Big Data
Volume 5 Number 2, 2017

DOI: 10.1089/big.2016.0054

REVIEW

Diversity in Big Data: A Review

Marina Drosou[1], H.V. Jagadish[2], Evaggelia Pitoura[1], and Julia Stoyanovich[3,*]

Abstract

Big data technology offers unprecedented opportunities to society as a whole and also to its individual members. At the same time, this technology poses significant risks to those it overlooks. In this article, we give an overview of recent technical work on diversity, particularly in selection tasks, discuss connections between diversity and fairness, and identify promising directions for future work that will position diversity as an important component of a data-responsible society. We argue that diversity should come to the forefront of our discourse, for reasons that are both ethical—to mitigate the risks of exclusion—and utilitarian, to enable more powerful, accurate, and engaging data analysis and use.

Keywords: data; diversity; empirical studies; models and algorithms; responsibly

Introduction

Big data technology holds incredible promise of improving people's lives, accelerating scientific discovery and innovation, and bringing about positive societal change. However, if not used responsibly, the same technology can exacerbate economic inequality, destabilize global markets, and reaffirm systemic bias. Considerable research attention is being paid to aspects of responsible data analysis and use such as fairness, accountability, transparency, and privacy. An important aspect that is often overlooked is *diversity*: ensuring that different kinds of objects are represented in the output of an algorithmic process. Unlike properties such as accuracy and relevance, which are usually assigned to individual items, diversity is an aspect of quality of a *collection* of items.

Diversity is widely acknowledged to be important in a range of contexts: It has been studied extensively in the social and biological sciences, with *index of diversity*, "a measure of the degree of concentration or diversity achieved when the individuals of a population are classified into groups," appearing as one of the first definitions in 1949.[1] More recently, sociologists and political scientists are studying the benefits of diversity both to small groups and to society as a whole,[2,3] whereas policy and legal scholars are drawing attention to the risks that big data technology poses to those it overlooks.[4–6] Diversity also appears in technical contexts as a means for "hedging bets." For example, search engines and recommender systems prefer to return a collection of diverse results rather than of very similar high-scoring results. Consequently, diversity has been considered extensively in information retrieval (IR)[7–11] and content recommendation.[12–15]

Diversity is a critical topic for big data systems research, for reasons that are both ethical—to mitigate the risks of exclusion—and utilitarian, to make human-centered data-intensive methods more powerful, accurate, and engaging. Consider the following scenarios where a lack of diversity among the results of an algorithmic task can cause significant harms.

Hiring

Statistical models are now routinely employed to select job candidates from a pool of resumes. As team diversity is increasingly recognized as a hiring goal,[16–19] these models need to be adapted to enforce diversity. Diversity of the workforce is desirable because it affords economic opportunity and leads to social mobility for people with varying demographic characteristics, and cultural and educational backgrounds. Importantly, diversity can also lead to higher productivity

[1]Department of Computer Science, University of Ioannina, Ioannina, Greece.
[2]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan.
[3]Department of Computer Science, Drexel University, Philadelphia, Pennsylvania.

*Address correspondence to: *Julia Stoyanovich, Department of Computer Science, Drexel University, Philadelphia, PA 19104*, E-mail: stoyanovich@drexel.edu

and enhance the problem-solving ability of teams. According to Diaz-Uda et al.,[16] "Diversity's definition has changed: In addition to creating a workplace inclusive of race, gender, and sexual orientation (to name a few), many organizations are seeking value in something even simpler, diversity of thought."

Crowdsourcing

Similar to hiring, crowdsourcing tasks benefit from diversity among the workers. Diversity of opinion, the requirement that each person has private information even if it is just an interpretation of the known facts,[3] is one of the necessary conditions for a "wise crowd." The extent to which diversity matters, in utilitarian terms, depends on the specifics of the task. It has been shown in Ref.[2] that, for problem-solving tasks, diversity is more important than the average ability of individuals, a phenomenon known as the "Diversity trumps ability" theorem.

Matchmaking

Consider an online application in which users search for dating partners. A common issue in such systems is that, because of subtle correlations between item attributes and the ranking function, top-ranked results are very similar to each other (e.g., are all between 35 and 40 years old), and they are not representative of the user's preferences (e.g., matches between 20 and 40 years old).[20] This is problematic because users do not find the results engaging. Perhaps a more severe effect is that opportunities are limited for the individuals who appear low in the list, despite matching the user's search criteria and being of potential interest.

