

The resulting intent hierarchy are called an *extended intent hierarchy* (EIH). As an OIH, the leaves of an EIH also correspond one to one with the atomic intents fig.

We define an atomic intent subset S as a subset of atomic intents which has at lease two atomic intents. Extending OIH to EIH is justified as follows: (1) An atomic intent subset S corresponds to jSj leaves in an OIH or EIH. The maximum depth of the nodes that are common ancestors of these leaves acts as a simple indicator of how much these intents are related to each other. Hence, we can refer to the maximum depth as the *redundant degree* of S; (2) When OIH is extended to EIH, the redundant degree remains the same for any atomic intent subset. This means that if OIH, which is built by experts, thinks that one atomic intent subset is more related than another, so does EIH and vise versa.

We consider the root of an intent hierarchy as the zeroth layer, the child nodes of the root as the first layer and so forth. If an intent hierarchy only has the zeroth layer and the first layer, the height of the intent hierarchy is one. In the paper, a *single-layer intent hierarchy* refers to an intent hierarchy whose height is one, while a *multilayer intent hierarchy* refers to that whose height is greater than one.

3.2 Weighting Intent Hierarchies

For an intent hierarchy, its nodes are weighted according to their relative popularity. The node weights should satisfy:

Property 6. $W(q) = 1$ and $W(n) = \sum_{c2C(n)} W(c)$ $(8n\ 2\ N)$

where q is the root node, N is the set of non-terminal nodes, and $C(n)$ is the set of child nodes of node n.

We propose four weighing schemes that make resulting node weights satisfy Property 6 as follows:

(1) *Uniformly top-down* (UT) can be used in any situation. It assumes that given a parent node, its child nodes are equally weighted. Starting from the root, the weight of node n is $W(n) = W(p)/jC(p)j$, where p is the parent node of n, and $C(p)$ is the set of child nodes of p.

(2) *Uniformly bottom-up* (UB) can be used in any situation. It assumes that given an intent hierarchy, its leaf nodes are equally weighted, i.e. $W(n) = \frac{1}{jBj}$ $(8n\ 2\ B)$ where B is the set of leaf nodes. Starting from the leaves, the weight of node n is defined as: $W(n) = \sum_{c2C(n)} W(c)$, where $C(n)$ is the set of child nodes of node n.

(3) *Nonuniformly top-down* (NT) can only be used when all the node weights in an intent hierarchy are known but they do not satisfy Property 6. This may happen when the atomic intents that receive no relevant documents are removed and node weights need renormalization. Starting from the root, the new weight of node n is defined as:

$$W(n) = \frac{W^{\theta}(n)}{\sum_{c2C(p)} W^{\theta}(c)} W(p) \tag{8}$$

where p is the parent of n, $C(p)$ is the children of p, $W^{\theta}(n)$ ($W^{\theta}(c)$) is the original weight of node n (c).

(4) *Nonuniformly bottom-up* (NB) can be used if only the weights of leaf nodes in an intent hierarchy are known. First the weights of leaf nodes are normalized, and then starting from the leaves, the weight of node n is $W(n) = \sum_{c2C(n)} W(c)$, where $C(n)$ is the set of child nodes of n.

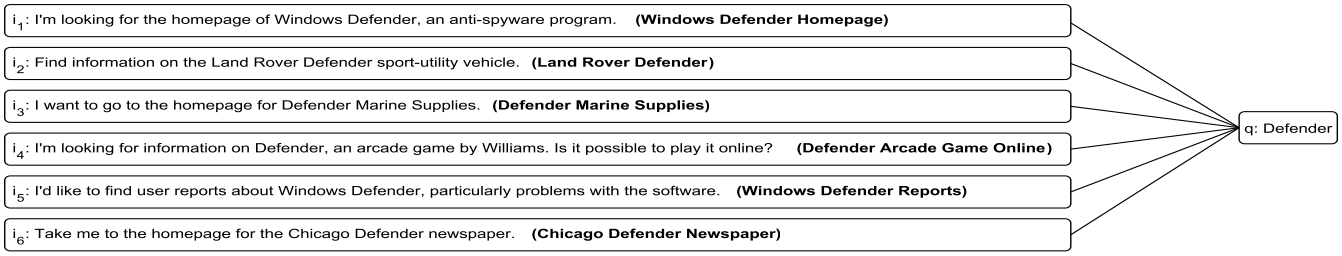
3.3 Creating Intent Hierarchies

Our proposed hierarchical measures, which will be introduced in 3.4, can work with any type of intent hierarchies. NTCIR-11 IMine test collection comes with multilayer intent hierarchies, and we directly use them in the experiments. In this section, we show the method for creating multilayer intent hierarchies from the predefined intents on TREC Web Track 2009-2013 diversity test collections. For each query in the test collections, the first intent is the same as the description of the query itself. Although the descriptions are the same, if a query has several different interpretations, the first intent is just one of these interpretations. A query's first intent does not refer to a more general concept than the other intents. So we do not treat the first intent differently.

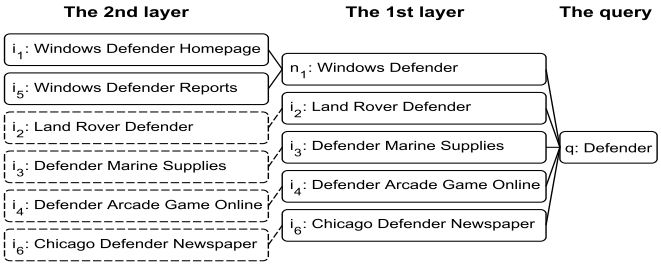
We use the official intents as atomic intents to avoid reassessing the relevance of documents. First we create OIH by manually grouping the official intents based on their semantic similarity. Then, we extend them to EIH. Figure 1 illustrates how we create OIH and EIH for the query "bobcat" in TREC 2010 Web Track. It can be seen from Figure 1(a) that this query has four official intents and intent i_1 and i_3 are related to the trade involving bobcat tractors. So we create a new node n_1 that stands for "bobcat tractors" as their parent node. Similarly, n_1 and i_4 are related to "bobcat company," hence we create another new node n_2 representing "bobcat company" as their parent. Finally, since n_2 ("bobcat company") and i_2 ("wild bobcat") are two distinct interpretations of query "bobcat," they are considered as the child nodes of the root node. The resultant OIH is shown in solid boxes in the left of Figure 1(b). Further, we extend the OIH by adding a child to i_2, and adding a child and a grandchild to i_4. The resultant EIH is shown in solid boxes plus dashed boxes in the left of Figure 1(b).

As for the OIH or EIH shown in Figure 1(b): (1) It is in a tree structure (Property 1); (2) Its root is query "bobcat" itself (Property 2); (3) The query is ambiguous, so the child nodes of root are its two different interpretations, i.e. "bobcat company" and "wild bobcat." A parent node refers to a more general concept than its children (Property 3), e.g. "bobcat company" is more general than "bobcat company homepage;" (4) The leaf nodes are exactly the official intents of query "bobcat" (Property 4). Further, the depth of all the leaf nodes in EIH is three (Property 5).

Only the relevance assessments for the original intents are available in TREC Web Track diversity test collections. For the intent hierarchies we create, document relevance judgments are just available for their leaf intents. As assessing document relevance is usually very time-consuming, it is not desirable to reassess the documents for the intermediate nodes. Fortunately, according to Property 3, a parent node of an intent hierarchy stands for a more general concept than its child nodes. Hence **it is reasonable to assume that if a document is relevant to a node, it would be relevant to the node's parent.** This means that we can



(a) Official intents of the query "defender".



	R			R'	S						
	i_1	i_4	i_5	n_1	D#-nDCG$_1$@1	D#-nDCG$_2$@1	D#-nDCG-LA$_E$@1	I-rec$_1$@1	I-rec$_2$@1	N-rec$_E$@1	N-rec$_O$@1
d_1	1	1	0	1	0.70	0.67	0.68	0.40	0.33	0.36	0.43
d_2	1	0	1	1	0.35	0.67	0.51	0.20	0.33	0.27	0.43
d_3	1	0	0	1	0.35	0.33	0.34	0.20	0.17	0.18	0.29
d^*	1	1	1	1	(Note that these measures are computed **at rank 1**)						

(b) LEFT: OIH is comprised of the solid boxes, whereas EIH includes both solid and dashed nodes. RIGHT: 1st column: document IDs (d^* : the ideal one), each of which is a ranked list of length 1. 2nd to 5th column (**R** and $\mathbf{R}^{\theta}$): relevance assessments for the official intents (in red) and derived relevance assessments for added nodes (in blue). 6th to 12th column (**S**): the scores computed by a measure at rank 1.

Fig. 2. The official intents, original intent hierarchy (OIH), and extended intent hierarchy (EIH) of No. 20 query "defender" in TREC Web Track 2009.

derive relevance assessments for the intermediate nodes starting from the leaves. In this paper, we simply let:

$$L_d(n) = \max_{c \in C(n)} L_d(c) \tag{9}$$

where $L_d(n)$ is the relevance rating assigned to document d for node n, and $C(n)$ is the set of child nodes of n.

We show an actual document (denoted by d in the following) from TREC Web Track 2010 diversity test collection in Figure 1(b). In the table, the officially provided relevance assessments are marked in blue, e.g. the relevance rating of d for i_1 is 1. Firstly, node n_1 has two child nodes, i_1 and i_3, and the relevance ratings of d for them are 1 and 0. According to Equation (9), the relevance rating of d for n_1 is 1. Similarly, we can derive the relevance rating for n_2 based on its child node i_4 and n_1. These derived relevance assessments are shown in red in the table of Figure 1(b).

