# Search Result Diversification

Marina Drosou
Dept. of Computer Science
University of Ioannina, Greece
mdrosou@cs.uoi.gr

Evaggelia Pitoura
Dept. of Computer Science
University of Ioannina, Greece
pitoura@cs.uoi.gr

## ABSTRACT

Result diversification has recently attracted much attention as a means of increasing user satisfaction in recommender systems and web search. Many different approaches have been proposed in the related literature for the diversification problem. In this paper, we survey, classify and comparatively study the various definitions, algorithms and metrics for result diversification.

## 1. INTRODUCTION

Today, most user searches are of an exploratory nature, in the sense that users are interested in retrieving pieces of information that cover many aspects of their information needs. Therefore, recently, *result diversification* has attracted considerable attention as a means of counteracting the over-specialization problem, i.e. the retrieval of too homogeneous results in recommender systems and web search, thus enhancing user satisfaction (e.g. [20, 16]). Consider, for example, a user who wants to buy a car and submits a related web search query. A diverse result, i.e. a result containing various brands and models with different horsepower and other technical characteristics is intuitively more informative than a result that contains a homogeneous result containing only cars with similar features.

Diversification is also useful in counter-weighting the effects of personalization. Personalization aims at tailoring results to meet the preferences of each specific individual (e.g. [10, 15]). However, this may lead to overly limiting the search results. Diversification can complement preferences and provide personalization systems with the means to retrieve more satisfying results (as in [14]).

In this paper, we survey the various approaches taken in the area of result diversification. We classify the ways that diverse items in the related literature are generally defined in three different categories, namely in terms of (i) *content* (or *similarity*), i.e. items that are dissimilar to each other (e.g. [18]), (ii) *novelty*, i.e. items that contain new information when compared to previously seen ones (e.g. [3, 19]) and (iii) *coverage*, i.e. items that belong to different categories (e.g. [1]). Then, we present various algorithms for result diversification and classify them into two main groups, namely (i) *greedy* (e.g. [20]) and (ii) *interchange* (e.g. [17]) algorithms. We also show the main metrics used for evaluating the performance of diversification systems.

The rest of this paper is structured as follows. In Section 2, we classify various definitions of the result diversification problem, while in Section 3, we see how diversity is combined with other ranking criteria. In Section 4, we review the proposed algorithms for efficiently retrieving diverse results and, in Section 5, we show measures used for evaluating the diversity of selected items. Finally, Section 6 concludes this paper.

## 2. DIVERSITY DEFINITIONS

Generally, the problem of selecting diverse items can be expressed as follows. Given a set[1] $\mathcal{X}$ of $n$ available items and a restriction $k$ on the number of wanted results, the goal is to select a subset $S^*$ of $k$ items out of the $n$ available ones, such that, the diversity among the items of $S^*$ is maximized.

In this section, we present various specific definitions of the result diversification problem that can be found in the research literature. We classify these definitions based on the way that diverse items are defined, i.e. (i) content, (ii) novelty and (iii) coverage. Note that, this classification is sometimes fuzzy, since these factors are related to each other and, therefore, a definition can affect more than one of them.

### 2.1 Content-based definitions

Content-based definitions interpret diversity as an instance of the *p-dispersion problem*. The ob-

[1]In some works, the term "set" is used loosely to denote a set with *bag semantics* or a *multiset*, where the same item may appear more than once in the set.

jective of the $p$-dispersion problem is to choose $p$ out of $n$ given points, so that the minimum distance between any pair of chosen points is maximized [6]. The $p$-dispersion problem has been studied in the field of Operations Research for locating facilities that should be dispersed; such as franchises belonging to a chain or nuclear power plants. Formally, the $p$-dispersion problem is defined as follows:

Given a set $\mathcal{X}$ of points, $\mathcal{X} = \{x_1, \ldots, x_n\}$, a distance metric $d(.,.)$ among points and an integer $k$, locate a subset $S^*$ of $\mathcal{X}$, such that:

$$S^* = \underset{\substack{S \subseteq \mathcal{X} \\ |S|=k}}{\operatorname{argmax}} f(S), \text{ where } f(S) = \min_{\substack{x_i, x_j \in \mathcal{X} \\ x_i \neq x_j}} d(x_i, x_j)$$

Content-based definitions of diversity have been proposed in the context of web search and recommender systems. Most often, however, the objective function that is maximized is the *average distance* of any two points, instead of the minimum one, that is:

$$f(S) = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j>i}^{k} d(x_i, x_j) \qquad (1)$$

This approach is followed in [20], where the diversity of a set of recommendations in a typical recommender system is defined based on their *intra-list similarity*, which is the application of Equation 1 along with a user-defined distance metric.

