Evaluating Search Result Diversity using Intent Hierarchies

Xiaojie Wang[1,2], Zhicheng Dou[*,1,2], Tetsuya Sakai[3], and Ji-Rong Wen[1,2,4]
[1]School of Information, Renmin University of China
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods, China
[3]Department of Computer Science and Engineering, Waseda University
[4]Key Laboratory of Data Engineering and Knowledge Engineering, MOE, China
[2]{wangxiaojie,dou}@ruc.edu.cn, [3]tetsuyasakai@acm.org, [4]jirong.wen@gmail.com

ABSTRACT

Search result diversification aims at returning diversified document lists to cover different user intents for ambiguous or broad queries. Existing diversity measures assume that user intents are independent or exclusive, and do not consider the relationships among the intents. In this paper, we introduce intent hierarchies to model the relationships among intents. Based on intent hierarchies, we propose several hierarchical measures that can consider the relationships among intents. We demonstrate the feasibility of hierarchical measures by using a new test collection based on TREC Web Track 2009-2013 diversity test collections. Our main experimental findings are: (1) Hierarchical measures are generally more discriminative and intuitive than existing measures using flat lists of intents; (2) When the queries have multilayer intent hierarchies, hierarchical measures are less correlated to existing measures, but can get more improvement in discriminative power; (3) Hierarchical measures are more intuitive in terms of diversity or relevance. The hierarchical measures using the whole intent hierarchies are more intuitive than only using the leaf nodes in terms of diversity and relevance.

Keywords

Ambiguity; Diversity; Evaluation; Novelty; Hierarchy

1. INTRODUCTION

Nowadays, people tend to meet their daily information needs by typing keywords into search engines like Google and Bing. However, these keywords, also known as queries, are often ambiguous or broad [14, 15, 28, 10]. 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 information needs. This poses a challenge to search engines when the targeted user intent cannot be known in advance.

To tackle this problem, a wide range of search result diversification algorithms ([1, 2, 5, 13, 18, 26, 27, 31, 25, 12, 11]) have been proposed over the past years. They aim at returning a diversified ranked document list that covers different intents of the queries. In the meantime, some researchers have introduced a variety of diversity measures, such as I-rec [22], α-nDCG [9], Intent-Aware measures [1], D♯-measures [24], etc. These measures evaluate ranked lists in terms of both diversity and relevance, and indicate which diversification algorithms are better to use. Existing diversity measures assume that the users' information need could be represented by a single layer of intents and these intents are either independent or exclusive. However, some of the intents are not independent and are related to each other.

We use the query "bobcat", which is a topic (No. 77) in Text Retrieval Conference(TREC) 2010 Web Track [8], 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 that 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 trade involving tractors of the company, whereas i_4 is about homepage the company. We argue that this kind of relationships among 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 ranks. 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 [22] 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 3.3.1. We think that this is due to their lack of recognition of the relationships among intents.

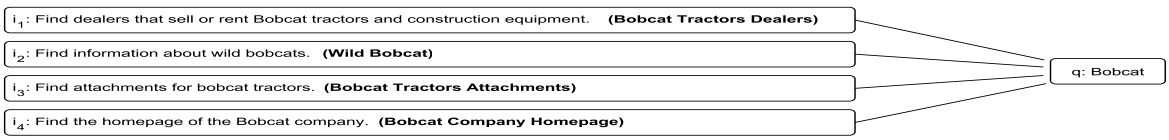
In light of the above observation, we introduce intent hierarchies to represent the relationships among intents. We

*Corresponding author

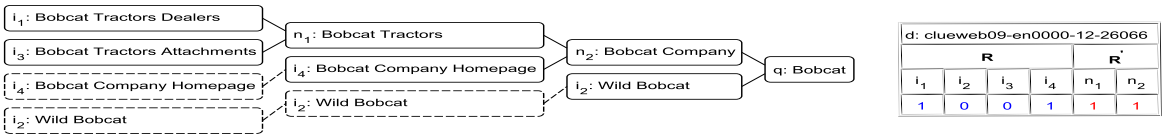
SIGIR '16, July 17 - 21, 2016, Pisa, Italy

ISBN 978-1-4503-4069-4/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2911451.2911497



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



(b) LEFT: Intent Hierarchies OIH and EIH. 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 $\mathbf{R}$ in blue).

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

design hierarchical measures using the intent hierarchies to solve the problems mentioned above. The main contributions of this paper are:

(1) To the best of our knowledge, this is the first work on modeling user intents as intent hierarchies and using the intent hierarchies for evaluating search result diversity.

(2) We propose hierarchical measures using intent hierarchies, including Layer-Aware measures, N-rec, LD♯-measures, LAD♯-measures, and HD♯-measures. We show several cases where hierarchical measures outperform existing measures in terms of discriminative power and intuitiveness.

(3) We present a method for creating intent hierarchies from existing diversity test collections, and reusing the relevance assessments. We create a new dataset based on the TREC Web Track 2009-2013 diversity test collections. The new dataset can be assessed online [1].

(4) We compare our measures with existing measures. We find that (i) Hierarchical measures are generally more discriminative and intuitive than existing measures, especially when using the intent hierarchies whose leaf nodes have the same depth; (ii) When the queries have multilayer intent hierarchies, hierarchical measures are less correlated to existing measures, but can get more improvement in discriminative power; (iii) The hierarchical measures using the whole intent hierarchies are more intuitive than only using the leaf nodes in terms of diversity and relevance.