Search and content recommendation

Poor diversity in algorithmic news feeds exacerbates "bubble effects"[21] that can polarize public opinion, increase ideological segregation, and, ultimately, undermine the democratic process. In Web search,[7,9] poor result diversity leads to poor user engagement, caters only to the most common information needs and fails to address less typical ones, and limits exposure of the "long tail" of results. In both domains, result diversity should be enforced to mitigate these problems.

Lack of diversity can have legal and ethical implications, and additional work is needed to support diversity in a data-responsible society. The goal of this survey is to initiate a discussion of a research agenda to mitigate the risks that a lack of diversity poses.

At the same time, we recognize that there are some circumstances where diversity may not be desired—for example, diverse opinions may not be helpful in determining answers to questions of fact. Even otherwise, there is usually some measure of utility or goodness that has to be combined with diversity. Although simply maximizing a utility score and ignoring diversity is usually not desirable, it is also usually not desirable to maximize diversity and ignore utility completely.

The remainder of this article is organized as follows. We outline the scope of this survey in Section 2. In Section 3, we present methodologies and results of empirical studies that measure the diversity of information produced and consumed by users in online social networks. In Sections 4 and 5, we discuss the state of the art in diversity models and algorithms. We discuss the trade-off between diversity and fairness in Section 6. We conclude this survey in Section 7, where we also outline directions for future work.

Scope

Diversity is a general term used to capture the quality of a collection of items, or of a composite item, S, with regards to the *variety* of its constituent elements. Depending on the application, S may stand for a set of people in a crowd-formation application, a bundle of items in a recommendation system, or a ranked list of Web search results. The elements of S may be unstructured (e.g., documents or news articles), structured (e.g., database tuples), or even complex objects (e.g., subgraphs).

The main focus of our survey is on the *selection task*—identifying a set or a bundle of elements that meet the quality or relevance criteria and that are also diverse. In many applications, the collection is ordered on a per-item quality score. We present diversity models in Section 4, and we explain how they are used for both unordered and ordered collections of results.

In this article, we survey models and algorithms that enforce diversity in the *output* of an algorithmic task. This is in contrast to diversity of the *input*, an important topic in machine learning that has been studied as volume-based diversity.[22] In volume-based diversity, the goal is to generate a diverse sub-sample from which to train a predictive model, or a diverse summary of a dataset. Input dataset X is made up of points in a continuous multi-dimensional space, and the objective is to select a set S of k points that maximizes the volume of the k-dimensional parallelepiped formed by the corresponding vectors in S. This problem formulation

gives rise to a probability distribution over subsets of X known as determinantal point processes, for which efficient sampling methods were developed in recent literature.[23,24] A recent approach has been proposed to combine volume-based diversity with so-called combinatorial diversity—a measure of entropy over a single discrete low-cardinality attribute.[25]

There is a rich body of literature in IR on diversity-based and novelty-based ranking of text documents or Web search results.[7–11] The primary motivation behind diversity work in IR is resolving query ambiguity:[10] When a user specifies the query "jaguar," which meaning of the term do they have in mind? Beyond resolving query ambiguity, novelty and serendipity are also considered important in user satisfaction with search results. Diversity work in IR uses a wide variety of techniques, including modifications of relevance and utility models, query reformulation, and optimization frameworks that directly incorporate diversity and novelty objectives. We use several examples from the IR domain, and also discuss some important models and algorithms. However, a comprehensive treatment of diversity in the IR literature is out of scope of the present survey.

Like fairness,[26–28] diversity is inherently a socio-technical concept that gives rise to a multitude of interpretations. Consequently, it is unlikely that a universal technical definition of diversity will emerge. An appropriate way to represent diversity requirements, to compute a result that meets those requirements, and to define the trade-offs between competing diversity objectives, and between diversity, utility, and fairness, depends on the application, and on the context of use.

Empirical Studies

Diversity is important in many scenarios, and it is discussed and studied in many contexts.

One scenario of particular interest is in employee hiring, especially when selection is at least partially based on algorithmic scoring of applicants, whether based on a supplied resume, on aptitude tests, or on other big data sources, such as social media presence. Many companies, particularly in high-tech industries, use big data tools to shortlist employment candidates, and they are also concerned about increasing the diversity of their workforce. See, for example, a popular press survey.[29]

Unfortunately, systematic investigation of diversity in data-rich algorithmic environments is in its early stages. We are not aware of rigorous empirical work on diversity in data-driven hiring algorithms. However, a number of empirical studies have been conducted to investigate diversity in the way that people are *exposed to* (what they see) and *interact with* (what they propagate, or "share") online information. We outline a few such studies in this section.