To conclude, we create a new dataset containing intent hierarchies by manually grouping the official intents from TREC Web track test collections. The good news is that we do not need to reassess document relevance with regards to the intent hierarchies. We directly leverage document relevance assessments for the leaf intents, and automatically assign relevance ratings for the intermediate intents. This also implies that when we want to create hierarchical intents for evaluating diversity, we just need to assess document relevance for the leaf nodes or atomic intents.

Most of the time of creating the new dataset is spent on grouping the original intents. On average, we spend about three minutes per query mainly in understanding the original intents with the assistance of Google and Bing.

3.4 Hierarchical Measures

We assume that: (1) The user who is interested in an intent will be interested in the broader intents. Based on this assumption, we automatically assign relevance assessment for the intermediate intents of an intent hierarchy according to Equation 9 in Section 3.3; (2) The user who is interested in an atomic intent will be likely to be more interested in the more related atomic intents, and be less interested in the less related atomic intents in an intent hierarchy. The relatedness between two atomic intents is defined by the length of the shortest path between the two intents: the shorter the length of the shortest path is, the more related the two intents are. This implies that the search results that cover more related atomic intents will only be of interested to a specific user group, and should be considered less diverse by the measures. N-rec and the measures based on N-rec reward search results that cover less related atomic intents in an intent hierarchy, which is illustrated in Section 3.4.1.

3.4.1 Node Recall

Given a query q, let V denote the nodes in its intent hierarchy except for its root. Let d_r denote the document at rank r, and let $N(d_r)$ denote the nodes in V to which d_r is relevant. Node recall (N-rec) at rank K is defined as:

$$N\text{-}rec@K = \frac{|\bigcup_{r=1}^{K} N(d_r)|}{|V|} \tag{10}$$

N-rec is a natural generalization of I-rec when using the intent hierarchy . They both are rank-insensitive and cannot handle graded relevance assessments. I-rec credits minor intents,while N-rec credits minor nodes, which contribute to rewarding wide coverage of users' information needs.

We use an example to show that N-rec outperforms I-rec in terms of discriminative power. In the right of Figure 2(b), I-rec$_1$@1 means only using the first layer, I-rec$_2$@1 means only using the second layer, and N-rec$_E$@1 means using EIH when computing N-rec. These measures are computed at

rank 1. Note that the original I-rec is equal to I-rec$_2$. We find that $d_1 > d_2 = d_3$ according to I-rec$_1$@1, $d_1 = d_2 > d_3$ according to I-rec$_2$@1, whereas $d_1 > d_2 > d_3$ according to N-rec$_E$@1. As we discussed in Section 3.4.2, The real preference is $d_1 > d_2 > d_3$. I-rec$_1$@1 fails to tell the difference between d_2 and d_3, while I-rec$_2$@1 fails to distinguish between d_1 and d_2. Only N-rec$_E$@1 can tell the difference between the three documents, and thus is more discriminative than I-rec. The ranking list that covers less related atomic intents is rewarded by N-rec, which is consistent with the second user assumption.

Another point worth noting is that **the types of intent hierarchies are crucial to N-rec.** In the right of Figure 2(b), N-rec$_O$@1 means using OIH instead of EIH. We find that N-rec$_O$@1 cannot determine which one of d_1 and d_2 is better because they have exactly the same score. This indicates that using EIH has higher discriminative power than using OIH.

We aim to retrieve documents that cover as many nodes of intent hierarchies as possible. Besides, we prefer the documents that are highly relevant to more popular nodes and layers. N-rec mainly rewards wide coverage of different nodes of intent hierarchies in the top ranks. In the following, we will discuss some measures to complement N-rec.

3.4.2 Layer-Aware measures

Our next proposal is to first evaluate a ranked list for each layer of an intent hierarchy using existing measures, then combine the evaluation scores across multiple layers.

Let H denote the height of the intent hierarchy, and let $L = \{l_1, l_2, ..., l_H\}$ denote its H layers. We define Layer-Aware measures ($LA\ measures$) at document cutoff K as:

$$M\text{-}LA@K = \sum_{i=1}^{H} w_i \quad M_i@K \tag{11}$$

Here, w_i is the weight of layer l_i, where $\sum_{i=1}^{H} w_i = 1$, and M_i is the evaluation score of measure M by using the intents of layer l_i. For example, ERR-IA-LA is computed as follows: (1) For each layer, compute the per-layer scores of ERR-IA; (2) Compute the weighted average of the per-layer scores using Equation (11).

We find that the combination of measures over layers of intent hierarchies could outperform the original measures. We use the query "defender", No. 20 topic in TREC Web Track 2009 [29], as an example. We choose this query because it has a relatively simple intent hierarchy. Its EIH is shown in the left of Figure 2(b). Suppose we have three documents, d_1-d_3, and each of them can be viewed as a ranked list containing only one document. Their relevance assessments for the EIH are displayed in blue in the right of Figure 2(b). To save space, the nodes that receive no relevant documents within the documents are not displayed. Assume that d is the first document within the ideal rank list and it is relevant to every node displayed. In the right of Figure 2(b), D♯-nDCG$_1$@1 is the evaluation score when only using the first layer of the EIH, D♯-nDCG$_2$@1 means only using the second layer, and D♯-nDCG-LA$_E$@1 is the average of D♯-nDCG$_1$@1 and D♯-nDCG$_2$@1. The original D♯-nDCG is equal to D♯-nDCG$_1$. We use the measures to score d_1 to d_3, i.e. evaluating at document cutoff 1.

We show the evaluation results in Figure 2(b): $d_1 > d_2 = d_3$ under D♯-nDCG$_1$@1, $d_1 = d_2 > d_3$ under D♯-nDCG$_2$@1, whereas $d_1 > d_2 > d_3$ under D♯-nDCG-LA$_E$@1. Here, ">" means the former document is preferred compared with the latter when evaluating them at rank 1, and "=" means neither is preferred. The real preference is $d_1 > d_2 > d_3$: (1) d_1 is more diversified than d_2 because d_1 covers both "windows defender" and "defender arcade game online," while d_2 only covers the former; (2) d_2 is more diversified than d_3 because d_2 covers both "windows defender homepage" and "windows defender reports" while d_3 just covers the former. Here, only D♯-nDCG-LA$_E$@1 is consistent with the real preference. D♯-nDCG$_1$@1 fails to tell the difference between d_2 and d_3, whereas D♯-nDCG$_2$@1 fails to tell the difference between d_1 and d_2. This indicates that the combination over layers has higher potential to reflect real user satisfaction than the use of a flat list of intents.

3.4.3 HD-measures

The global gain of an intent hierarchy at rank r is given by:

$$GG_h(r) = \sum_{i=1}^{H} w_i \quad GG_i(r) \tag{12}$$

where w_i is the weight of layer l_i and $GG_i(r)$ is the global gain for layer l_i at rank r. Let $CGG_h(r) = \sum_{k=1}^{r} GG_h(k)$, which is the cumulative global gain for the intent hierarchy at rank r. Further, let $GG_h(r)$ and $CGG_h(r)$ denote the global gain and the cumulative global gain for the intent hierarchy at rank r in the ideal ranked list. The ideal list is obtained by listing up all the judged documents in descending order of global gains for the intent hierarchy. Let $J(r) = 1$ if the document at rank r is relevant to the intent hierarchy, and $J(r) = 0$ otherwise. Let $C(r) = \sum_{k=1}^{r} J(k)$. We define $HD\text{-}nDCG$ and $HD\text{-}Q$ at document cutoff K as:

$$HD\text{-}nDCG@K = \frac{\sum_{r=1}^{K} GG_h(r)/\log(r+1)}{\sum_{r=1}^{K} GG_h(r)/\log(r+1)} \tag{13}$$

$$HD\text{-}Q@K = \frac{1}{\min(K;R)} \sum_{r=1}^{K} J(r) \frac{C(r) + \ CGG_h(r)}{r + \ CGG_h(r)} \tag{14}$$

where R is the number of judged documents relevant to the intent hierarchy.

3.4.4 LD♯-measures

We use the leaf nodes of intent hierarchies to compute D-measures, such as D-nDCG and D-Q. Then, $LD♯\text{-}measure$ is defined as:

$$LD]\text{-}measure@K = \ N\text{-}rec@K + (1 \quad)D\text{-}measure@K \tag{15}$$

where γ is a parameter controlling the tradeoff between diversity and relevance. Since D-measures only use the leaf nodes, LD♯-measures mainly reward high relevance with more popular leaves Also, LD♯-measures cannot handle the weights of layers. To tackle these, we propose HD♯-measures and LAD♯-measures in the next two sections.

3.4.5 HD♯-measures

We define $HD♯\text{-}measure$ as:

$$HD]\text{-}measure@K = \ N\text{-}rec@K + (1 \quad)HD\text{-}measure@K \tag{16}$$

where $HD\text{-}measure$ can be HD-nDCG or HD-Q, and γ is a parameter between 0 and 1.

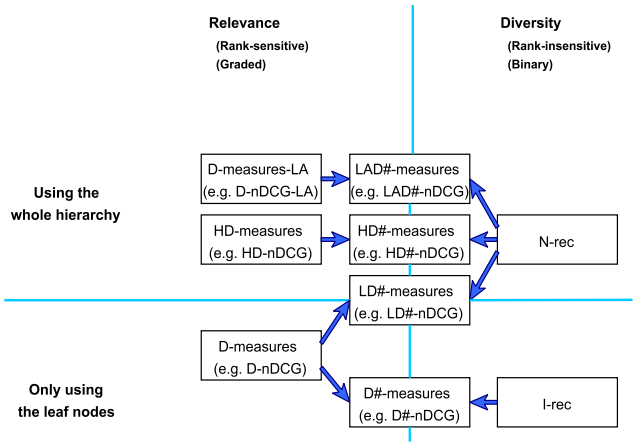


Fig. 3. Relationships of D*♯*-measures, LD*♯*-measures, HD*♯*-measures, and LAD*♯*-measures.