The remainder of this paper is organized as follows. Section 2 describes some existing diversity measures and the methods for testing evaluation measures. In Section 3, we introduce intent hierarchies, and our method for creating a new test collection based on TREC Web Track 2009-2013 diversity test collections. We then propose several new diversity measures that can utilize the intent hierarchies. Section 4 describes experimental results and analysis. We conclude our work in Section 5.

2. RELATED WORK

Given a query q, most existing measures evaluate a ranked document list by modeling users' information need as a flat list of intents $\{i\}$. Some measures can handle intent probability $Pr(i|q)$ and graded relevance assessments but some cannot. In this section, we briefly summarize the previous work on designing and testing diversity measures.

[1]http://www.playbigdata.com/dou/heval/

2.1 Diversity Measures

2.1.1 Intent Recall

Intent recall(I-rec) [22], also known as subtopic recall [30] 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-rec* for a certain cutoff K can be expressed as:

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

Note that I-rec does not take the positions of relevant documents into account, 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, α-*nDCG* [9] is defined as:

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

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. α-nDCG tends to disregard unpopular intents and hence can be counterintuitive sometimes [24].

2.1.3 Intent-Aware measures

Intent-Aware measures (*IA measures*) [1] is a general framework to evaluate ranked document lists. Assuming that $\sum_{i\in\{i\}} Pr(i|q) = 1$, *M-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 [16], ERR [4], nERR [7], etc.

2.1.4 D♯-measures

D♯-measures [24] aim to boost intent recall, and to reward documents that are highly relevant to more popular intents. 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 given by:

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

Let $CGG(r) = \sum_{k=1}^{r} GG(k)$, which is the cumulative global gain at rank r. Further, let $GG^*(r)$ and $CGG^*(r)$ denote the global gain and the cumulative global gain respectively 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-nDCG* and *D-Q* at document cutoff K are defined as:

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

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

where R is the number of judged relevant documents. Then *D♯-measure* is defined as:

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

where *D-measure* can be D-nDCG or D-Q, and γ is a parameter controlling the tradeoff between diversity and relevance. D♯-measures are free of the under-normalization problem of α-nDCG and IA measures.

The diversity 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 among intents into consideration*, which is what we aim to deal with in this paper.

2.2 Measure Evaluation

Given a certain significance level, *discriminative power* measures the stability of measures across queries and experiments based on significance tests, e.g. paired bootstrap test [20], Tukey's Honestly Significant Differences(HSD) [3] test, etc. Discriminative power can be used to estimate the performance difference required to achieve statistical significance between two retrieval systems [21].

Concordance test [21] 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.

Rank correlation compares two rankings, which are two ranked system lists in our case. Kendall's τ [17] is a widely-used statistic to measure rank correlation. However τ lacks the property of top heaviness, which means the exchanges near the top of a ranked list and those near the bottom are treated equally, even though the swaps near the top is generally more important in the context of IR evaluation. τ_{ap} [29] 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.

In this paper, we use discriminative power, concordance test, and rank correlation to evaluate diversity measures.

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

[3]http://research.nii.ac.jp/ntcir/index-en.html

3. PROPOSED METHODS

In this section, we define two types of intent hierarchies to represent the relationships among user intents and discuss their properties. We then introduce our method for creating such intent hierarchies and obtaining relevance assessments for the intent hierarchies based on TREC Web Track 2009-2013 diversity test collections. Last, we propose several diversity measures based on intent hierarchies, and demonstrate that in some cases, the new measures outperform their corresponding existing measures.

3.1 Intent Hierarchies

Given a query q, the users' information need is represented as a set of intents $\{i\}$. We assume these intents cannot be further subdivided, and refer to them as *atomic intents*. We aim to build an *intent hierarchy* based on the semantic similarity or relatedness of the intents. The intent hierarchy should possess some basic properties as follows:

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 user's information need as a whole. The root is a dummy node only 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 is one interpretation of the query, 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.

We call an intent hierarchy that satisfies the properties specified above is called an *original intent hierarchy* (OIH). OIH can be extended so as to satisfy an extra property as:

Property 5. These atomic intents are in the same layer of the intent hierarchy. In other words, all leaf nodes of the intent hierarchy have the same depth because the atomic intents correspond to leaves of the intent hierarchy (see Property 4).

An intent hierarchies that satisfies all five properties are called an *extended intent hierarchy* (EIH). If a query's OIH satisfies Property 5, then its EIH is the same as the OIH.

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 Creating Intent Hierarchies

In this paper, we create intent hierarchies based on TREC Web Track 2009-2013 diversity test collections. Note that for each query in TREC Web Track 2010-2013, the description of its first intent is the same as the description of the query itself. We find that 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 relevance of the documents. First we create original intent hierarchies (OIH) by manually grouping the official intents based on their semantic similarity or relatedness. Then, we extend them to extended intent hierarchies (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 tractors of the "bobcat company." 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 more child nodes to i_2 and i_4 to make sure that all the leaf nodes have the same height. The resultant EIH is shown in solid boxes plus dashed boxes in the left of Figure 1(b).

For a leaf node, we use its original weight of the corresponding official intent as its initial weight. For an intermediate node, we set its original weight to the sum of its child node weights. We then normalize the weights for each layer to make sure that these weights sum to 1. For TREC Web Track 2009-2013 test collections, because of the lack of official intent weights, we assume that each official intent for a query is equally important.