Another work that defines diverse recommendations based on content is [17]. The distance between recommendations is measured based on their *explanations*. Given a set of items $\mathcal{X}$ and a set of users $\mathcal{U}$, the explanation of an item $x \in \mathcal{X}$ recommended to a user $u \in \mathcal{U}$ can be defined in a content-based approach as:

$$Expl(u, x) = \{x' \in \mathcal{X} | sim(x, x') > 0 \wedge x' \in Items(u)\}$$

where $sim(x, x')$ is the similarity of $x$, $x'$ and $Items(u)$ is the set of all items rated in the past by user $u$. A non content-based collaborative filtering approach is also considered, in which:

$$Expl(u, x) = \{u' \in \mathcal{U} | sim'(u, u') > 0 \wedge x \in Items(u')\}$$

where $sim'(.,.)$ is a similarity metric between two users. The similarity $sim(.,.)$ between two items $x$ and $x'$ can be defined based on the Jaccard similarity coefficient, the cosine similarity or any other similarity measure. The diversity of a set of items $S \subseteq \mathcal{X}$ is defined as the average distance of all pairs of items (as in Equation 1). A similar Jaccard-based similarity measure is also used in [7]. In that case, each document is described by a sketch produced by a number of hash functions. Another alternative distance metric used in that work is a taxonomy-based categorical distance when this can be applied (e.g. in the case of documents).

A content-based definition of diversity has also been applied in the context of publish/subscribe systems [5, 4]. Here, given a period or a window of matching events and an integer $k$, only the $k$ most diverse of them (based on Equation 1) are delivered to the related subscribers.

Another definition that can be classified in this category is the one used in [16] in the context of database systems. Given a database relation $\mathcal{R} = (A_1, \ldots, A_m)$, a *diversity ordering* of $\mathcal{R}$, denoted $\prec_{\mathcal{R}}$, is a total ordering of its attributes based on their importance, say $A_1 \prec \ldots \prec A_m$. Also, a *prefix with respect to* $\prec_{\mathcal{R}}$, denoted $\rho$, is defined as a sequence of attribute values in the order given by $\prec_{\mathcal{R}}$, moving from higher to lower priority. Let $\rho$ be a prefix of length $l$ and $t$, $t'$ be two tuples of $\mathcal{R}$ that share $\rho$. The similarity between $t$ and $t'$ is defined as:

$$sim_{\rho}(t, t') = \begin{cases} 1 & \text{if } t.A_{l+1} = t'.A_{l+1} \\ 0 & \text{otherwise} \end{cases}$$

Now, given an integer $k$, a subset $S$ of $\mathcal{R}$ with cardinality $k$ is defined to be diverse with respect to $\rho$ if all tuples in $S$ share the prefix $\rho$ and the sum of their pair-wise similarities, as defined above, is minimized. $S$ is also said to be diverse with respect to $\mathcal{R}$ if it is diverse with respect to every possible prefix for $\mathcal{R}$.

Finally, a content-based definition of diversity is used in [8] to extend the *k-nearest neighbor* problem, so that, given an item $x$, the $k$ spatially closest results that are sufficiently different from the rest of the answers are retrieved. In this case, the distance between two items is based on the Gower coefficient, i.e. a weighted average of the respective attribute differences of the items. Assuming $\delta_i$ to be equal to the difference between the $i^{\text{th}}$ attributes of two items $x$, $x'$ then:

$$d(x, x') = \sum_i w_i \delta_i$$

where $w_i$ is a weight corresponding to the $i^{\text{th}}$ dimension of the items. Two items are considered diverse if their distance $d(.,.)$ is greater than a given threshold and a set $S$ is considered diverse if all the pairs of items in it are diverse.