3.4.6 LAD♯-measures

We define $LAD\sharp\text{-}measure$ as:

$$LAD\sharp\text{-}measure@K = \gamma\, N\text{-}rec@K + (1-\gamma)\, D\text{-}measure\text{-}LA@K \quad (17)$$

where γ is a parameter balancing diversity with relevance, and D-measure-LA is the LA version of D-measure.

To measure the relevance of ranked lists, HD♯-measures use HD-measures, while LAD♯-measures use D-measures-LA. HD-measures and D-measures-LA reward high relevance to more popular nodes, and can handle layer weights. The difference between them 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. Take HD-nDCG and D-nDCG-LA as an example:

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

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

where $GG_i(r)$ is the global gain for layer l_i at rank r, and $D\text{-}nDCG_i$ means only using the nodes of layer l_i.

3.4.7 Summarization and Discussion

We call the proposed diversity measures *hierarchical measures*. LA measures mainly reward high relevance to major nodes, and thus take little account of minor nodes. This can be seen if we transform the definition of M-IA-LA, e.g., into

$$M\text{-}IA\text{-}LA@K = \sum_{n \in V} w_n M_n@K \quad (18)$$

where V is the nodes of an intent hierarchies except for the root, and w_n is the weight of node n. M_n is the per-node version of measure M, which can be nDCG, ERR, etc. This problem does not exist for the hierarchical measures which rely on N-rec because N-rec can credit minor intents.

Each of D♯-measures, LD♯-measures, HD♯-measures, and LAD♯-measures is a linear combination of two measures: one mainly rewards search result diversity, whereas another mainly rewards the relevance. We show their relationships in Figure 3. The figure shows that: (1) To reward the diversity, LD♯-measures, HD♯-measures, and LAD♯-measures use the whole intent hierarchy, but D♯-measures only use the leaves; (2) To reward the relevance, HD♯-measures and LAD♯-measures use the whole intent hierarchy, but D♯-measures and LD♯-measures only use the leaves.

Hierarchical measures work more easily with EIH than with OIH because they require that for every layer, the node weights sum up to 1. EIH can guarantee this no matter which weighing scheme is used, while OIH cannot. So one drawback of using OIH is that when the node weights of one layer do not sum up to 1, renormalization is required. Besides, some hierarchical measures work in a layer wise way, like LA measures and HD-measures, but the document relevance assessments for some atomic intents may not be considered in the deep layers when using OIH. This means that hierarchical measures using OIH may neglect some atomic intents, and be counterintuitive sometimes.

TABLE 1
The types of intent hierarchies in TREC Web Track 2009-2013 and NTCIR-11 IMine

	(A) TREC (Uniform)		(B) NTCIR (EIH)	
	OIH	EIH	Uniform	Nonuniform
Top-down	OUT	EUT	EUT	ENT
Bottom-up	OUB	EUB	EUB	ENB

4 Experiments

First, we show that hierarchical measures using EIH weighted bottom-up, i.e. computing the node weights upwards given the leaf node weighs, have advantages over existing measures in discriminative power and intuitiveness. Then, we present the experimental results of comparing the performance of hierarchical measures using different intent hierarchies, i.e. OIH or EIH weighted top-down or bottom-up uniformly or nonuniformly. Last, we show the benefits of using nonuniform weights, which are usually more costly than simple uniform weights.

4.1 Settings

We experiment with diversity measures on TREC Web Track 2009-2013 diversity test collections and NTCIR-11 IMine test collection. We build a new test collections from the TREC test collections , which contains intent hierarchies and the relevance assessments. The new dataset is publicly available on[4]. Available types of intent hierarchies in TREC Web Track 2009-2013 and NTCIR-11 IMine are summarized in Table 1. TREC Web Track 2009-2013 do not provide nonuniform intent weights, so only uniformly top-down and uniformly bottom-up weighting schemes can be applied. The intent hierarchies in NTCIR-11 IMine are EIH already. Since NTCIR-11 IMine provides nonuniform intent weights, the other two weighting schemes can also be applied, i.e. nonuniformly top-down and nonuniformly bottom-up. In the following, we use subscripts to denote the types of intent hierarchies.

Except for concordance test, we use document cutoff $K = 20$ for all the measures. In concordance test, we use document cutoff $K = 10$ instead because: (1) When doing case studies in Section 4.2.3, the details of top 20 documents

4. http://www.playbigdata.com/dou/heval/

5

TABLE 2
Discriminative power of existing measures and LA measures based on the paired bootstrap test at = 0:05.

(A) TREC Web Track 2009-2013 diversity test collections.			
Existing measure	disc. power	LA measure	disc. power
-nDCG	58.11%	-nDCG-LA	58.32%
ERR-IA	53.26%	ERR-IA-LA	53.79%
nDCG-IA	54.63%	nDCG-IA-LA	55.37%
Q-IA	47.47%	Q-IA-LA	48.95%
D*J*-nDCG	57.05%	D*J*-nDCG-LA	57.47%
D*J*-Q	56.21%	D*J*-Q-LA	56.32%
(B) NTCIR-11 IMine test collection.			
measure	disc. power	LA measure	disc. power
-nDCG	65.63%	-nDCG-LA	63.13%
ERR-IA	62.50%	ERR-IA-LA	65.00%
nDCG-IA	68.75%	nDCG-IA-LA	73.13%
Q-IA	65.00%	Q-IA-LA	71.88%
D*J*-nDCG	68.75%	D*J*-nDCG-LA	69.38%
D*J*-Q	70.63%	D*J*-Q-LA	71.25%

6

TABLE 3
Discriminative power of D*J*-measures, LD*J*-measures, HD*J*-measures, and LAD*J*-measures based on the paired bootstrap test at = 0:05.

(A) TREC Web Track 2009-2013 diversity test collections.			
measure	disc. power	measure	disc. power
D*J*-nDCG	57.05%	D*J*-Q	56.21%
LD*J*-nDCG	57.58%	LD*J*-Q	56.42%
HD*J*-nDCG	57.26%	HD*J*-Q	56.53%
LAD*J*-nDCG	57.26%	LAD*J*-Q	56.53%
(B) NTCIR-11 IMine test collection.			
measure	disc. power	measure	disc. power
D*J*-nDCG	68.75%	D*J*-Q	70.63%
LD*J*-nDCG	68.75%	LD*J*-Q	71.25%
HD*J*-nDCG	69.38%	HD*J*-Q	71.25%
LAD*J*-nDCG	69.38%	LAD*J*-Q	71.25%

cannot be fitted into one page; (2) The conclusions drew when $K = 10$ are almost the same as those drew when $K = 20$. We set $\gamma = 0.5$ in Equation (7), (15), (16), and (17).

4.2 Hierarchical Measures Outperform Existing Measures

In this section, we fix the type of intent hierarchies used by hierarchical measures: (1) On TREC Web Track 2009-2013 test collections, hierarchical measures use EIH weighted uniformly bottom-up; (2) On NTCIR-11 IMine test collection, hierarchical measures use EIH weighted nonuniformly bottom-up. In the following sections, we will give explanation for selecting such types of intent hierarchies.

4.2.1 Discriminative Power

Following the previous work [17], [24], [26], [30], [31], we use the paired bootstrap test and set $B = 1,000$ (B is the number of bootstrap samples). On TREC test collections we conduct the experiments as follows: (1) Sampling 20 submitted runs every year (2009-2013), which produces 950 pairs of sampled runs; (2) With the 950 pairs of sampled runs, computing the discriminative power using all the queries in TREC test collections. On the NTCIR test collection, we conduct the experiments as follows: (1) Using all the runs in NTCIR-11 IMine test collection because there are only 11 runs in Chinese and 15 runs in English. This gives us 160 pairs of runs; (2) With the 160 pairs of runs, computing the discriminative power using all the queries in the NTCIR test collection. The results are shown in Tables 2 and 3.

TABLE 4
Intuitiveness based on preference agreement with gold standard measures. For each measure pair, the higher score is shown in bold.

(A) TREC Web Track 2009-2013. Gold standard: N-rec and Precision.				
	-nDCG-LA	ERR-IA-LA	nDCG-IA-LA	Q-IA-LA
LD*J*-nDCG	:**673**/.237	:**715**/.185	:**385**/.051	:**375**/.227
HD*J*-nDCG	:**667**/.233	:**712**/.181	:**392**/.086	:**372**/.229
LAD*J*-nDCG	:**672**/.236	:**714**/.184	:**385**/.054	:**374**/.226
LD*J*-Q	:**787**/.106	:**814**/.080	:**678**/.151	:**578**/.111
HD*J*-Q	:**787**/.105	:**814**/.079	:**678**/.152	:**578**/.111
LAD*J*-Q	:**787**/.106	:**814**/.080	:**678**/.151	:**578**/.111
(B) NTCIR-11 IMine. Gold standard: N-rec and Precision.				
	-nDCG-LA	ERR-IA-LA	nDCG-IA-LA	Q-IA-LA
LD*J*-nDCG	:**705**/.171	:**806**/.131	:**653**/.112	:**654**/.144
HD*J*-nDCG	:**704**/.145	:**813**/.108	:**624**/.130	:**633**/.144
LAD*J*-nDCG	:**721**/.151	:**815**/.114	:**657**/.119	:**657**/.130
LD*J*-Q	:**785**/.096	:**835**/.077	:**718**/.091	:**728**/.067
HD*J*-Q	:**792**/.070	:**845**/.058	:**718**/.100	:**736**/.082
LAD*J*-Q	:**782**/.096	:**837**/.078	:**720**/.094	:**730**/.070

By comparing the discriminative power of existing measures and the corresponding LA measures in each row of Table 2, we find that except α-nDCG-LA, LA measures using EIH weighted bottom-up usually have higher discriminative power than (or have the same discriminative power as) the corresponding existing measures, especially in the case of IA measures. For example, (1) On TREC test collections, Q-IA-LA (49.16%) outperforms Q-IA (47.47%) in terms of discriminative power; (2) On the NTCIR test collection with nonuniform weights, Q-IA-LA (71.88%) beats Q-IA (65.00%) in terms of discriminative power.