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).

Note that we only have document relevance assessments for the original intents appeared in TREC Web Track diversity test collections. In other words, for the intent hierarchies we create, document relevance judgments are just available for their leaf intents. We do not have document relevance assessments for intermediate intents. As assessing document relevance is usually very time-consuming, it is not desirable to reassess the documents for intermediate nodes of the intent hierarchies. 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 derive relevance assessments for the intermediate nodes starting from the leaves. In this paper, we simply let:

$$L_d(n) = \max_{c \in C(n)} L_d(c) \tag{8}$$

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 (8), 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.

The new test collection has 250 queries, and 105 topics have multilayer intent hierarchies. 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 search engines such as Bing and Google.

3.3 Hierarchical Measures

3.3.1 *Layer-Aware measures*

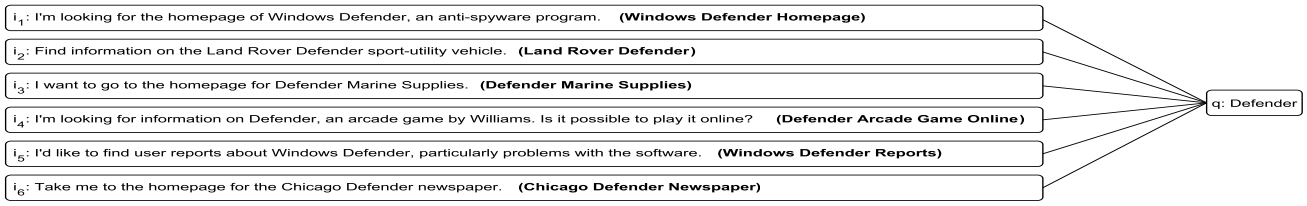
Given a query q and its intent hierarchy, our first proposal for evaluating a ranked list is to first evaluate the ranked list for each layer using existing measures, then combine the evaluation scores.

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

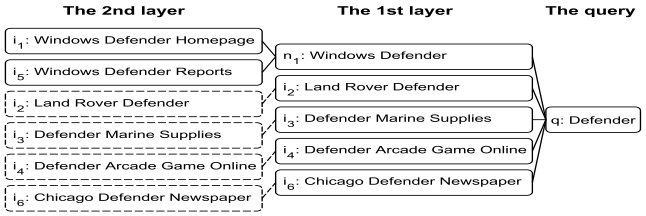
$$M\text{-}LA@K = \sum_{i=1}^{H} w_i * M_i@K \tag{9}$$

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 (9).

We find that the combination of measures over layers of intent hierarchies could outperform the original measures using a flat list of intents. We use the query "defender", which is a topic (No. 20) in TREC Web Track 2009 [6], as an example. We choose this query because it has a relatively simple intent hierarchy. Its extended intent hierarchy (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



(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: Intent Hierarchies OIH and EIH. 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 document is equal to a ranked list of length 1. 2nd to 5th column ($\mathbf{R}$ and $\mathbf{R}'$): relevance assessments for the official intents (in red) and derived relevance assessments for added nodes (in blue). 6th to 12th column ($\mathbf{S}$): the measures are computed at rank 1 (subscript $_1$ means only using the first layer of EIH and subscript $_2$ means only using the second layer. subscript $_O$ means using OIH and subscript $_E$ means using EIH.), e.g. d_2 get 0.35 when using D♯-nDCG on the first layer of EIH. Note that the original D♯-nDCG is equal to D♯-nDCG$_2$ and the original I-rec is equal to I-rec$_2$.

Figure 2: The official intents, original intent hierarchy (OIH), and extended intent hierarchy (EIH) of No. 20 query "defender" in TREC Web Track 2009. In the right table, if two documents have the same scores under a measure (in green below), it means that this measure cannot tell which one is better, e.g. D♯-nDCG$_1$@1 treats d_2 and d_3 as equally good, but d_3 is better because of its relevance to an extra intent i_5.

Figure 2(b). Note that the nodes that receive no relevant documents within the documents are not displayed to save space. Assume 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. Note that the original D♯-nDCG is equal to D♯-nDCG$_2$. We use the measures to score d_1 to d_3, which is equivalent of evaluating at document cutoff 1. We show the evaluation results in Figure 2(b), e.g. d_2 gets 0.35 when using D♯-nDCG$_1$@1.

We find that d_1>d_2=d_3 in terms of D♯-nDCG$_1$@1, d_1=d_2>d_3 in terms of D♯-nDCG$_2$@1, whereas d_1>d_2>d_3 in terms of 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 should be d_1>d_2>d_3. This is because (1) d_1 is more diversified than d_2 because d_1 refers to two interpretations of query "defender", i.e. "windows defender" and "defender arcade game online," while d_2 only refers to the former; (2) d_2 is more diversified than d_3 because d_2 refers to two aspects of "windows defender", i.e. "windows defender homepage" and "windows defender reports" while d_3 just refers to 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 in some cases.

3.3.2 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 set of nodes in V to which d_r is relevant. Given a document cutoff K, we define node recall (N-rec) as:

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

which 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 intent hierarchy rather than a flat list of intents. They both are rank-insensitive and cannot handle graded relevance assessments.

We use an example to show that N-rec is able to outperform 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 the extended intent hierarchy (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 3.3.1, The real preference should be 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.

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 the original intent hierarchy (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. At the same time, we pre-

fer 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 discuss some measures to complement N-rec.