## 2.2 Novelty-based definitions

Novelty is a notion closely related to that of diversity, in the sense that items which are diverse from all items seen in the past are likely to contain novel information, i.e. information not seen before.

A distinction between novelty and diversity in the context of information retrieval systems is made in

[3], where novelty is viewed as the need to avoid redundancy, whereas diversity is viewed as the need to resolve ambiguity. Each document $x$ and query $q$ are considered as a collection of *information nuggets* from the space of all possible nuggets $\mathcal{O} = \{o_1, \dots, o_m\}$. Given a binary random variable $R_x$ that denotes whether a document $x$ is considered relevant to a given query $q$, then:

$$P(R_x = 1|q, x) = P(\exists o_i, \text{ such that } o_i \in x \cap q)$$

Now, given an ordered list of documents $x_1, \dots, x_n$ retrieved by an IR system for $q$, the probability that the $k^{\text{th}}$ document is both novel and diverse from the $k-1$ first ones, i.e. $R_{x_k} = 1$, is equal to the probability of that document containing a nugget that cannot be found in the previous $k-1$ documents. Given a list of $k-1$ preceding documents, the probability that a nugget $o_i \in \mathcal{O}$ is novel for a query $q$ is:

$$P(o_i \in q|q, x_1, \dots, x_{k-1}) = P(o_i \in q) \prod_{j=1}^{k-1} P(o_i \notin x_j)$$

Assuming that all nuggets are independent and equally likely to be relevant for all queries, then:

$$P(R_{x_k} = 1|q, x_1, \dots, x_k) = 1 - \prod_{i=1}^{m}(1 - \gamma\alpha J(x_k, o_i)(1-\alpha)^{r_{o_i,k-1}}) \quad (2)$$

where $J(x_k, o_i) = 1$ if some human judge has determined that $x_k$ contains the nugget $o_i$ (or zero otherwise), $\alpha$ is a constant in $(0, 1]$ reflecting the possibility of a judge error in positive assessment, $\gamma = P(o_i \in q)$ and $r_{o_i,k-1}$ is the number of documents ranked up to position $k-1$ that have been judged to contain $o_i$, i.e. $r_{o_i,k-1} = \sum_{j=1}^{k-1} J(x_j, o_i)$. This approach requires prior knowledge of the nuggets and also considerable amount of human effort for judging the relevance of documents in order to compute the related probabilities.

Another work based on novelty is [19], which aims at enhancing adaptive filtering systems with the capability of distinguishing novel and redundant items. Such systems should identify documents that are similar to previously delivered ones, in the sense of having the same topic, but also dissimilar to them, in the sense of containing novel information. The redundancy $R$ of each document $x$ is measured with respect to its *similarity* to all previously delivered documents, denoted $D(x)$, as follows:

$$R(x|D(x)) = \operatorname*{argmax}_{x' \in D(x)} R(x|x')$$

where $R(x|x')$ is the redundancy (similarity) of $x$ with respect to another document $x'$. Three different ways for measuring $R(x|x')$ are considered, namely the *set difference*, the *geometric distance* and the *distributional distance*. The set difference is based on the number of new terms that appear in $x$:

$$R(x|x') = \left|Set(x) \cap \overline{Set(x')}\right|$$

In the above formula, given a term $w$ and a document $x$, it holds that $w \in Set(x)$, if and only if, $Count(w, x) > h$, where $h$ is a constant and $Count(w, x) = \alpha_1 tf_{w,x} + \alpha_2 df_w + \alpha_3 rdf_w$. $tf_{w,x}$ is the frequency of $w$ in $x$, $df_w$ is the number of all filtered documents that contain $w$, $rdf_w$ is the number of delivered documents that contain $w$ and $\alpha_1, \alpha_2, \alpha_3$ are constants with $\alpha_1 + \alpha_2 + \alpha_3 = 1$. The geometric distance is based on the cosine similarity between $x$ and $x'$: If we represent each document $x$ as a vector $\mathbf{x} = (tf_{w_1,x}, tf_{w_2,x}, \dots, tf_{w_m,x})^T$, where $w_1, w_2, \dots, w_m$ are all the available terms, then:

$$\begin{aligned} R(x|x') &= \cos(\mathbf{x}, \mathbf{x'}) \\ &= \frac{\mathbf{x}^T\mathbf{x'}}{\|\mathbf{x}\| \, \|\mathbf{x'}\|} \end{aligned}$$