By comparing the results of discriminative power of D♯-measures, LD♯-measures, HD♯-measures, and LAD♯-measures (each column in Table 3), we find that (1) When using EIH weighted bottom-up, LD♯-measures, HD♯-measures and LAD♯-measures are consistently better than (or as good as) D♯-measures in terms of discriminative power. For example, on TREC test collections, LD♯-nDCG, HD♯-nDCG, and LAD♯-nDCG (57.58%, 57.26%, and 57.26% respectively) are all more discriminative than D♯-nDCG (57.05%) in terms of discriminative power; (2) On the NTCIR test collection, the results of discriminative power of most measures are the same, which is due to the relatively small size of NTCIR-11 IMine test collection.

4.2.2 Concordance Test

In Section 3, we argue that LA measures may be less intuitive among hierarchical measures. On TREC test collections (or the NTCIR test collection), we compare the intuitiveness of two measures, M_1 and M_2, as follows: (1) N-rec (or Precision) is used as the gold standard: The relative intuitiveness of M_1 (or M_2) is the ratio of ranked list pairs, for which M_1 (or M_2) and the gold standard have the same preference, to those for which M_1 and M_2 have different preference; (2) Both N-rec and Precision are used as the gold standard measures: The relative intuitiveness of M_1 (or M_2) is the ratio of ranked list pairs, for which M_1 (or M_2) has the same preference with N-rec and Precision, to those for which M_1 and M_2 have different preference. The results are shown in Table 4. N-rec is used as the gold standard measure for diversity because: (1) It is a simple binary measure; (2) It measures search result diversity better than I-rec, which is

TABLE 5
Intuitiveness based on preference agreement with gold standard measures. For each measure pair, the higher score is shown in bold.

(A) TREC Web Track 2009-2013. Gold standard: N-rec					
	ERR-IA	D*J*-nDCG	LD*J*-nDCG	HD*J*-nDCG	LAD*J*-nDCG
-nDCG	:**987**/.390	.646/:**979**	.632/:**985**	.643/:**980**	.640/:**982**
ERR-IA	-	.561/:**984**	.551/:**990**	.559/:**987**	.557/:**988**
D*J*-nDCG	-	-	.303/:**741**	.646/:**789**	.607/:**794**
LD*J*-nDCG	-	-	-	:**905**/.762	:**895**/.777
HD*J*-nDCG	-	-	-	-	.682/:**950**
(B) TREC Web Track 2009-2013. Gold standard: Precision.					
	ERR-IA	D*J*-nDCG	LD*J*-nDCG	HD*J*-nDCG	LAD*J*-nDCG
-nDCG	:**773**/.373	.493/:**698**	.494/:**698**	.492/:**705**	.493/:**704**
ERR-IA	-	.456/:**725**	.456/:**725**	.455/:**731**	.455/:**729**
D*J*-nDCG	-	-	.580/:**612**	.556/:**692**	.560/:**682**
LD*J*-nDCG	-	-	-	.544/:**736**	.547/:**730**
HD*J*-nDCG	-	-	-	-	:**762**/.529
(C) TREC Web Track 2009-2013. Gold standard: N-rec and Precision					
	ERR-IA	D*J*-nDCG	LD*J*-nDCG	HD*J*-nDCG	LAD*J*-nDCG
-nDCG	:**761**/.082	.245/:**681**	.236/:**684**	.243/:**686**	.241/:**686**
ERR-IA	-	.187/:**712**	.181/:**715**	.186/:**718**	.185/:**717**
D*J*-nDCG	-	-	.068/:**372**	.300/:**491**	.274/:**486**
LD*J*-nDCG	-	-	-	.457/:**504**	.450/:**513**
HD*J*-nDCG	-	-	-	-	.456/:**485**
(A) NTCIR-11 IMine. Gold standard: N-rec.					
	ERR-IA	D*J*-nDCG	LD*J*-nDCG	HD*J*-nDCG	LAD*J*-nDCG
-nDCG	:**973**/.159	.385/:**885**	.420/:**867**	.458/:**862**	.419/:**883**
ERR-IA	-	.256/:**944**	.284/:**933**	.319/:**922**	.292/:**935**
D*J*-nDCG	-	-	.548/:**644**	:**643**/.622	.540/:**698**
LD*J*-nDCG	-	-	-	:**719**/.626	.605/:**721**
HD*J*-nDCG	-	-	-	-	.475/:**852**
(B) NTCIR-11 IMine. Gold standard: Precision.					
	ERR-IA	D*J*-nDCG	LD*J*-nDCG	HD*J*-nDCG	LAD*J*-nDCG
-nDCG	:**886**/.227	.348/:**841**	.466/:**812**	.467/:**829**	.467/:**811**
ERR-IA	-	.287/:**886**	.361/:**868**	.363/:**870**	.366/:**860**
D*J*-nDCG	-	-	:**788**/.587	:**720**/.636	:**762**/.619
LD*J*-nDCG	-	-	-	.662/:**727**	.663/:**698**
HD*J*-nDCG	-	-	-	-	:**738**/.639
(C) NTCIR-11 IMine. Gold standard: N-rec and Precision.					
	ERR-IA	D*J*-nDCG	LD*J*-nDCG	HD*J*-nDCG	LAD*J*-nDCG
-nDCG	:**868**/.050	.100/:**733**	.182/:**698**	.213/:**707**	.183/:**710**
ERR-IA	-	.067/:**834**	.116/:**811**	.139/:**802**	.120/:**805**
D*J*-nDCG	-	-	:**385**/.346	:**413**/.350	.365/:**405**
LD*J*-nDCG	-	-	-	:**439**/.403	.337/:**442**
HD*J*-nDCG	-	-	-	-	.295/:**525**

traditionally used as the gold standard measure for diversity. We find that LA measures are consistently less intuitive than LD♯-measures HD♯-measures, and LAD♯-measures.

We compare α-nDCG, ERR-IA, D♯-nDCG, LD♯-nDCG, HD♯-nDCG, and LAD♯-nDCG in terms of intuitiveness. We do the concordance test on all the queries in TREC test collections or all the queries in the NTCIR test collection, and show the results in Table 5. In Table 5, blocks (A) and blocks (B) use N-rec and Precision as the gold standard measure respectively, whereas in blcoks (C), both N-rec and Precision are used as the gold standard measures.

Table 5 shows that: (1) In terms of the diversity, LD♯-nDCG, HD♯-nDCG, and LAD♯-nDCG are usually more intuitive than existing measures. This is expected because these hierarchical measures depend on N-rec by means of Equation (15) and the like; (2) D♯-nDCG is consistently more intuitive than α-nDCG and ERR-IA, but generally less intuitive than LD♯-nDCG, HD♯-nDCG, and LAD♯-nDCG; (3) Among the hierarchical measures, LD♯-nDCG is most intuitive in terms of diversity; HD♯-nDCG is most intuitive in terms of relevance; LAD♯-nDCG is the most intuitive measure in terms of both diversity and relevance.

Table 5 shows that LD♯-nDCG, HD♯-nDCG and LAD♯-nDCG, which use the whole intent hierarchy to measure diversity, are more intuitive than D♯-nDCG in terms of diversity. HD♯-nDCG and LAD♯-nDCG, which use the whole intent hierarchy to measure relevance, are more intuitive than D♯-nDCG and LD♯-nDCG in terms of relevance. We get the same result when both diversity and relevance are considered. This indicates that **using the whole intent hierarchy instead of using the leaf nodes only can improve the intuitiveness of measures.** This is because: (1) Diversity is considered as covering as many nodes of an intent hierarchy as possible, so the measures using the whole intent hierarchy are more intuitive; (2) When using the whole intent hierarchy to compute diversity, using the whole intent hierarchy to compute relevance achieves a better balance between diversity and relevance than only using the leaves.

4.2.3 Case Studies

D♯-nDCG, LD♯-nDCG, HD♯-nDCG, and LAD♯-nDCG are closely related (shown in Sections 3.4.7 and 4.2.4). Besides, they are consistently more intuitive than the others. We further examine their differences in terms of intuitiveness by looking at some real examples from the submitted runs in TREC Web Track 2009-2013 diversity task.

Specifically, we select five pairs of real ranked lists from TREC Web Track diversity runs in Table 6, and refer to them as **Case A-E**. For example, **Case A** stands for two runs cmuFuTop10D and THUIR10DvNov for No. 77 query; The middle column shows the relevance assessments of the top ten documents in each run (e.g. the first document retrieved by cmuFuTop10D is relevant to intent i_4 with a relevance rating 1); The last four columns show the 's for each query (e.g. score of cmuFuTop10D minus that of THUIR10DvNov) where arrows indicate which run has higher score under each measure. Note that in this section, the measures are computed for a document cutoff $K = 10$ because we only have space to show top 10 documents in Table 6. We categorize five cases into two classes from the viewpoint of diversity (**Case A-C**) or relevance (**Case D-E**).