3.3.3 LD♯-measures

We use the leaf nodes of intent hierarchies to compute D-measures. Then, *LD♯-measure* is defined as:

$$LD\sharp\text{-}measure@K = \gamma N\text{-}rec@K + (1-\gamma) D\text{-}measure@K \quad (11)$$

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

3.3.4 HD♯-measures and LAD♯-measures

Inspired by D♯-measures, we define the global gain for an intent hierarchy at rank r as:

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

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-nDCG* and *HD-Q* at document cutoff K as:

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

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

where R is the number of judged documents relevant to the intent hierarchy. We define *HD♯-measure* as:

$$HD\sharp\text{-}measure@K = \gamma N\text{-}rec@K + (1-\gamma) HD\text{-}measure@K \quad (15)$$

where *HD-measure* can be HD-nDCG or HD-Q, and γ is a parameter controlling the tradeoff between diversity and relevance. Besides, We define *LAD♯-measure* as:

$$LAD\sharp\text{-}measure@K = \gamma N\text{-}rec@K + (1-\gamma) D\text{-}measure\text{-}LA@K \quad (16)$$

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:

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

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

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.

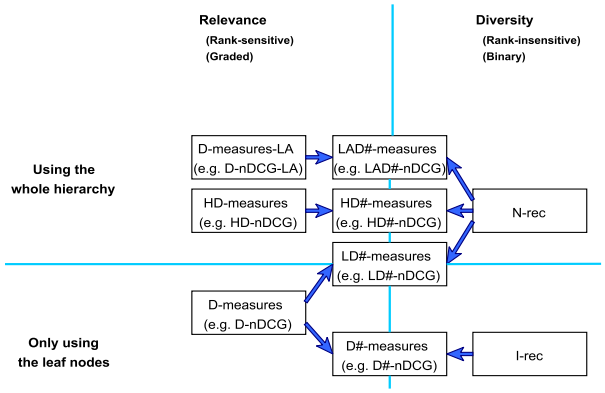


Figure 3: Relationships of D♯-measures, LD♯-measures, HD♯-measures, and LAD♯-measures.

3.3.5 Summarization

Since our measures use intent hierarchies, we call them *hierarchical measures*. Each of D♯-measures, LD♯-measures, HD♯-measures and LAD♯-measures is a linear combination of two measures: one measure mainly rewards the diversity of ranked lists, whereas another measure mainly rewards the relevance. We show their relationships in Figure 3. It can be seen that (1) To reward the diversity, LD♯-measures, HD♯-measures and LAD♯-measures use the whole intent hierarchy, whereas D♯-measures only use the leaf nodes; (2) To reward the relevance, HD♯-measures and LAD♯-measures use the whole intent hierarchy, whereas D♯-measures and LD♯-measures only use the leaf nodes.

4. EXPERIMENTS

4.1 Settings

We experiment with the proposed measures on the TREC Web Track 2009-2013 diversity test collections and the new test collection mentioned in Section 3.2. The new test collection has two types of intent hierarchies, i.e. the original intent hierarchies (OIH), and the extended intent hierarchies (EIH). The results of our measures using OIH are different from those using EIH. In the following, subscript $_O$ means using the OIH, while subscript $_E$ means using the EIH.

We use uniform probabilities for official intents like in TREC Web Track 2009-2013 diversity task. We use uniform layer weights when computing our measures, and leave the investigation of nonuniform weights to future. Unless stated otherwise, we use document cutoff $K = 20$ for all measures, and $\gamma = 0.5$ in Equation (7, 11, 15, and 16).

4.2 Discriminative Power Results

Following the previous work [19, 20, 23, 24, 21], we use the paired bootstrap test and set $B = 1,000$ (B is the number of bootstrap samples). When the queries have single-layer intent hierarchies: (1) LA measures are reduced to their corresponding existing measures. For example, D♯-measures-LA are reduced to D♯-measures; (2) LD♯-measures, HD♯-measures, and LAD♯-measures are reduced to D♯-measures. We conduct the experiments as follows: (1) Sampling 20 submitted runs every year (2009-2013), which produces 950 pairs of

Table 1: Discriminative power and performance Δ of diversity measures based on the paired bootstrap test at $\alpha = 0.05$. The leftmost column shows existing measures' results; the middle column shows their corresponding LA measures' results using original intent hierarchies (denoted by subscript $_O$); the rightmost column shows their corresponding LA measures' results using extended intent hierarchies (denoted by subscript $_E$). For each row, the greatest value is in bold.