Finally, the distributional distance is based on a probabilistic language model. Each document $x$ is represented by a unigram word distribution $\theta_x$ and the distance among two documents is measured via the Kull- back-Leibler formula:

$$\begin{aligned} R(x|x') &= -KL(\theta_x, \theta_{x'}) \\ &= -\sum_{w_j} P(w_j|\theta_x) \log \frac{P(w_j|\theta_x)}{P(w_j|\theta_{x'})} \end{aligned}$$

A mixture-model approach is considered in order to find the language models for the $\theta$ distributions.

## 2.3 Coverage-based definitions

Some works view diversity in a different way, that of selecting items that cover many different interpretations of the user's information need. For example, [1] considers typical web search and, given a query $q$ and a taxonomy $C$ of independent information categories, aims at retrieving $k$ documents that cover many interpretations of $q$, especially interpretations that are considered important. The result diversification problem in this context is formally defined as follows: Given a query $q$, a set of documents $\mathcal{X}$, a taxonomy $C$, a probability distribution $P(c|q)$ of each category $c \in C$ being relevant to $q$, the probability $V(x|q, c)$ of each document $x \in \mathcal{X}$ being relevant to each category $c$ for $q$ and an integer $k$, find a set of documents $S^*$, such that:

$$S^* = \operatorname*{argmax}_{\substack{S \subseteq \mathcal{X} \\ |S|=k}} P(S|q)$$

where:

$$P(S|q) = \sum_c P(c|q)(1 - \prod_{x \in S}(1 - V(x|q, c))) \quad (3)$$

The probability of $x$ *not* covering a relevant to the query $q$ category $c$ is equal to $(1-V(x|q,c))$. Therefore, the above equation, in essence, maximizes the probability of each relevant category $c$ being covered by at least one document in $S$. This method requires prior knowledge of the taxonomy and the learning of the probability distributions.

[12] also makes use of a cover-based definition of diversity to locate and highlight diverse concepts in documents. Given a set of *sentences* $S$, the *cover* of $S$ is the union of all terms $t$ appearing in any sentence $x$ in them, that is:

$$Cov(S) = \bigcup_{x \in S} \bigcup_{t \in x} t$$

Assuming a function $g(i)$ that measures the benefit we have by covering a term exactly $i$ times, the *gain* of $S$ is:

$$Gain(S) = \sum_{i=0}^{|S|} \sum_{t \in \mathcal{T}_i} w(t) g(i)$$

where $\mathcal{T}_i$ is the set of terms appearing in exactly $i$ sentences in $S$ and $w(t)$ is a weight for the term $t$. Now, the result diversification problem is defined as follows: Given a document consisting of $n$ sentences $\mathcal{X} = \{x_1, \ldots, x_n\}$ and an integer $k$, locate a set of sentences $S^*$, such that:

$$S^* = \underset{\substack{S \subseteq \mathcal{X} \\ |S| \leq k}}{\operatorname{argmax}}\, Gain(S) \tag{4}$$

## 3. COMBINATION OF DIVERSITY WITH OTHER CRITERIA

Diversity is most commonly used along with some other ranking criterion, most commonly that of *relevance* to the user's query. To the best of our knowledge, the first work in which the two measures were combined is [2], in which *marginal relevance*, i.e. a linear combination of relevance and diversity, is proposed as a criterion for ranking results retrieved by IR systems. A document has high marginal relevance if it is both relevant to the user query $q$ and also exhibits minimal similarity to previously selected documents. Formally, given the set of all retrieved documents $\mathcal{X}$ and the set of already selected ones, denoted $S$, the document $x^* \in \mathcal{X} \backslash S$ that has the maximum marginal relevance to $S$ is:

$$x^* = \underset{x \in \mathcal{X} \backslash S}{\operatorname{argmax}} \left[ \lambda(rel(x) - (1-\lambda) \max_{x' \in S} d(x, x')) \right]$$

where $rel(x)$ is the relevance of $x$ to the query and $\lambda \in [0,1]$. This approach has also been applied in [14] as a means to reformulate queries submitted in web search. The above formulation of the problem is called *max-sum diversification*. The objective function that is maximized this case is:

$$f(S) = (k-1) \sum_{x \in S} rel(x) + 2\lambda \sum_{x, x' \in S} d(x, x')$$

where $\lambda > 0$. Other variations of combining relevance and diversity are the *max-min diversification*, where:

$$f(S) = \min_{x \in S} rel(x) + \lambda \min_{x, x' \in S} d(x, x')$$

and also a *mono-objective* formulation of the problem in which:

$$f(S) = \sum_{x \in S} \left[ rel(x) + \frac{\lambda}{|\mathcal{X} - 1|} \sum_{x' \in \mathcal{X}} d(x, x') \right]$$

[7] considers the combination of relevance and diversity and presents eight intuitive axioms that diversification systems should satisfy. However, it is shown that not all of them can be satisfied simultaneously.

The combination of these two criteria has also been studied in [18] as an optimization problem. Let once again $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a set of items and $D$ be an $n \times n$ distance matrix with the $(i,j)^{\text{th}}$ element being equal to $d(x_i, x_j)$. Let also $\mathbf{m}$ be an $n$-dimensional vector with the $i^{\text{th}}$ element being equal to $rel(x_i)$. Consider, finally, an integer $k$ and a binary $n$-dimensional vector $\mathbf{y}$ with the $i^{\text{th}}$ element being equal to 1, if and only if, $x_i$ belongs to the $k$ most highly relevant and diverse items. Now, given a diversification factor $\lambda \in [0,1]$, we can define the problem of selecting $k$ items that are both as relevant and diverse as possible as follows:

$$\begin{aligned} \mathbf{y}^* \quad &= \quad \underset{\mathbf{y}}{\operatorname{argmax}} (1-\lambda)\alpha \mathbf{y}^T D \mathbf{y} + \lambda \beta \mathbf{m}^T \mathbf{y} \\ \text{s.t.} \quad & \quad \mathbf{1}^T \mathbf{y} = k \text{ and} \\ & \quad y(i) \in \{0,1\}, 1 \leq i \leq n \end{aligned} \tag{5}$$

where $\alpha$ and $\beta$ are normalization parameters.

Diversity is also combined with spatial distance, as a relevance characterazation, when solving the $k$-nearest diverse neighbors problem in [8].

Finally, threshold-based techniques can also be employed as in [18], where variations of the optimization problem of Equation 5 are considered (e.g. maximize the diversity of the selected items given a relevance threshold and, the dual, maximize the relevance of the selected items given a minimum required diversity). Placing a threshold on diversity however may be hard, since it requires an estimation of the achievable diversity.

## 4. ALGORITHMS

Given a set $\mathcal{X}$ of items, $\mathcal{X} = \{x_1, \ldots, x_n\}$, a distance metric $d(.,.)$ among items and an integer $k$, the diversification problem is to locate a subset $S^*$ of $\mathcal{X}$, such that the diversity among the selected items is maximized, where the diversity of a set of items is defined based on some specific definition of Section 2.

Generally, the diversification problem has been shown to be NP-hard. Thus, to solve large instances of the problem, we need to rely on heuristics. Many heuristic algorithms have been used in the research literature and have been employed for solving variations of the problem in more than one research fields. We can classify these heuristics into two main categories: (i) *greedy* and (ii) *interchange* (or *swap*). In following, we describe heuristics in each category and their applications.