In **Case A**, we argue that D♯-nDCG is less intuitive than the other three. THUIR10DvNov covers both "bobcat company" and "wild bobcat" while cmuFuTop10D only covers the former (Please refer to the detailed description for the official intents of No. 77 query shown in Figure 1) although both runs cover three leaf intents. In this sense, THUIR10DvNov is more diversified than cmuFuTop10D and should be preferred. Note that this is also a case where I-rec cannot tell which run is better but N-rec can. The rightmost column of Table 6 shows that only D♯-nDCG disagrees with this intuition. In **Case B**, we argue that D♯-nDCG and HD♯-nDCG are less intuitive than the other two. Similar to **Case A**, UAMSD10aSRfu covers both "bobcat company" and "wild bobcat," whereas THUIR10DvQEW fails to cover the latter. So UAMSD10aSRfu should be preferred, and only LAD♯-nDCG and LD♯-nDCG agree with this. In **Case C**, we argue that LD♯-nDCG is the most intuitive among the four measures. In this case, both msrsv2div and qirdcsuog3 cover "bobcat company" and "wild bobcat". However, Figure 1 shows that msrsv2div covers both "bobcat tractors" and "bobcat company homepage," which are sub intents of

TABLE 6
Five ranked list pairs from TREC Web Track 2009-2013 diversity test collections. 1st column: case IDs (query IDs). 2nd column: run IDs. 3rd column: number of official intents covered by each run. 4th column: number of nodes in EIH covered by each run. 5th column: relevance ratings for each intent at ranks 1-10. The rightmost column: performance differences using each measure and arrows point to its preferred run.

				Document rank (i: official intents) 1	2	3	4	5	6	7	8	9	10	Δ in D*J*-nDCG	Δ in LD*J*-nDCG	Δ in HD*J*-nDCG	Δ in LAD*J*-nDCG
A (77)	cmuFuTop10D	3	6	i_4L1				i_3L1					i_1L1	0.0013 *	-0.1098 +	-0.0977 +	-0.0988 +
	THUIR10DvNov	3	8	i_4L1					i_1L1				i_2L1				
B (77)	THUIR10DvQEW	2	5	i_4L1					i_3L1					0.0300 *	-0.0256 +	0.0011 *	-0.0019 +
	UAMSD10aSRfu	2	6						i_4L1			i_2L1					
C (77)	msrsv2div	3	8	i_4L1			i_2L1	i_2L1	i_2L1			i_3L1	i_2L1	-0.0329 +	0.0226 *	-0.0115 +	-0.0085 +
	qirdcsuog3	3	7	i_3L1			i_1L1 i_3L1	i_1L1		i_1L1 i_3L1	i_1L1	i_2L1					
D (117)	qutir11a	3	5	i_1L1 i_2L1	i_2L1	i_2L1 i_3L1	i_1L1 i_3L1	i_1L2 i_2L1 i_3L1	i_2L1	i_1L2 i_2L2	i_1L3 i_2L1	i_2L1	i_2L1	-0.0030 +	-0.0030 +	0.0171 *	0.0148 *
	uwBBadhoc	3	5	i_1L3 i_2L2 i_3L3	i_3L1								i_1L1 i_2L2 i_3L1				
E (128)	2011SiftR2	3	5	i_1L2 i_2L2 i_3L1			i_1L2 i_3L2							0.0087 *	0.0087 *	-0.0005 +	0.0004 *
	UWatMDSdm	3	5	i_1L1 i_2L1	i_1L1		i_1L1 i_3L1		i_1L1	i_1L2 i_2L2			i_1L2 i_2L2				

TABLE 7
Kendall's / Symmetric $_{ap}$ by averaging over TREC Web track 2009-2013 or NTCIR-11 IMine. Values ($\geq$ .950) are shown in bold.

(A) TREC Web Track 2009-2013 diversity test collections					
	ERR-IA	D*J*-nDCG	LD*J*-nDCG	HD*J*-nDCG	LAD*J*-nDCG
-nDCG	.923/.870	.840/.796	.845/.796	.843/.792	.844/.793
ERR-IA	-	.772/.699	.780/.706	.779/.704	.779/.706
D*J*-nDCG	-	-	:**976**/:**959**	:**976**/:**957**	:**977**/:**960**
LD*J*-nDCG	-	-	-	:**991**/:**988**	:**995**/:**993**
HD*J*-nDCG	-	-	-	-	:**995**/:**994**
(B) NTCIR-11 IMine test collection					
	ERR-IA	D*J*-nDCG	LD*J*-nDCG	HD*J*-nDCG	LAD*J*-nDCG
-nDCG	.869/.763	.899/.922	.825/.843	.835/.878	.835/.878
ERR-IA	-	.787/.767	.713/.704	.723/.728	.723/.728
D*J*-nDCG	-	-	.926/.913	.936/.949	.936/.949
LD*J*-nDCG	-	-	-	:**990**/:**964**	:**990**/:**964**
HD*J*-nDCG	-	-	-	-	:**999**/:**999**

"bobcat company," while qirdcsuog3 does not cover "bobcat company homepage." Because of this, msrsv2div should be preferred and only LD♯-nDCG agrees with this.

In summary, from the viewpoint of diversity, LD♯-nDCG is the most intuitive measure. HD♯-nDCG is less intuitive than LAD♯-nDCG, but is more intuitive than D♯-nDCG.

The two runs in **Case D** and in **Case E** have the same I-rec and N-rec, hence the measures' preference is determined by their Precision part (e.g. D-nDCG if it is D♯-nDCG, and HD-nDCG if it is HD♯-nDCG). In **Case D**, we argue that D♯-nDCG and LD♯-nDCG are less intuitive than the other two. No matter whether measuring by I-rec or by N-rec, qutir11a and uwBBadhoc are equally good in terms of diversity. However, qutir11a should be preferred because its top ten documents are all relevant, whereas uwBBadhoc only has three. From the rightmost column of Table 6, we find that D♯-nDCG and LD♯-nDCG fail to reflect this. In **Case E**, we argue that HD♯-nDCG is the most intuitive. UWatMDSdm should be preferred because it returns much more relevant documents than 2011SiftR2. In this case, only HD♯-nDCG successfully recognizes this.

Generally, from the viewpoint of relevance, LAD♯-nDCG is more intuitive than LD♯-nDCG. LAD♯-nDCG is able to measure the relevance of ranked lists more accurately by considering the whole intent hierarchy, and thus make the measures more consistent with Precision than LD♯-nDCG.

Case A serves as a good example showing that more diverse search results score higher under by hierarchical measures: THUIR10DvNov that covers both "bobcat company" and "wild bobcat" has higher score than cmuFuTop10D that only covers "bobcat company" under by LD♯-nDCG, HD♯-nDCG, as well as LAD♯-nDCG. Hierarchical measures achieve this by considering intents i_1, i_3, and i_4 are more related than intent i_2 (see Figure 1 for the descriptions of the intents). Intuitively, hierarchical measures consider search result diversity as covering as many intents that are less related to each other as possible. This is the main difference between hierarchical measures and existing measures, which considers search result diversity as covering as many intents as possible

4.2.4 Rank Correlation Results

We compute Kendall's τ and τ_{ap} for different pairs of measures, and the results are shown in Table 7. We find that: (1) LD♯-nDCG, HD♯-nDCG, and LAD♯-nDCG are more correlated to D♯-nDCG than α-nDCG and ERR-IA. This is because they are different kinds of extensions of D♯-nDCG. Similar to D♯-nDCG, they model diversity and relevance in different components separately. They yield the same evaluation results when the queries only have single-layer intent hierarchies; (3) LD♯-nDCG and HD♯-nDCG are less correlated. As discussed in Section 4.2.3, LD♯-nDCG prefers highly diversified ranked lists, whereas HD♯-nDCG prefers highly relevant ranked lists.

4.3 Using EIH is Better than Using OIH

The key factor influencing the performance of hierarchical measures is using OIH or using EIH. We have discussed some drawbacks of using OIH in Section 3.4. Since the intent hierarchies in NTCIR-11 IMine test collection have already satisfied the five properties of EIH, we only use TREC Web Track 2009-2013 test collections to compare the hierarchical measures using OIH and those using EIH.

TABLE 8
Discriminative power of diversity measures based on the paired bootstrap test at = 0:05 on TREC Web Track 2009-2013 test collections.