(a) 250 queries in TREC Web Track 2009-2013								
existing measures			measures based on OIH			measures based on EIH		
measure	disc.power	required Δ	measure	disc.power	required Δ	measure	disc.power	required Δ
I-rec	49.1%	0.14	I-rec-LA$_O$	48.5%	0.13	I-rec-LA$_E$	**49.7%**	0.13
α-nDCG	**56.8%**	0.11	α-nDCG-LA$_O$	56.4%	0.12	α-nDCG-LA$_E$	56.0%	0.11
ERR-IA	52.0%	0.12	ERR-IA-LA$_O$	51.1%	0.14	ERR-IA-LA$_E$	**52.1%**	0.14
nDCG-IA	53.4%	0.07	nDCG-IA-LA$_O$	52.4%	0.06	nDCG-IA-LA$_E$	**53.5%**	0.06
Q-IA	43.1%	0.06	Q-IA-LA$_O$	42.7%	0.06	Q-IA-LA$_E$	**43.5%**	0.06
D♯-nDCG	55.1%	0.09	D♯-nDCG-LA$_O$	54.4%	0.10	D♯-nDCG-LA$_E$	**55.3%**	0.09
D♯-Q	52.7%	0.09	D♯-Q-LA$_O$	52.7%	0.09	D♯-Q-LA$_E$	**53.7%**	0.08
(b) 105 queries that have multilayer intent hierarchies (out of 250)								
I-rec	36.4%	0.23	I-rec-LA$_O$	33.4%	0.26	I-rec-LA$_E$	**36.6%**	0.26
α-nDCG	**38.7%**	0.22	α-nDCG-LA$_O$	37.8%	0.26	α-nDCG-LA$_E$	37.9%	0.18
ERR-IA	31.9%	0.19	ERR-IA-LA$_O$	31.2%	0.23	ERR-IA-LA$_E$	**32.9%**	0.21
nDCG-IA	29.9%	0.11	nDCG-IA-LA$_O$	29.2%	0.13	nDCG-IA-LA$_E$	**32.3%**	0.13
Q-IA	20.2%	0.14	Q-IA-LA$_O$	20.3%	0.13	Q-IA-LA$_E$	**22.4%**	0.14
D♯-nDCG	38.9%	0.17	D♯-nDCG-LA$_O$	36.4%	0.19	D♯-nDCG-LA$_E$	**39.7%**	0.18
D♯-Q	38.7%	0.17	D♯-Q-LA$_O$	36.3%	0.16	D♯-Q-LA$_E$	**39.3%**	0.16

Table 2: Discriminative power (shown in columns A) and performance Δ (shown in columns B) of diversity measures ranked by their discriminative power based on the paired bootstrap test at $\alpha = 0.05$. Baseline measures are marked by *.

(a) 250 queries in TREC Web Track 2009-2013					
measure	A	B	measure	A	B
HD♯-nDCG$_E$	55.9%	0.11	HD♯-Q$_E$	53.4%	0.09
LD♯-nDCG$_O$	55.5%	0.10	LAD♯-Q$_O$	53.4%	0.09
LD♯-nDCG$_E$	55.3%	0.09	LD♯-Q$_O$	53.3%	0.09
LAD♯-nDCG$_E$	55.2%	0.10	HD♯-Q$_O$	53.1%	0.09
*D♯-nDCG	55.1%	0.09	LAD♯-Q$_E$	52.9%	0.10
HD♯-nDCG$_O$	54.7%	0.10	LD♯-Q$_E$	52.7%	0.09
LAD♯-nDCG$_O$	54.5%	0.10	*D♯-Q	52.7%	0.09
(b) 105 queries that have multilayer intent hierarchies (out of 250)					
HD♯-nDCG$_E$	40.5%	0.16	LD♯-Q$_E$	39.4%	0.19
LD♯-nDCG$_E$	40.0%	0.17	LAD♯-Q$_E$	39.4%	0.19
LAD♯-nDCG$_E$	39.1%	0.19	HD♯-Q$_E$	38.9%	0.17
*D♯-nDCG	38.9%	0.17	*D♯-Q	38.7%	0.17
LD♯-nDCG$_O$	38.1%	0.19	HD♯-Q$_O$	38.1%	0.17
LAD♯-nDCG$_O$	36.8%	0.21	LD♯-Q$_O$	37.3%	0.19
HD♯-nDCG$_O$	36.5%	0.17	LAD♯-Q$_O$	37.1%	0.15

sampled runs in total; (2) With the 950 pairs of sampled runs, computing the discriminative power and performance Δ using all 250 queries in TREC Web Track 2009-2013 diversity test collections; (3) With the 950 pairs of sampled runs, computing the discriminative power and performance Δ using the 105 queries that have multilayer intent hierarchies. Performance Δ is the required value to achieve statistical significance, and is computed following [21]. The results are shown in Table 1 and Table 2.

By comparing the discriminative power scores of existing measures and their corresponding LA measures based on OIH or EIH in each row of Table 1, we find that: (1) Except α-nDCG-LA, LA measures using EIH are more discriminative than their corresponding existing measures, especially in the case of IA measures. For example, when experimenting with 105 queries that have multilayer intent hierarchies, Q-IA-LA$_E$ (22.4%) outperforms Q-IA (20.2%) in terms of discriminative power; (2) The measures using OIH are less discriminative than the measures using EIH. For example, nDCG-IA-LA$_O$ is 29.2% while nDCG-IA-LA$_E$ is 32.3% when experimenting with 105 queries that have multilayer intent hierarchies.

By comparing the discriminative power results of D♯-measures, LD♯-measures, HD♯-measures, and LAD♯-measures (each block in Table 2), we find that: (1) The measures using EIH are generally more discriminative than the measures using OIH; (2) When using EIH, LD♯-measures, HD♯-measures and LAD♯-measures are better than (or at least as good as) D♯-measures in terms of discriminative power.

By comparing the results using all 250 queries in TREC Web Track 2009-2013 (shown in Table 1(a) or Table 2(a)) and the results only using the queries that have multilayer intent hierarchies (shown in Table 1(b) or Table 2(b)), we find that hierarchical measures have greater improvement of discriminative power than existing measures for queries that have multilayer intent hierarchies. This is reasonable because our measures have potential to recognize the difference between ranked lists by utilizing the hierarchies whereas existing measures cannot. Another justification is that our measures are equivalent to existing measures when the queries only have single-layer intent hierarchies.