## 4.1 Greedy Heuristics

The greedy heuristics are the ones most commonly used since they are intuitive and some of them are also relatively fast. Greedy heuristics generally make use of two sets: the set $\mathcal{X}$ of available items and the set $S$ which contains the selected ones. Items are iteratively moved from $\mathcal{X}$ to $S$ and vice versa until $|S| = k$ and $|\mathcal{X}| = n - k$. In most works, $S$ is initialized with some item, e.g. the most relevant one, and then items are moved one-by-one from $\mathcal{X}$ to $S$ until $k$ of them have been selected. The item that is moved each time is the one that has the maximum *item-set distance* from $S$. The item-set distance, denoted $setdist(x_i, S)$, between an item $x_i$ and a set of items $S$ is defined based on its distance from the items in $S$, for example:

$$setdist(x_i, S) = \min_{x_j \in S} d(x_i, x_j)$$

or

$$setdist(x_i, S) = \frac{1}{|S|} \sum_{x_j \in S} d(x_i, x_j)$$

Ties are generally broken arbitarily.

This greedy approach is, for example, used in [20] in the context of recommender systems, where, given a set of recommendations $\mathcal{X} = \{x_1, \ldots, x_n\}$ and their degrees of relevance $rel(x_i)$, $1 \leq i \leq n$, to a user query, diverse recommendations are produced. $S$ is initialized with the most relevant recommendation. Then, recommendations are added one-by-one to $S$ as follows: For each recommendation $x_i$ not yet added to $S$, its item-set distance from the recommendations already in $S$ is computed. These "candidate" recommendations are then sorted in order of (i) relevance to the query and (ii) item-set distance to $S$. The rank of each recommendation is a linear combination of its positions in the two sorted lists. The recommendation with the minimum rank is added to $S$ and the process is repeated until $S$ has $k$ recommendations. Note that the recommender system has to produce a larger number of recommendations ($n$) out of which the final $k$ ones will be selected. The larger this number, the higher the possibility that more diverse recommendations will be located (at the cost of higher computation cost).

A greedy heuristic is also employed in [12] for locating diverse sentences in documents. At each round, the sentence which has the highest gain for $S$, as defined in Equation 4, is added to $S$. [1] also follows the greedy approach. In that case, an algorithm is proposed that, given the set of the top-$k$ most relevant documents to a query, it re-orders them in a way, such that, the objective function of Equation 3 is maximized. [7] also employs another greedy variation, first presented in [9] as a solution to the $p$-dispersion problem, in which, at each iteration, the two remaining items with the largest pair-wise distance are added to $S$. A greedy solution is also used in [17] for recommenders. However, in that case, threshold values are also used to determine when two recommendations are considered distant. [8] also uses a greedy algorithm for locating the $k$-nearest diverse neighbors to a given item.

A special case of greedy heuristics are neighborhood heuristics. These algorithms start with a solution $S$ containing one random item and then iteratively add items to the solution. The items to be considered at each iteration are limited based on the notion of *r-neighborhood* of an item $x_i \in \mathcal{X}$, $N(x_i, \mathcal{X}, r)$, defined as:

$$N(x_i, \mathcal{X}, r) = \{x_j \in \mathcal{X} : d(x_i, x_j) \leq r\}$$

In other words, all items that have a smaller or equal to $r$ distance to $x_i$ belong to its $r$-neighborhood. At each iteration, only items outside the $r$-neighborhoods of all already selected items are considered. Out of these items, one is chosen to be added to the solution. This can be the fist located item outside those $r$-neighborhoods, the one that has the smallest sum of distances to the already selected items or the one that has the largest sum of distances to the already selected items [6]. Note that the selection of $r$ plays an important role as it restricts the number of items that are considered at each iteration. In fact, given a value of $r$, a solution $S$ with $|S| = k$ may not even exist.

## 4.2 Insterchange (Swap) Heuristics

Interchange (or Swap) heuristics have also been used in the literature for solving the diversification problem. Generally, these heuristics are initialized with a random solution $S$ and then iteratively attempt to improve that solution by interchanging an item in the solution with another item that is not in the solution. At each round, possible interchanges are the first met one that improves the solution or the one that improves the solution the most.

An interchange heuristic that combines the relevance and diversity criteria is proposed in [17]. In this approach, $S$ is initialized with the $k$ most relevant items. At each iteration, the item of $S$ that contributes the least to the diversity of the entire set, i.e. the one with the minimum item-set distance, is interchanged with the most relevant item in $\mathcal{X} \backslash S$. Interchanges stop when there are no more items in $\mathcal{X} \backslash S$ with higher relevance than a given threshold.