OIH uniform top-down		EIH uniform top-down		OIH uniform bottom-up		EIH uniform bottom-up	
measure	disc.power	measure	disc.power	measure	disc.power	measure	disc.power
-nDCG-LA$_{OUT}$	58.21%	-nDCG-LA$_{EUT}$	**58**:**32**%	-nDCG-LA$_{OUB}$	58.21%	-nDCG-LA$_{EUB}$	**58**:**32**%
ERR-IA-LA$_{OUT}$	52.32%	ERR-IA-LA$_{EUT}$	**53**:**79**%	ERR-IA-LA$_{OUB}$	52.32%	ERR-IA-LA$_{EUB}$	**53**:**79**%
nDCG-IA-LA$_{OUT}$	54.11%	nDCG-IA-LA$_{EUT}$	**55**:**37**%	nDCG-IA-LA$_{OUB}$	53.79%	nDCG-IA-LA$_{EUB}$	55.16%
Q-IA-LA$_{OUT}$	48.21%	Q-IA-LA$_{EUT}$	48.95%	Q-IA-LA$_{OUB}$	47.89%	Q-IA-LA$_{EUB}$	**49**:**16**%
D*J*-nDCG-LA$_{OUT}$	56.32%	D*J*-nDCG-LA$_{EUT}$	**57**:**47**%	D*J*-nDCG-LA$_{OUB}$	55.68%	D*J*-nDCG-LA$_{EUB}$	57.37%
D*J*-Q-LA$_{OUT}$	54.42%	D*J*-Q-LA$_{EUT}$	56.32%	D*J*-Q-LA$_{OUB}$	54.53%	D*J*-Q-LA$_{EUB}$	**56**:**53**%
LD*J*-nDCG$_{OUT}$	57.16%	LD*J*-nDCG$_{EUT}$	**57**:**58**%	LD*J*-nDCG$_{OUB}$	56.63%	LD*J*-nDCG$_{EUB}$	57.37%
HD*J*-nDCG$_{OUT}$	56.53%	HD*J*-nDCG$_{EUT}$	**57**:**26**%	HD*J*-nDCG$_{OUB}$	56.21%	HD*J*-nDCG$_{EUB}$	**57**:**26**%
LAD*J*-nDCG$_{OUT}$	56.53%	LAD*J*-nDCG$_{EUT}$	**57**:**26**%	LAD*J*-nDCG$_{OUB}$	56.00%	LAD*J*-nDCG$_{EUB}$	**57**:**26**%
LD*J*-Q$_{OUT}$	55.79%	LD*J*-Q$_{EUT}$	56.42%	LD*J*-Q$_{OUB}$	55.68%	LD*J*-Q$_{EUB}$	**56**:**63**%
HD*J*-Q$_{OUT}$	55.79%	HD*J*-Q$_{EUT}$	**56**:**53**%	HD*J*-Q$_{OUB}$	55.47%	HD*J*-Q$_{EUB}$	**56**:**53**%
LAD*J*-Q$_{OUT}$	54.42%	LAD*J*-Q$_{EUT}$	**56**:**53**%	LAD*J*-Q$_{OUB}$	54.42%	LAD*J*-Q$_{EUB}$	**56**:**53**%

TABLE 9
Intuitiveness results on TREC Web Track 2009-2013 test collections. For each measure pair (one measure using OIH and the other using EIH), the higher score is shown in bold.

	Uniform top-down		Uniform bottom-up	
	OIH	EIH	OIH	EIH
(A) Gold standard measure: I-rec				
-nDCG-LA	1.000	1.000	1.000	1.000
ERR-IA-LA	.663	:**729**	.663	:**724**
nDCG-IA-LA	.669	:**724**	624	:**748**
Q-IA-LA	.662	:**698**	.653	:**696**
D*J*-nDCG-LA	.788	:**836**	.710	:**867**
D*J*-Q-LA	.760	:**761**	.758	:**776**
LD*J*-nDCG	.826	:**846**	.781	:**865**
HD*J*-nDCG	.809	:**856**	.735	:**882**
LAD*J*-nDCG	.827	:**852**	.753	:**886**
LD*J*-Q	.738	:**801**	.725	:**793**
HD*J*-Q	.750	:**793**	.739	:**795**
LAD*J*-Q	:**814**	.761	:**815**	.774
(B) Gold standard measure: Precision				
-nDCG-LA	1.000	1.000	1.000	1.000
ERR-IA-LA	.539	:**655**	.539	:**655**
nDCG-IA-LA	.593	:**660**	472	:**744**
Q-IA-LA	.610	:**659**	.510	:**743**
D*J*-nDCG-LA	.562	:**659**	.520	:**666**
D*J*-Q-LA	.509	:**718**	.496	:**730**
LD*J*-nDCG	:**617**	.614	.557	:**672**
HD*J*-nDCG	.577	:**655**	.517	:**668**
LAD*J*-nDCG	.569	:**655**	.511	:**676**
LD*J*-Q	.631	:**633**	.602	:**631**
HD*J*-Q	.620	:**649**	.584	:**648**
LAD*J*-Q	.500	:**761**	.483	:**754**
(C) Gold standard measure: I-rec and Precision				
-nDCG-LA	1.000	1.000	1.000	1.000
ERR-IA-LA	.349	:**481**	.349	:**481**
nDCG-IA-LA	.365	:**460**	247	:**534**
Q-IA-LA	.379	:**451**	.304	:**510**
D*J*-nDCG-LA	.390	:**510**	.298	:**545**
D*J*-Q-LA	.298	:**486**	.285	:**510**
LD*J*-nDCG	.462	:**473**	.366	:**545**
HD*J*-nDCG	.406	:**520**	.296	:**556**
LAD*J*-nDCG	.415	:**526**	.307	:**568**
LD*J*-Q	.374	:**436**	.333	:**426**
HD*J*-Q	.375	:**444**	.330	:**445**
LAD*J*-Q	.323	:**503**	.309	:**529**

4.3.1 Discriminative Power

On TREC test collections, we apply the same method as described in Section 4.2.1 to compute the discriminative power of hierarchical measures using four types of intent hierarchies, i.e. OUT, EUT, OUB, and EUB (shown in Table 1(A)). The results of discriminative power are shown in Table 8. By comparing the first column with the second column, and comparing the third column with the fourth column in Table 8, we find that the hierarchical measures using EIH are consistently more discriminative than those using OIH. This finding holds true no matter which weighting scheme is used (UT or UB).

4.3.2 Concordance Test

To compare the intuitiveness of hierarchical measures using OIH and those using EIH, we use I-rec and Precision as the gold standard measures. This is because: (1) They are simple binary measures, and are traditionally used as the gold standard measures; (2) They do not depend on intent hierarchies being OIH or EIH. We use all the queries in TREC test collections to compute the intuitiveness, and show the results in Table 9. By comparing the first column with the second column, and comparing the third column with the fourth column in Table 9, we find that the hierarchical measures using EIH are mostly more intuitive than those using OIH regardless of the weighting scheme used. This is because the hierarchical measures using OIH may reward high relevance to some official intents, and fail to reward wide coverage of the official intents.

To sum up, we find that hierarchical measures using EIH outperform those using OIH in terms of discriminative power and intuitiveness. This means that **hierarchical measures work better with EIH than with OIH**. This is because for an atomic intent, every layer of EIH either include it or one of its ancestors, but this is untrue for some layers of OIH. This makes hierarchical measures bias towards some atomic intents and be counterintuitive sometimes. For this reason, we use EIH when computing hierarchical measures on TREC test collections in Section 4.2.

4.4 Weighting Bottom-up is Better than Weighting Top-down

Another factor that affects the performance of hierarchical measures is weighting intent hierarchies top-down or bottom-up. We conduct the experiments on TREC Web Track 2009-2013 test collections and NTCIR-11 IMine test collection, and find that the conclusions are consistent.

4.4.1 Discriminative Power

We compute the discriminative power of hierarchical measures using different types of intent hierarchies on TREC Web Track test collections and NTCIR-11 IMine test collection, and show the results in Tables 8 and 10. We find that: (1) Hierarchical measures using OIH weighted top-down are mostly more discriminative than those using OIH weighted bottom-up; (2) Hierarchical measures using EIH weighted top-down are generally less discriminative than those using EIH weighted bottom-up.

TABLE 10
Discriminative power of diversity measures based on the paired bootstrap test at = 0:05 on NTCIR-11 IMine test collection.

EIH uniform top-down		EIH nonuniform top-down		EIH uniform bottom-up		EIH nonuniform bottom-up	
measure	disc.power	measure	disc.power	measure	disc.power	measure	disc.power
-nDCG-LA$_{EUT}$	**63**:**13**%	-nDCG-LA$_{ENT}$	**63**:**13**%	-nDCG-LA$_{EUB}$	**63**:**13**%	-nDCG-LA$_{ENB}$	**63**:**13**%
ERR-IA-LA$_{EUT}$	**65**:**00**%	ERR-IA-LA$_{ENT}$	**65**:**00**%	ERR-IA-LA$_{EUB}$	**65**:**00**%	ERR-IA-LA$_{ENB}$	**65**:**00**%
nDCG-IA-LA$_{EUT}$	70.63%	nDCG-IA-LA$_{ENT}$	71.25%	nDCG-IA-LA$_{EUB}$	72.50%	nDCG-IA-LA$_{ENB}$	**73**:**13**%
Q-IA-LA$_{EUT}$	69.38%	Q-IA-LA$_{ENT}$	68.75%	Q-IA-LA$_{EUB}$	**71**:**88**%	Q-IA-LA$_{ENB}$	**71**:**88**%
D$\sharp$-nDCG-LA$_{EUT}$	68.75%	D$\sharp$-nDCG-LA$_{ENT}$	69.38%	D$\sharp$-nDCG-LA$_{EUB}$	**70**:**00**%	D$\sharp$-nDCG-LA$_{ENB}$	69.38%
D$\sharp$-Q-LA$_{EUT}$	70.63%	D$\sharp$-Q-LA$_{ENT}$	70.63%	D$\sharp$-Q-LA$_{EUB}$	**71**:**25**%	D$\sharp$-Q-LA$_{ENB}$	**71**:**25**%
LD$\sharp$-nDCG$_{EUT}$	68.75%	LD$\sharp$-nDCG$_{ENT}$	68.75%	LD$\sharp$-nDCG$_{EUB}$	**69**:**38**%	LD$\sharp$-nDCG$_{ENB}$	68.75%
HD$\sharp$-nDCG$_{EUT}$	**69**:**38**%	HD$\sharp$-nDCG$_{ENT}$	68.75%	HD$\sharp$-nDCG$_{EUB}$	**69**:**38**%	HD$\sharp$-nDCG$_{ENB}$	**69**:**38**%
LAD$\sharp$-nDCG$_{EUT}$	**69**:**38**%	LAD$\sharp$-nDCG$_{ENT}$	68.75%	LAD$\sharp$-nDCG$_{EUB}$	**69**:**38**%	LAD$\sharp$-nDCG$_{ENB}$	**69**:**38**%
LD$\sharp$-Q$_{EUT}$	70.63%	LD$\sharp$-Q$_{ENT}$	70.63%	LD$\sharp$-Q$_{EUB}$	**71**:**25**%	LD$\sharp$-Q$_{ENB}$	**71**:**25**%
HD$\sharp$-Q$_{EUT}$	**71**:**25**%	HD$\sharp$-Q$_{ENT}$	**71**:**25**%	HD$\sharp$-Q$_{EUB}$	**71**:**25**%	HD$\sharp$-Q$_{ENB}$	**71**:**25**%
LAD$\sharp$-Q$_{EUT}$	**71**:**25**%	LAD$\sharp$-Q$_{ENT}$	**71**:**25**%	LAD$\sharp$-Q$_{EUB}$	**71**:**25**%	LAD$\sharp$-Q$_{ENB}$	**71**:**25**%

TABLE 11
Discriminative power of diversity measures based on the paired bootstrap test at = 0:05 on NTCIR-11 IMine test collection.