The above observations suggest that it is preferable to use EIH when computing hierarchical measures. We think that the hierarchical measures using EIH have higher discriminative power than the hierarchical measures using OIH. For example, Figure 2(b) shows that d_1 is more diversified than d_2 because d_1 refers to two interpretations of the query, while d_2 only refers to one of them. N-rec$_E$ agrees with this but N-rec$_O$ cannot tell which one is more diversified.

4.3 Intuitiveness

4.3.1 *Difference between using OIH and EIH*

The hierarchical measures using EIH are more intuitive than using OIH in terms of diversity. Following the previous work [21], we use I-rec as the gold standard measure for the diversity because it does not depend on intent hierarchies OIH and EIH. Table 3 shows the intuitiveness when using all the queries in TREC Web Track 2009-2013 diversity test collections. We find that for a document cutoff $K = 10$, the hierarchical measures using EIH are more intuitive than using OIH. For a document cutoff $K = 20$, there is only one exception (LD♯-nDCG@20).

This is because the hierarchical measures using OIH may reward high relevance to some official intents, and fail to

Table 3: Intuitiveness based on preference agreement with I-rec. For each measure pair (using OIH or EIH), the higher score is shown in bold and the numbers of disagreements between this pair are shown in parentheses below.

	ERR-IA-LA	nDCG-IA-LA	Q-IA-LA	LD♯-nDCG	HD♯-nDCG	LAD♯-nDCG
(a) Document cutoff $K = 10$. Gold standard measure: I-rec						
OIH	.663	.624	.653	.802	.025	.753
EIH	**.724**	**.748**	**.696**	**.841**	**.999**	**.886**
	(4973)	(5492)	(4660)	(2601)	(3421)	(4948)
(b) Document cutoff $K = 20$. Gold standard measure: I-rec						
OIH	.656	.677	**.821**	.823	.827	
EIH	**.732**	**.748**	**.729**	.739	**.837**	**.840**
	(5362)	(6688)	(5273)	(2511)	(5329)	(5645)

reward wide coverage of the official intents. Take the OIH in Figure 2 as an example. Since we assume that the documents that are relevant to a node are relevant to its parent node, the relevance assessments for intent i_1 or i_5 are reflected in the relevance assessments for their parent node n_1. Though the first layer of the OIH excludes i_1 or i_5, it indirectly considers them through their parent node n_1. By including the other four intents, the first layer considers all six official intents, but the second layer of the OIH only has i_1 or i_5. When combining the two layers, the relevance assessments for i_1 or i_5 are considered twice, once in the first layer and again in the second layer. However, the relevance assessments for the other four intents are only considered once in the first layer. This means that when using the OIH, hierarchical measures mainly reward higher relevance to i_1 and i_5 than other intents. The EIH in Figure 2 solves this problem by extending i_2, i_3, i_4, and i_6 to the second layer so that every official intent can be considered in each layer when evaluating the ranking quality.

In the remaining part of the section, we will only report experimental results using EIH due to space limitation. In most experiments, using EIH yields higher discriminative power and intuitiveness than OIH.

4.3.2 Intuitiveness of Hierarchical Measures

In Section 4.2, we show that LD♯-nDCG$_E$, HD♯-nDCG$_E$, and LAD♯-nDCG$_E$ are highly discriminative among hierarchical measures. In this section, we further compare their intuitiveness with some existing measures, including α-nDCG, ERR-IA, and D♯-nDCG. We do the concordance test based on all the queries in TREC Web Track 2009-2013 diversity test collections, and show the results in Table 4. In Table 4(a) and Table 4(b), we use N-rec$_E$ and Precision as the gold standard measure respectively, whereas in Table 4(c), both N-rec$_E$ and Precision are used as the gold standard measures. We use N-rec$_E$ as a gold standard measure in terms of the diversity because: (1) It is a simple binary measures; (2) It measures diversity better than I-rec, which is traditionally used as the gold standard measure for diversity.

Table 4 shows that (1) In terms of the diversity, LD♯-nDCG$_E$, HD♯-nDCG$_E$, and LAD♯-nDCG$_E$ are more intuitive than existing measures. This is expected because these hierarchical measures directly depend on N-rec$_E$ by means of Equation (11) and the like; (2) In terms of diversity, LD♯-nDCG$_E$ is most intuitive; (3) In terms of relevance, HD♯-nDCG$_E$ is most intuitive; (4) In terms of both diversity and relevance, LAD♯-nDCG$_E$ is the most intuitive measure.

Table 4 shows that using the whole intent hierarchies instead of only using the leaf nodes can improve the intuitiveness of measures. HD♯-nDCG$_E$ and LAD♯-nDCG$_E$ use the whole intent hierarchy to measure both diversity and relevance of ranked lists. LD♯-nDCG$_E$ uses the whole intent hierarchy to measure the diversity but only uses the leaf nodes to measure the relevance. D♯-nDCG only uses the leaf nodes to measure the diversity and relevance. Table 4 shows that LD♯-nDCG$_E$, HD♯-nDCG$_E$ and LAD♯-nDCG$_E$ are more intuitive than D♯-nDCG in terms of diversity. HD♯-nDCG$_E$ and LAD♯-nDCG$_E$ are more intuitive than D♯-nDCG and LD♯-nDCG$_E$ in terms of relevance. We get the same result when both diversity and relevance are considered.