A common concern is that users tend to be connected with like-minded people in social networks, a practice possibly leading to the creation of "information bubbles" or "echo chambers"—a situation where users are not exposed to diverse information in their feeds. Facebook recently conducted a study on the matter[21] by using a subset of its users who self-identified as liberals or conservatives. Users share content, such as news articles or opinions. The authors of the study classified stories as either "hard" (such as politics, national news, or world affairs) or "soft" content (such as sports, entertainment, or travel) by training a support vector machine on unigram, bigram, and trigram text features. Self-reported ideological affiliations of the users who shared a particular piece of content were averaged to derive ideological alignment of the piece.

The authors of the study observed substantial polarization among hard content shared by users, with the most frequently shared links clearly aligned with largely liberal or conservative populations. One of the striking findings of the study was that, of the hard news stories shared by liberals' friends, 24% were cross-cutting, compared with 35% for conservatives. To provide context, consider that if an individual acquired information from random others, between 40% and 45% of the content would be cross-cutting. Overall, it was established that about 30% of shared news content and about 29% of consumed (seen) content cuts across ideological lines. Finally, about 25% of the content on which users click in their feeds cuts across ideological lines.

According to the authors of the study, "Although partisans tend to maintain relationships with like-minded contacts, on average more than 20% of an individual's Facebook friends who report an ideological affiliation are from the opposing party, leaving substantial room for exposure to opposing viewpoints." Further, on comparing both availability of content and content consumption, the authors concluded that individual choices, more than algorithms, limit exposure to attitude-changing content in the context of Facebook. That said, additional studies are needed to inform the discussion about how our media exposure and attitude formation are shaped by our social networks.

Even if one assumes that all content generated by a user's friends is satisfactorily propagated and seen throughout the network, there is still the concern of how diverse the social network around a specific user actually is.[30] A 2009 study on Facebook data[31] determined that users maintain a relationship with no more than 10% of their friends, irrespective of the size of their friendship group. Here, maintaining a relationship is broadly construed, and it corresponds primarily to consuming content shared by a user's friends.

Another study[32] investigated the diversity of content to which social media users are exposed. The study found that people who actively seek out news and information from social media are at a higher risk of becoming trapped in "information bubbles" compared with those who use search engines, in the sense that they are exposed to a significantly narrower spectrum of information sources. The study was conducted on a large collection of Web click data spanning a three-and-a-half-year period, and it showed that the diversity of information sources reached from social media is significantly lower than that of those reached from search engines. Diversity of an information source s was measured by considering all targets (Websites) t reachable from s, and factoring in the amount of traffic (clicks) from s to t. That is, the study was based on traffic patterns rather than on content analysis.

The Web click data used in Ref.[32] were anonymized and so did not distinguish between users, and they were used to establish the presence of a collective "information bubble." However, low collective diversity does not necessarily mean that individual diversity is also low. That is, low collective diversity could be consistent with high individual diversity, and vice versa. To investigate individual diversity further, the authors used two additional datasets—link shares on Twitter and AOL—and showed a strong correlation between collective diversity and average user diversity. This finding suggests that information sharing and consumption in social media give rise to both a "collective bubble" and "individual bubbles."

Another important question is: *How* and *by whom* is diverse content being generated? Does the popularity of the sources generating diverse content affect its propagation? A study on Twitter data[33] showed that hashtags (or topics) that become popular are those being adopted by a diverse audience at the same time. One interpretation of this finding is that audience diversity makes it more likely that a hashtag will reach more users. On the other hand, the study found that users gain more followers—and greater social impact—by tweeting only on a specific topic. In other words, highly popular users are not likely to contribute toward diversity. As summarized by the authors of the study: "In short, diverse messages and focused messengers are more likely to gain impact." In this study, as observed in Refs.[21,32] topical diversity of content and of users was determined based on network structure and on information directly reported by the users, and not on content analysis.

The studies discussed here are important both because of the specific insights they bring and because they develop methodologies for measuring diversity of information in the social Web environment. Additional studies are certainly warranted in this space, proposing alternative methodologies and focusing on other contexts, such as hiring and matchmaking. Such studies contribute to making large-scale algorithmic processes transparent and help hold the dominant online platforms that are accountable for the societal effects they create.[34]

Importantly, conclusions that can be drawn from studies that are based solely on data traces, and that have no knowledge of the underlying algorithmic process, are limited. We discuss the state of the art in diversity models and algorithms in the following sections.