EIH uniform top-down		EIH nonuniform top-down		EIH uniform bottom-up		EIH nonuniform bottom-up	
measure	disc.power	measure	disc.power	measure	disc.power	measure	disc.power
D-nDCG$_{EUT}$	77.50%	D-nDCG$_{ENT}$	77.50%	D-nDCG$_{EUB}$	77.50%	D-nDCG$_{ENB}$	77.50%
HD-nDCG$_{EUT}$	77.50%	HD-nDCG$_{ENT}$	**78**:**13**%	HD-nDCG$_{EUB}$	76.88%	HD-nDCG$_{ENB}$	76.25%
D-nDCG-LA$_{EUT}$	76.88%	D-nDCG-LA$_{ENT}$	**78**:**13**%	D-nDCG-LA$_{EUB}$	76.88%	D-nDCG-LA$_{ENB}$	77.50%
D-Q$_{EUT}$	81.88%	D-Q$_{ENT}$	81.88%	D-Q$_{EUB}$	81.88%	D-Q$_{ENB}$	81.88%
HD-Q$_{EUT}$	**81**:**88**%	HD-Q$_{ENT}$	**81**:**88**%	HD-Q$_{EUB}$	81.25%	HD-Q$_{ENB}$	81.25%
D-Q-LA$_{EUT}$	**81**:**88**%	D-Q-LA$_{ENT}$	**81**:**88**%	D-Q-LA$_{EUB}$	81.25%	D-Q-LA$_{ENB}$	81.25%

4.4.2 Concordance Test

To compare the intuitiveness of hierarchical measures using intent hierarchies weighted top-down or bottom-up, I-rec and Precision are used as the gold standard measures for the same reasons as mentioned in 4.3.2. We use all the queries in TREC test collections or all the queries the NTCIR test collection to compute the intuitiveness. The results are shown in Table 12. By comparing the first column with the second column in Table 12, we find that hierarchical measures using OIH weighted top-down are generally more intuitive than those using OIH weighted bottom-up. However, by comparing the third column with the fourth column, comparing the fifth column with the sixth column, and comparing the seventh column with the eighth column in Table 12, we find that hierarchical measures using EIH weighted top-down are generally less intuitive than those using EIH weighted bottom-up.

To summarize, it is preferable for OIH to be weighted top-down, whereas **it is preferable for EIH to be weighted bottom-up.** This is due to that when weighting EIH top-down, the leaves of a subtree that has many leaves will end up with minor weights, and may be neglected by hierarchical measures. Because of this, intent hierarchies are weighted bottom-up on NTCIR-11 IMine test collection in Section 4.2 for they are EIH. Sections 4.3 and 4.4 suggest that when nonuniform intent weights are not available, it is preferable for hierarchical measures to use EIH weighted bottom-up. Since TREC Web Track 2009-2013 do not provide nonuniform intent weights, hierarchical measures use EIH weighted bottom-up on the test collections in Section 4.2.

4.5 Benefits of Using Nonuniform Weights

LD$\sharp$-measures, HD$\sharp$-measures, and LAD$\sharp$-measures use N-rec to model diversity. Since N-rec does not take weights into consideration, weighting schemes have no effect on its discriminative power. D-measures, HD-measures, and D-measures-LA model relevance and consider weights. We compare their discriminative power on NTCIR-11 IMine test collections with different weighting schemes, and show the results in Table 11. By comparing the first column with the second column and comparing the third column and the fourth column in Table 11, we find that **D-measures, HD-measures, and D-measures-LA using nonuniform weights tend to have higher discriminative power than those using uniform weights.**

4.6 Hierarchical Measures and Diversification Algorithms

Hu et al. [32] proposes two hierarchical diversification algorithms HxQuAD and HPM2 that use hierarchical intents. These two algorithms are demonstrated to outperform traditional diversification algorithms xQuAD, PM2, TxQuAD, and TPM2 that use a flat list of intents. Hu et al. [32] automatically generates hierarchical intents whose height is two. xQuAD, PM2, TxQuAD, and TPM2 can use the first-layer intents only, the second-layer intents only, or all the intents in the first-layer and second-layer, which results in 12 different algorithms. Given two measures M_1 and M_2, the percentage of run pairs that are significantly different in the total run pairs is computed as follows: (1) The topic set in TREC Web Track 2009-2012 diversity test collections is randomly partitioned into five equal sized subsets. We fix the specific partition of the original topic set afterwards. For each of the five subsets, the remaining four subsets are used as the training data, and M_1 is used to decide the best configuration of an algorithm in the training data, which gives us the final run for the spared subset. In this way, we produce a final run for each of the two hierarchical algorithms and 12 traditional algorithms; (2) We use M_2 to evaluate the final runs produced before, one for each algorithm. Two-tailed paired t-test with significant level equal 0.05 is then used to test whether the run of a hierarchical algorithm

TABLE 12
Intuitiveness results on TREC Web Track 2009-2013 and NTCIR-11 IMine. For each measure pair, the higher score is shown in bold.

	TREC OIH uniform		TREC EIH uniform		NTCIR-11 EIH uniform		NTCIR-11 EIH nonuniform	
	Top-down	Bottom-up	Top-down	Bottom-up	Top-down	Bottom-up	Top-down	Bottom-up
(A) Gold standard measure: I-rec								
-nDCG-LA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ERR-IA-LA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
nDCG-IA-LA	**0:775**	0.613	**0:727**	0.685	0.653	**0:721**	0.669	**0:673**
Q-IA-LA	**0:731**	0.651	**0:722**	0.667	0.654	**0:681**	**0:681**	0.620
D*J*-nDCG-LA	**0:967**	0.641	0.814	**0:877**	0.636	**0:821**	0.753	**0:771**
D*J*-Q-LA	**0:855**	0.825	0.727	**0:840**	0.818	**0:864**	0.864	**0:909**
LD*J*-nDCG	**0:894**	0.847	0.821	**0:893**	0.565	**0:768**	0.736	**0:753**
HD*J*-nDCG	**0:961**	0.689	0.840	**0:871**	0.705	**0:795**	**0:813**	0.731
LAD*J*-nDCG	**0:966**	0.669	0.829	**0:881**	0.691	**0:819**	0.779	**0:785**
LD*J*-Q	0.764	**0:867**	0.786	**0:813**	**0:882**	0.647	**0:912**	0.794
HD*J*-Q	0.790	**0:849**	0.768	**0:840**	0.837	0.837	0.800	**0:855**
LAD*J*-Q	**0:884**	0.824	0.775	**0:839**	0.837	0.837	0.843	0.843
(B) Gold standard measure: Precision								
-nDCG-LA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ERR-IA-LA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
nDCG-IA-LA	**0:746**	0.483	0.618	**0:662**	0.620	**0:697**	0.618	**0:687**
Q-IA-LA	**0:780**	0.497	0.634	**0:663**	0.598	**0:657**	0.576	**0:629**
D*J*-nDCG-LA	**0:644**	0.526	0.633	**0:638**	0.649	**0:709**	0.620	**0:753**
D*J*-Q-LA	**0:805**	0.468	**0:740**	0.711	0.864	**0:886**	0.864	**0:886**
LD*J*-nDCG	**0:739**	0.505	0.566	**0:661**	0.464	**0:761**	0.545	**0:781**
HD*J*-nDCG	**0:667**	0.506	0.635	**0:647**	**0:705**	0.659	0.684	**0:720**
LAD*J*-nDCG	**0:652**	0.513	0.635	**0:647**	0.691	**0:698**	0.674	**0:727**
LD*J*-Q	**0:760**	0.613	0.671	**0:731**	0.412	**0:765**	0.735	**0:824**
HD*J*-Q	**0:807**	0.495	**0:739**	0.712	**0:857**	0.816	**0:836**	0.782
LAD*J*-Q	**0:790**	0.485	**0:733**	0.725	**0:857**	0.816	**0:863**	0.804
(C) Gold standard measure: I-rec and Precision								
-nDCG-LA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ERR-IA-LA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
nDCG-IA-LA	**0:554**	0.251	**0:436**	0.429	0.407	**0:502**	0.429	**0:502**
Q-IA-LA	**0:578**	0.307	**0:460**	0.428	0.413	**0:492**	0.437	**0:445**
D*J*-nDCG-LA	**0:613**	0.257	0.467	**0:519**	0.377	**0:543**	0.446	**0:566**
D*J*-Q-LA	**0:660**	0.310	0.487	**0:551**	0.727	**0:750**	0.750	**0:795**
LD*J*-nDCG	**0:638**	0.370	0.408	**0:559**	0.152	**0:536**	0.354	**0:567**
HD*J*-nDCG	**0:629**	0.268	0.486	**0:521**	0.438	**0:472**	**0:518**	0.497
LAD*J*-nDCG	**0:618**	0.263	0.476	**0:530**	0.403	**0:523**	0.477	**0:552**
LD*J*-Q	**0:524**	0.483	0.458	**0:543**	0.353	**0:471**	0.647	0.647
HD*J*-Q	**0:597**	0.359	0.508	**0:552**	**0:694**	0.653	0.636	0.636
LAD*J*-Q	**0:673**	0.314	0.510	**0:564**	**0:694**	0.653	**0:706**	0.647