Table 4: Intuitiveness based on preference agreement with gold standard measures. For each measure pair, the higher score is shown in bold and the numbers of disagreements between this pair are shown in parentheses below.

	ERR-IA	D♯-nDCG	LD♯-nDCG$_E$	HD♯-nDCG$_E$	LAD♯-nDCG$_E$
(a) Gold standard measure: N-rec$_E$ ("diversity")					
α-nDCG	**.988**/.362	.661/**.983**	.656/**.986**	.663/**.984**	.661/**.984**
	(14215)	(43908)	(44098)	(44522)	(44444)
ERR-IA	-	.577/**.987**	.573/**.991**	.578/**.990**	.577/**.990**
	-	(56060)	(56011)	(56245)	(56209)
D♯-nDCG	-	-	.428/**.612**	.700/**.741**	.677/**.738**
	-	-	(2124)	(3822)	(3586)
LD♯-nDCG$_E$	-	-	-	**.898**/.799	**.895**/.811
	-	-	-	(2356)	(2026)
HD♯-nDCG$_E$	-	-	-	-	.724/**.915**
	-	-	-	-	(330)
(b) Gold standard measure: Precision ("relevance")					
	ERR-IA	D♯-nDCG	LD♯-nDCG$_E$	HD♯-nDCG$_E$	LAD♯-nDCG$_E$
α-nDCG	**.749**/.345	.359/**.746**	.358/**.749**	.358/**.751**	.357/**.751**
	(14215)	(43908)	(44098)	(44522)	(44444)
ERR-IA	-	.348/**.754**	.346/**.756**	.345/**.758**	.345/**.758**
	-	(56060)	(56011)	(56245)	(56209)
D♯-nDCG	-	-	.488/**.592**	.502/**.625**	.499/**.625**
	-	-	(2124)	(3822)	(3586)
LD♯-nDCG$_E$	-	-	-	.523/**.629**	.518/**.633**
	-	-	-	(2356)	(2026)
HD♯-nDCG$_E$	-	-	-	-	**.603**/.552
	-	-	-	-	(330)
(c) Gold standard measures: N-rec$_E$ and Precision					
	ERR-IA	D♯-nDCG	LD♯-nDCG$_E$	HD♯-nDCG$_E$	LAD♯-nDCG$_E$
α-nDCG	**.738**/.085	.156/**.731**	.154/**.735**	.159/**.734**	.157/**.735**
	(14215)	(43908)	(44098)	(44522)	(44444)
ERR-IA	-	.126/**.742**	.124/**.747**	.127/**.748**	.127/**.749**
	-	(56060)	(56011)	(56245)	(56209)
D♯-nDCG	-	-	.036/**.217**	.267/**.371**	.247/**.368**
	-	-	(2124)	(3822)	(3586)
LD♯-nDCG$_E$	-	-	-	.432/**.438**	.424/**.449**
	-	-	-	(2356)	(2026)
HD♯-nDCG$_E$	-	-	-	-	.370/**.476**
	-	-	-	-	(330)

4.3.3 Case Studies

D♯-nDCG, LD♯-nDCG$_E$, HD♯-nDCG$_E$, and LAD♯-nDCG$_E$ are closely related (shown in Section 3.3.5 and Section 4.4). We 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 5, 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 Δ'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 5. We categorize five cases into two classes from the viewpoint of diversity (**Case A-C**) or relevance (**Case D-E**).

Table 5: Five ranked list pairs from TREC Web Track 2009-2013 diversity test collections, document cutoff $K = 10$. 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 extended intent hierarchies 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♯-nDCG	Δ in LD♯-nDCG$_E$	Δ in HD♯-nDCG$_E$	Δ in LAD♯-nDCG$_E$
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				

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$_E$ can. The rightmost column of Table 5 shows that only D♯-nDCG disagrees with this intuition. In **Case B**, we argue that D♯-nDCG and HD♯-nDCG$_E$ 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$_E$ and LD♯-nDCG$_E$ agree with this. In **Case C**, we argue that LD♯-nDCG$_E$ 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 "bobcat company," while qirdcsuog3 does not cover "bobcat company homepage." Because of this, msrsv2div should be preferred and only LD♯-nDCG$_E$ agrees with this.

In summary, from the viewpoint of diversity, LD♯-nDCG$_E$ is the most intuitive measure. HD♯-nDCG$_E$ is less intuitive than LAD♯-nDCG$_E$, 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$_E$, hence the measures' preference is determined by their Precision part (e.g. D-nDCG if it is D♯-nDCG, and HD-nDCG$_E$ if it is HD♯-nDCG$_E$). In **Case D**, we argue that D♯-nDCG and LD♯-nDCG$_E$ are less intuitive than the other two. No matter whether measuring by I-rec or by N-rec$_E$, 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 5, we find that D♯-nDCG and LD♯-nDCG$_E$ fail to reflect this. In **Case E**, we argue that HD♯-nDCG$_E$ is the most intuitive among the four measures. UWatMDSdm should be preferred because it returns much more relevant documents than 2011SiftR2. In this case, only HD♯-nDCG$_E$ successfully recognizes this.

Table 6: Kendall's τ / Symmetric τ_{ap} by averaging over TREC Web track 2009-2013. Values greater than .950 are shown in bold.