Another work that employs an interchange algorithm is [13], where, given a set of structured search results, the goal is to identify a subset of their features that are able to differentiate them. Starting with a random subset of features, at each iteration, one of these features is interchanged with a better candidate feature.

### 4.3 Other Heuristics

An algorithm for achieving diversity in database systems based on a tree index structure, i.e. the Dewey tree, is presented in [16]. Each tuple of a database relation is represented by a path in the tree. Higher levels of the tree represent more important attributes, according to the diversity ordering of the relation (see Section 2). Diverse tuples are retrieved by traversing this tree.

Motivated by the fact that the one-dimensional $p$-dispersion problem can be solved optimally in polynomial time, [6] considers a dimensionality-reduction heuristic that projects items in one dimension only. However, in practice, this approach does not result in good solutions.

A hybrid greedy/interchange heuristic is used in [4] in the context of continuous data. In this case, a diverse subset $S$ is located using a greedy approach and then its diversity is further improved by performing interchanges.

Another related approach is that of [11], where, given the set of a database query results, these results are grouped in $k$ clusters and the corresponding $k$ medoids are retrieved as a subset of $k$ representative and diverse results.

Finally, in [18], where the diversification problem is formulated as an optimization one, a solution is approximated via optimization techniques that include problem relaxation and quantization.

## 5. EVALUATION MEASURES

The diversity of a set $S$ of selected items can be evaluated by the value of the objective function $f(S)$ based on which the diversity problem is defined, e.g. Equation 1. This approach is used in most of the related work (e.g. [20, 17, 5, 12, 18]). The computed value can be normalized by the corresponding value for the set $S^*$, i.e. the optimal solution to the diversification problem. This, however, is not always feasible due to the high cost of computing the optimal solution.

In the field of IR systems, there has been an effort to adapt traditional IR evaluation measures so as to become diversity-aware. A key difference of these approaches is that the retrieved results are usually viewed as an *ordered list* instead of a set. These adapted measures are usually applied along with novelty-based or coverage-based diversity definitions.

For example, [3] proposes evaluating retrieved results through a weighted *Normalized Discounted Cumulative Gain Measure* (denoted $\alpha$-NDCG), a measure often used in the context of IR systems that measures the gain of an item being at a specific position of the list given the items that precede it. Given an ordered list of items, the $k^{\text{th}}$ element of the list's gain vector, denoted $\mathbf{G}$, is computed based on Equation 2 as:

$$\mathbf{G}[k] = \sum_{i=1}^{m} J(x_k, o_i)(1-\alpha)^{r_{o_i, k-1}}$$

and the corresponding cumulative gain vector, denoted $\mathbf{GC}$, is computed as:

$$\mathbf{CG}[k] = \sum_{j=1}^{k} \mathbf{G}[j]$$

Usually, the elements of the cumulative gain vector are weighted according to their position in the list, so the discounted cumulative gain vector, denoted $\mathbf{DGC}$, is computed as:

$$\mathbf{DCG}[k] = \sum_{j=1}^{k} \frac{\mathbf{G}[j]}{\log_2(1+j)}$$

The discounted cumulative gain vector computed for a list is finally normalized by the ideal discounted cumulative gain. However, the computation of this is an NP-complete problem and, in practice, its value is approximated via heuristics.

The adaptation of the NDCG measure is also considered in [1], where NDCG is aggregated over all available categories that a document may be related to (see Section 2). This variation is called *Intent-Aware Normalized Discounted Cumulative Gain Measure* (denoted NDCG-IA). Its value for the $k^{\text{th}}$ element of a list $S$ of items retrieved for a query $q$ is:

$$NDCG\text{-}IA(S, k) = \sum_{c} P(c|q) NDCG(S, k|c)$$

The same aggregation method can be applied to other IR measures as well, such as *Mean Reciprocal Rank* (MRR) and *Mean Average Precision* (MAP).