Models

In this section, we present several formal models of diversity. Different applications pose different diversification objectives, but let us start with the following simple formulation of the *diversification problem* for the *selection task*. Given a measure of diversity *div*, a set $\mathcal{I}$ of n elements, and an integer k, $k \leq n$, we want to form a set S containing k of the n elements of $\mathcal{I}$, such that:

$$S = \underset{S' \subseteq \mathcal{I},\ |S'|=k}{\operatorname{argmax}} div(S') \quad (1)$$

In the rest of this section, we first discuss extensions of this problem that incorporate utility, ranking, and aggregate diversity. Then, we present alternative definitions for the diversity measure *div*. Although there is no single formal definition of *div*, the various measures can be roughly classified as (a) distance based, (b) coverage based, and (c) novelty based.[35]

Utility and ranking

In most cases, diversity is just one of many requirements. For example, in a search, we want results that are not only diverse but also relevant to the information

needs of the user. In recommendation systems, besides being diverse, recommendations should also be accurate, meaning that the predicted user rating of an item must be close to the actual user rating. When forming crowds or teams, the appropriateness of a candidate for the specific task is also central to the selection process, in addition to diversity.

Thus, in general, a compromise must be made between diversity and some form of application-dependent quality or *utility*. Then, given a measure of utility $u(S)$, the overall objective for forming S should combine requirements regarding both the utility $u(S)$ and the diversity $div(S)$. Observe that, although utility is a property of a single item, this is not the case for diversity. For example, we can measure the relevance of a single document, the accuracy of a single recommendation, or the aptitude of a single job candidate; whereas measuring diversity must be done with respect to one or more other documents, recommendations, or candidates, respectively.

In the presence of utility, the input $\mathcal{I}$ in our problem formulation (Equation 1) may not be a set but instead a ranked list of elements that are ordered by utility. Furthermore, the output elements in S must be selected so that they have both high utility and high diversity. Diversity measures presented in the remainder of this section and diversification algorithms in Section 5 refer to a set of results S. Clearly, the results in S can be sorted by utility. Furthermore, in some cases, S is formed incrementally by selecting at each step an item that, if added to the current set, will maximize some measure f. The order in which items are added to S induces a ranking.

A common way of combining utility and diversity is through a function f on both u and div. A function that is used frequently is some linear combination of u and div, for example, $f(S) = \alpha\, u(S) + (1-\alpha)\, div(S)$, where parameter α, $0 \leq \alpha \leq 1$, controls the relative importance of diversity and utility. Setting $\alpha = 1$ results in selecting elements solely on utility, whereas setting $\alpha = 0$ selects elements solely on diversity. The authors cited in Ref.[36] consider a set of natural axioms that an appropriate function f for combining utility and diversity must satisfy. Interestingly, they show that there exists no function f that satisfies all of them.

Note, that it may also be possible to define a *goodness* measure f that captures both utility and diversity without explicitly referring to a separate measure of utility and a separate measure of diversity. When a combined measure f of utility and diversity is used, we can formulate our problem as selecting the set of k elements with the largest f value.

Besides looking for a single unified measure, another way to combine utility and diversity is by maximizing one of the two desired properties, subject to a threshold on the other side. For example, given a constant threshold on the minimum utility that the selected elements must satisfy, we look for the set S that maximizes diversity. Conversely, given a threshold on the value of diversity, we may ask for the set S with the maximum utility. As a specific example, take hiring, where we may be given as input the number of qualifying candidates (i.e., candidates with acceptable utility) and want to select a diverse subset of them. As another example, in crowdsourcing, we may express a requirement regarding the diversity of our workers, and then select the best among those who satisfy the diversity requirement.

Aggregate diversity

So far, we considered forming a single set (or ranked list) S. An interesting type of diversity is *aggregate diversity*, where the goal is to return several such sets. For example, take recommendations, where each user receives a different ordered list S of recommended items. In this case, we may want to increase both the diversity of the items recommended to each individual user and the diversity of the recommended items across users, to avoid recommending the same items to all users.[37] Other possible examples include matchmaking, where we may want to diversify the proposed dating partners across users, and graduate school admissions, where we may opt to diversify across schools to avoid having the exact same admissions lists.

In these cases, the elements of $\mathcal{I}$ in Equation 1 are not individual items but sets (or ranked lists) of items, and we ask to select the k most diverse among them.