(a) 250 queries in TREC Web Track 2009-2013

	ERR-IA	D♯-nDCG	LD♯-nDCG$_E$	HD♯-nDCG$_E$	LAD♯-nDCG$_E$
α-nDCG	.923/.870	.840/.796	.845/.796	.843/.792	.844/.793
ERR-IA	-	.772/.699	.780/.706	.779/.704	.779/.706
D♯-nDCG	-	-	**.976/.959**	**.976/.957**	**.977/.960**
LD♯-nDCG$_E$	-	-	-	**.991/.988**	**.995/.993**
HD♯-nDCG$_E$	-	-	-	-	**.995/.994**

(b) 105 queries that have multilayer intent hierarchies (out of 250)

	ERR-IA	D♯-nDCG	LD♯-nDCG$_E$	HD♯-nDCG$_E$	LAD♯-nDCG$_E$
α-nDCG	.872/.802	.812/.747	.821/.755	.821/.758	.822/.759
ERR-IA	-	.701/.609	.714/.624	.712/.626	.714/.628
D♯-nDCG	-	-	**.964**/.941	**.959**/.933	**.958**/.932
LD♯-nDCG$_E$	-	-	-	**.984/.977**	**.986/.981**
HD♯-nDCG$_E$	-	-	-	-	**.996/.995**

Generally, from the viewpoint of relevance, LAD♯-nDCG$_E$ is more intuitive than LD♯-nDCG$_E$. LAD♯-nDCG$_E$ 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$_E$.

4.4 Rank Correlation Results

We compute Kendall's τ and τ_{ap} for different pairs of measures to check the correlation between these measures. Results are shown in Table 6. The table shows that: (1) LD♯-nDCG$_E$, HD♯-nDCG$_E$ and LAD♯-nDCG$_E$ are less correlated to existing measures, especially when only using the queries that have multilayer intent hierarchies. This is because our measures are able to recognize the subtle difference between ranked lists when multilayer intent hierarchies are used, whereas the existing measures may not. This indicates that our measures are useful and could be supplementary to the existing measures; (2) LD♯-nDCG$_E$, HD♯-nDCG$_E$, as well as LAD♯-nDCG$_E$ 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$_E$ and HD♯-nDCG$_E$ are less correlated. As discussed in 4.3, LD♯-nDCG$_E$ prefers highly diversified ranked lists, whereas HD♯-nDCG$_E$ prefers highly relevant ranked lists.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we argued that user intents of a query could be hierarchical. We described the concept of hierarchical intents and proposed hierarchical measures that could work with intent hierarchies. We created a new test collection containing intent hierarchies based on the existing TREC Web Track 2009-2013 diversity test collections by grouping the official intents into original intent hierarchies and extending them to extended intent hierarchies. Our experimental results showed that our proposed hierarchical measures can be more discriminative than existing measures which use a flat list of intents and assume the independence among intents. We revealed that 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 both diversity and relevance are important.

In this paper, we simply assume that the official intents provided in TREC Web Track 2009-2013 diversity test collections are atomic intents. It is possible that some of these intents can be further divided into sub intents. We will investigate this in the future.

6. ACKNOWLEDGMENTS

This work was supported by the National Key Basic Research Program (973 Program) of China under grant No. 2014CB340403, and the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China No. 15XNLF03, the National Natural Science Foundation of China (Grant No. 61502501, 61502502, and 61502503)

7. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM*, 2009.

[2] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.

[3] B. A. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *TOIS*, 2012.

[4] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM*, 2009.

[5] H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR*, 2006.

[6] C. L. Clarke, N. Craswell, and I. Soboroff. Overview of the trec 2009 web track. In *TREC*, 2009.

[7] C. L. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *WSDM*, 2011.

[8] C. L. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Overview of the trec 2010 web track. In *TREC*, 2010.

[9] 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.

[10] C. L. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *ICTIR*, 2009.

[11] V. Dang and B. W. Croft. Term level search result diversification. In *SIGIR*, 2013.

[12] V. Dang and W. B. Croft. Diversity by proportionality: an election-based approach to search result diversification. In *SIGIR*, 2012.

[13] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *WSDM*, 2011.

[14] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *WWW*, 2007.

[15] 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.

[16] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR*, 2000.

[17] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 1938.

[18] F. Radlinski and S. Dumais. Improving personalized web search using result diversification. In *SIGIR*, 2006.

[19] T. Sakai. Bootstrap-based comparisons of ir metrics for finding one relevant document. In *AIRS*, 2006.

[20] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *SIGIR*, 2006.

[21] T. Sakai. Evaluation with informational and navigational intents. In *WWW*, 2012.

[22] T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C.-Y. Lin. Simple evaluation metrics for diversified search results. In *EVIA*, 2010.

[23] T. Sakai and S. Robertson. Modelling a user population for designing information retrieval metrics. In *EVIA*, 2008.

[24] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. In *SIGIR*, 2011.

[25] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *WWW*, 2010.

[26] R. L. Santos, C. Macdonald, and I. Ounis. Selectively diversifying web search results. In *CIKM*, 2010.

[27] R. L. Santos, C. Macdonald, and I. Ounis. Intent-aware search result diversification. In *SIGIR*, 2011.

[28] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 1999.

[29] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *SIGIR*, 2008.

[30] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR*, 2003.

[31] X. Zhu, A. B. Goldberg, J. Van Gael, and D. Andrzejewski. Improving diversity in ranking using absorbing random walks. In *HLT-NAACL*, 2007.