Fianlly, a redundancy-aware variation of the traditional *precision* and *recall* measures is considered in [19]:

$$Redundancy\text{-}Precision = \frac{R^-}{R^- + N^-}$$

and

$$Redundancy\text{-}Recall = \frac{R^-}{R^- + R^+}$$

where $R^-$ is the set of non-delivered redundant documents, $N^-$ is the set of non-delivered non-redundant ones and $R^+$ is the set of delivered redundant ones.

Besides deriving appropriate measures, user studies are also central in evaluating the usefulness of diversification. In a recent study, two thousand volunteers from the BookCrossing[2] community were asked to rate recommendations produced by using diversification techniques [20]. The results vary according to the method used to acquire the initial recommendations, but overall users rated the diversified recommendations higher than the non-diversified ones in all cases, as long as diversity contributed up to 40% to the linear combination of the relevance and diversity measures. A higher contribution led to a lower overall rating by the users. An interesting finding is that, when diversified results were presented to the users, the individual recommendations where generally rated lower but the overall rating of the recommendation list as a whole was higher.

## 6. CONCLUSIONS

In this work, we presented the various definitions of the result diversification problem proposed in the research literature and classified them into three main categories, namely content-based, novelty-based and cover- age-based. These three factors are closely related and, therefore, most related work considers more than one of them. We also reviewed different approaches taken for the combination of diversity with other ranking criteria, most commonly that of relevance, to the user's information need. We classified the algorithms used in the literature for locating diverse items into two main categories (greedy and interchange) and also discussed other used approaches. Finally, we showed how diversity is evaluated.

## 7. REFERENCES


[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM*, pages 5–14, 2009.

[2] J. G. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336, 1998.

[3] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR*, pages 659–666, 2008.

[4] M. Drosou and E. Pitoura. Diversity over continuous data. *IEEE Data Eng. Bull.*, 32(4):49–56, 2009.

[5] M. Drosou, K. Stefanidis, and E. Pitoura. Preference-aware publish/subscribe delivery with diversity. In *DEBS*, 2009.

[6] E. Erkut, Y. Ülküsal, and O. Yeniçerioglu. A comparison of -dispersion heuristics. *Computers & OR*, 21(10):1103–1113, 1994.

[7] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *WWW*, pages 381–390, 2009.

[8] J. R. Haritsa. The KNDN problem: A quest for unity in diversity. *IEEE Data Eng. Bull.*, 32(4):15–22, 2009.

[9] R. Hassin, S. Rubinstein, and A. Tamir. Approximation algorithms for maximum dispersion. *Operations Research Letters*, 21(3):133 – 137, 1997.

[10] G. Koutrika and Y. E. Ioannidis. Personalized queries under a generalized preference model. In *ICDE*, pages 841–852, 2005.

[11] B. Liu and H. V. Jagadish. Using trees to depict a forest. *PVLDB*, 2(1):133–144, 2009.

[12] K. Liu, E. Terzi, and T. Grandison. Highlighting diverse concepts in documents. In *SDM*, pages 545–556, 2009.

[13] Z. Liu, P. Sun, and Y. Chen. Structured search result differentiation. *PVLDB*, 2(1):313–324, 2009.

[14] F. Radlinski and S. T. Dumais. Improving personalized web search using result diversification. In *SIGIR*, pages 691–692, 2006.

[15] K. Stefanidis, M. Drosou, and E. Pitoura. PerK: personalized keyword search in relational databases through preferences. In *EDBT*, pages 585–596, 2010.

[16] E. Vee, U. Srivastava, J. Shanmugasundaram, P. Bhat, and S. Amer-Yahia. Efficient computation of diverse query results. In *ICDE*, pages 228–236, 2008.

[17] C. Yu, L. V. S. Lakshmanan, and S. Amer-Yahia. It takes variety to make a world: diversification in recommender systems. In *EDBT*, pages 368–378, 2009.

[18] M. Zhang and N. Hurley. Avoiding monotony: improving the diversity of recommendation lists. In *RecSys*, pages 123–130, 2008.

[19] Y. Zhang, J. P. Callan, and T. P. Minka. Novelty and redundancy detection in adaptive filtering. In *SIGIR*, pages 81–88, 2002.

[20] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *WWW*, pages 22–32, 2005.


[2] http://www.bookcrossing.com