is significantly better than that of a traditional algorithm. The significance test is performed between a hierarchical algorithm and a traditional algorithm, which gives us 24 run pairs in total. Finally, we compute the percentage of run pairs, which have significant difference, in the 24 run pairs for the measures M_1 and M_2. The results are shown in Table 13, where the measures in the leftmost column are used as training measures and the measures in the topmost row are used as validation measures. We find that: (1) No matter which measure is used to tune parameters, it is easier to show the significant improvements of hierarchical algorithms when using hierarchical measures rather than D-nDCG. This means that in some cases, existing measures cannot find the advantages of hierarchical algorithms but hierarchical measures can; (2) α-nDCG remains a good choice to tune parameters of hierarchical algorithms.

TABLE 13
The percentage of run pairs that pass the two-tailed paired t-test with significant level equal 0.05 when comparing the runs by hierarchical algorithms and those by traditional algorithms. The leftmost measures are used to tune algorithm parameters in the training data, and the resulting final runs are evaluated by the topmost measures.

	-nDCG	ERR-IA	D*J*-nDCG	LD*J*-nDCG	HD*J*-nDCG	LAD*J*-nDCG
-nDCG	66.67%	58.33%	58.33%	62.50%	62.50%	62.50%
ERR--IA	66.67%	66.67%	33.33%	45.83%	45.83%	45.83%
D*J*-nDCG	29.17%	12.50%	70.83%	75.00%	75.00%	75.00%
LD*J*-nDCG	50.00%	25.00%	62.50%	70.83%	70.83%	70.83%
HD*J*-nDCG	33.33%	12.50%	70.83%	75.00%	70.83%	70.83%
LAD*J*-nDCG	33.33%	12.50%	70.83%	75.00%	75.00%	70.83%

5 Conclusions And Future Work

In this paper, we argued that flat lists are not expressive enough to model the relationships between user intents. In view of this, we introduced intent hierarchies with four different weighting schemes. Then we proposed hierarchical measures that could work with intent hierarchies, and illustrated their advantages over existing measures. We experimented with a new test collection based on TREC Web Track 2009-2013 diversity test collections, and the NTCIR-11 IMine test collection. Our main experimental findings are: (1) Hierarchical measures can be more discriminative than existing measures which use flat lists of intents; (2) LD♯-nDCG should be used when the diversity of search results is more valued than the relevance, whereas HD♯-nDCG should be used when the relevance is more important. LAD♯-nDCG is a better choice when diversity and relevance are equally important; (3) The performance of hierarchical measures depends on the types of intent hierarchies. When nonuniform weights are unavailable, it is preferable for hierarchical measures to use EIH weighted bottom-up uniformly. When nonuniform weights are available, it is preferable for hierarchical measures to use EIH weighted bottom-up nonuniformly; (4) The gain of using hierarchical intents to

diversify search results may be vague to existing measures based on flat lists. It is important to evaluate hierarchical diversification algorithms using hierarchical measures.

ACKNOWLEDGMENTS

Zhicheng Dou is the corresponding author of the paper. This work was supported by the National Natural Science Foundation of China under Grant No. 61502501 and No. 61402155, and the National Key Basic Research Program (973 Program) of China under Grant No. 2014CB340403.

REFERENCES

[1] Z. Dou, R. Song, and J.-R. Wen, "A large-scale evaluation and analysis of personalized search strategies," in *WWW*, 2007.
[2] B. J. Jansen, A. Spink, and T. Saracevic, "Real life, real users, and real needs: a study and analysis of user queries on the web," *Information Processing & Management*, 2000.
[3] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, "Analysis of a very large web search engine query log," *SIGIR Forum*, 1999.
[4] C. L. Clarke, M. Kolla, and O. Vechtomova, "An effectiveness measure for ambiguous and underspecified queries," in *ICTIR*, 2009.
[5] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in *WSDM*, 2009.
[6] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *SIGIR*, 1998.
[7] H. Chen and D. R. Karger, "Less is more: probabilistic models for retrieving fewer relevant documents," in *SIGIR*, 2006.
[8] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen, "Multi-dimensional search result diversification," in *WSDM*, 2011.
[9] F. Radlinski and S. Dumais, "Improving personalized web search using result diversification," in *SIGIR*, 2006.
[10] R. L. Santos, C. Macdonald, and I. Ounis, "Intent-aware search result diversification," in *SIGIR*, 2011.
[11] X. Zhu, A. B. Goldberg, J. Van Gael, and D. Andrzejewski, "Improving diversity in ranking using absorbing random walks." in *HLT-NAACL*, 2007.
[12] R. L. Santos, C. Macdonald, and I. Ounis, "Exploiting query reformulations for web search result diversification," in *WWW*, 2010.
[13] V. Dang and W. B. Croft, "Diversity by proportionality: an election-based approach to search result diversification," in *SIGIR*, 2012.
[14] V. Dang and B. W. Croft, "Term level search result diversification," in *SIGIR*, 2013.
[15] T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C.-Y. Lin, "Simple evaluation metrics for diversified search results," in *EVIA*, 2010.
[16] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," in *SIGIR*, 2008.
[17] T. Sakai and R. Song, "Evaluating diversified search results using per-intent graded relevance," in *SIGIR*, 2011.
[18] C. L. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack, "Overview of the trec 2010 web track," in *TREC*, 2010.
[19] X. Wang, Z. Dou, T. Sakai, and J.-R. Wen, "Evaluating search result diversity using intent hierarchies," in *SIGIR*, 2016.
[20] C. X. Zhai, W. W. Cohen, and J. Lafferty, "Beyond independent relevance: methods and evaluation metrics for subtopic retrieval," in *SIGIR*, 2003.
[21] K. Järvelin and J. Kekäläinen, "Ir evaluation methods for retrieving highly relevant documents," in *SIGIR*, 2000.
[22] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *CIKM*, 2009.
[23] C. L. Clarke, N. Craswell, I. Soboroff, and A. Ashkan, "A comparative analysis of cascade measures for novelty and diversity," in *WSDM*, 2011.
[24] T. Sakai, "Evaluating evaluation metrics based on the bootstrap," in *SIGIR*, 2006.
[25] B. A. Carterette, "Multiple testing in statistical analysis of systems-based information retrieval experiments," *TOIS*, 2012.
[26] T. Sakai, "Evaluation with informational and navigational intents," in *WWW*, 2012.
[27] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, 1938.
[28] E. Yilmaz, J. A. Aslam, and S. Robertson, "A new rank correlation coefficient for information retrieval," in *SIGIR*, 2008.
[29] C. L. Clarke, N. Craswell, and I. Soboroff, "Overview of the trec 2009 web track," in *TREC*, 2009.
[30] T. Sakai, "Bootstrap-based comparisons of ir metrics for finding one relevant document," in *AIRS*, 2006.
[31] T. Sakai and S. Robertson, "Modelling a user population for designing information retrieval metrics," in *EVIA*, 2008.
[32] S. Hu, Z. Dou, X. Wang, T. Sakai, and J.-R. Wen, "Search result diversification based on hierarchical intents," in *CIKM*, 2015.

Xiaojie Wang is a Ph.D. candidate in the School of Computing and Information Systems at The University of Melbourne, Australia. He received his B.S degrees in Applied Mathematics and Computer Science from Renmin University of China in 2016. His research interests include information retrieval, data mining, and machine learning.

Ji-Rong Wen is a Professor and the Dean in the School of Information at Renmin University of China (RUC). He was awarded the prestigious National 1000 Plan Expert of China. He is the director of Big Data Research Centre and is leading Big Data Analytics and Intelligence Lab in RUC. His main research interest lies on big data management, information retrieval (especially web search), data mining and machine learning.

Zhicheng Dou is an Associate Professor in the School of Information at Renmin University of China. He received his B.S. and Ph.D. degrees in Computer Science from the Nankai University in 2003 and 2008, respectively. He worked at Microsoft Research as a researcher from July 2008 to September 2014. His research interests include natural language processing, information retrieval, and data mining.

Tetsuya Sakai is a professor at the Department of Computer Science and Engineering, Waseda University, Japan. He is also an Associate Dean at the IT Strategies Division of Waseda, and a visiting professor at National Institute of Informatics. He obtained a Ph.D from Waseda in 2000. His previous employment includes Toshiba and Microsoft Research Asia.

Rui Zhang is a Professor and Reader in the School of Computing and Information Systems at The University of Melbourne, Australia. His research interest is big data and information management in general, particularly in areas of intent tracking, recommendation systems, spatial and temporal data analytics, moving object management, indexing techniques, high performance computing and data streams.