**Search Result Diversity Evaluation based on Intent Hierarchies**

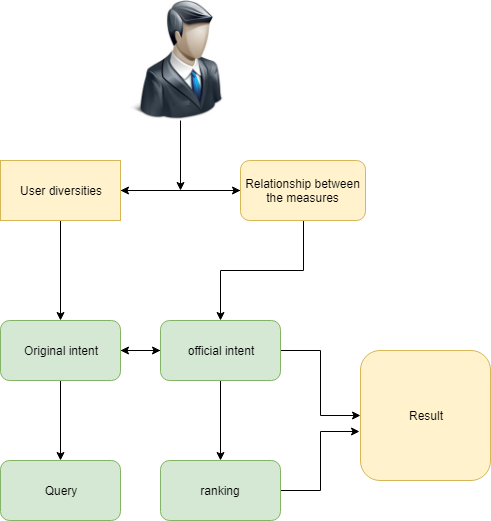
**Abstract:**

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.

**Architecture:**

****

**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. 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 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, 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 i1-i4. The figure shows that except intent i2 which is about “wild bobcat,” the remaining ones, i1, i3, and i4, are all about “bobcat company.” This indicates that i1, i3, and i4 are more related to each other, but are less related to i2. Even within the three intents about “bobcat company,” i1 and i3 are closer because they are about the businesses involving tractors of the company, whereas i4 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 i1, i3, and i4, while THUIR10DvNov covers i1, i2, and i4 in their top ten documents. Since i1, i3, and i4 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. We think that this is due to their lack of recognition of the relationships among intents.

**Existing System:**

Specifically, we find two submitted runs for the query, cmuFuTop10D and THUIR10DvNov, in TREC Web Track 2010 diversity task. cmuFuTop10D covers i1, i3, and i4, while THUIR10DvNov covers i1, i2, and i4 in their top ten documents. Since i1, i3, and i4 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 treats them as equally good because they cover the same number of intents. Some other existing measures also have similar problems. 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 to solve the problems mentioned above. This paper is the extended version of the SIGIR 2016 paper . In the original work, 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.

**Proposed System:**

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. 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. 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. 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. Describes some existing diversity measures and the methods for testing them. 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. 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

**Modules:**

* **Discriminative Power**
* **Concordance Test**
* **Hierarchical Measures and Hierarchical Diversification Algorithms**
* **Creating Intent Hierarchies**

**Discriminative Power:**

On TREC test collections, we apply the same method as described 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 , we find that the hierarchical measures using EIH are consistently more discriminative than those using OIH. This finding holds true no matter which weight-ing scheme is used.

**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. By comparing the first column with the second column, and comparing the third column, 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.

**Hierarchical Measures and Hierarchical Diversification Algorithms:**

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.Many financial products today are still based on linear models and the Markowitz framework. It is lauded as Nobel prize winning and scientific. However, the math of choice may be inadequate as has recently been pointed out by several articles, e.g. in the Journal of Portfolio Management. They argue that economics has become a theory that describes an idealized economic and financial reality without any strong connection to data.

**Creating 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:

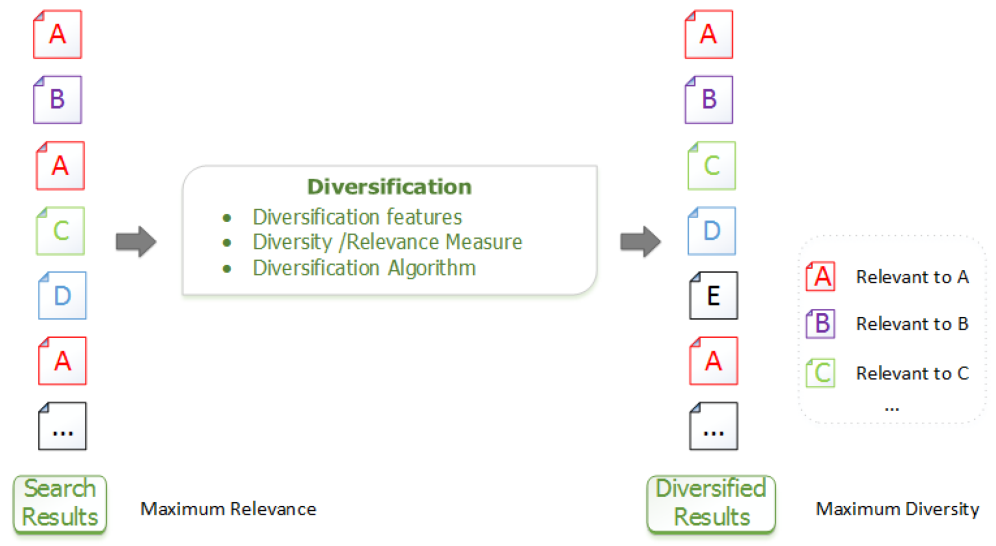
Property 1. The intent hierarchy is in a tree structure, where every child has only one parent.

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.

**Algorithm:**

**Hierarchical Diversification Algorithms:**

Many financial products today are still based on linear models and the Markowitz framework. It is lauded as Nobel prize winning and scientific. However, the math of choice may be inadequate as has recently been pointed out by several articles, e.g. in the Journal of Portfolio Management. They argue that economics has become a theory that describes an idealized economic and financial reality without any strong connection to data. For example, Mean-Variance (MV) optimisation was created in the 1950s where computation time was expensive and data was rare. It does not address the high level of uncertainty in financial markets leading to instability, concentration and underperformance. The consequences of using overly simplistic and outdated statistical methods are palpable.



**Future 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 {i}. Some measures can handle intent probability P r(i|q) and graded relevance assessments but some cannot. In this section, we summarize the previous work on designing and evaluating diversity measures.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV or Later Version

➢ RAM - 4 GB (min)

➢ Hard Disk - 40 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP or Later Version
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL

**Conclusion:**

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.