Distance-based measures

Distance-based diversity measures rely on a pairwise distance (or similarity) measure between the elements of S (see Refs.[9,14,15,38,39] and the S-model in Ref.[40]). Given such a pairwise measure $d(i,j)$ between two elements i and j of S, the diversity $div(S)$ is expressed by using an aggregation function of the pairwise distances between its elements. Specifically, the diversity of S is most commonly defined as either the *average*

$$div(S) = \frac{1}{|S|} \sum_{i,j \in S,\ i \neq j} d(i,j), \tag{2}$$

or the *minimum*

$$div(S) = \min_{i,j \in S,\ i \neq j} d(i,j) \quad (3)$$

distance between the elements of S.

Maximizing the average diversity is known as MAX-SUM diversification, whereas maximizing the minimum diversity is known as MAXMIN diversification. MAXMIN diversification can be seen as an instance of the p-dispersion problem, a well-studied problem in operations research. The goal of the p-dispersion problem is selecting p out of n given locations for placing facilities, so that the minimum distance between any pair of the chosen locations is maximized.[38]

The definition of an appropriate pairwise distance measure is key for effective diversification and is highly application dependent. For example, distance d may correspond to a distance between feature vectors under a mapping of $\mathcal{I}$ into a feature space. The definition of the distance measure also affects the efficiency of the diversification algorithm. For example, it has been shown that when *div* is a *metric*, there are efficient approximation algorithms for the diversification problem.[36]

To showcase the flexibility of distance-based measures, we present two examples. In the first one, diversification is considered in the context of structured databases, where each element of the ground set $\mathcal{I}$ is a tuple with m attributes A.[39] For instance, take a set $\mathcal{I}$ of cars, where attributes correspond to features such as color, brand, the existence of cruise control, navigation system, etc. The goal is to present to the user a subset S of k cars that not only satisfy the requirements of the user (i.e., the utility requirements) but also have different feature values. The interesting aspect of this model is that the attributes are ordered, say $A_1 \prec \cdots \prec A_m$, based on their importance with regards to diversification. For example, *brand* $\prec$ *color* means that the user would like to see cars of all different brands and then, when all brands are covered, cars with different colors. Similarity is defined by using prefixes. A *prefix with respect to* $\prec$, denoted ρ, is a sequence of attribute values in the order given by $\prec$, moving from higher to lower priority. If ρ is a prefix of length l, the similarity $sim_\rho(i,j)$ between two tuples i and j is equal to 1 if i and j agree on their $l+1$ attribute, and it is 0 otherwise. A set $S \subseteq \mathcal{I}$ is called diverse with respect to ρ if

$$S = \operatorname*{argmin}_{S' \subseteq \mathcal{I},\ |S'|=k} \sum_{i,j \in S'} sim_\rho(i,j) \quad (4)$$

S is called a diverse result set if it is diverse with respect to every prefix.

The second example comes from recommendation systems and measures the distance between recommendations based on their *explanations*.[14] To recommend an item i to a user u, content-based recommenders exploit how u has rated items that are similar to i, whereas recommenders based on collaborative filtering exploit the ratings of i by users similar to u. Thus, given a set of items $\mathcal{I}$ and a set of users $\mathcal{U}$, the explanation $expl(u,i)$ of an item $i \in \mathcal{I}$ recommended to a user $u \in \mathcal{U}$ is defined as: (a) the set of items similar to i that user u rated in the past, in a content-based approach, and (b) the set of users similar to u who rated item i, in a collaborative filtering approach. The pairwise distance between two items i and j recommended to user u is defined as the distance (e.g., Jaccard or cosine) between $expl(u,i)$ and $expl(u,j)$. The more diverse the explanations based on which the recommendation of an item is made, the more diverse the item.

Coverage-based measures

Coverage-based diversity measures rely on the existence of a predefined number of aspects, that is, topics, interpretations, or opinions (see Refs.[7,10,41,42] and the T-model in Ref.[40]). For example, these aspects may correspond to interpretations of a keyword in a query, or to different skills in team formation. In coverage-based approaches, the diversity $div(S)$ of S measures the extent to which the elements of S cover these aspects. The coverage of an aspect is often represented by using a probabilistic model, for example, in Refs.[7,10,43] and the T-model in Ref.[40] Consequently, the corresponding measure of diversity is a probabilistic one.

Coverage-based models differ along several dimensions, including (a) what are the aspects to be covered, and (b) how coverage is measured. We now present a few examples. The first two examples refer to a search in IR, where the ground set $\mathcal{I}$ is a set of documents.

In the first example, the authors assume the existence of topic categories.[7] Given a user search query q, $P(c|q)$ denotes the probability that query q belongs to topic c, and $V(i|q,c)$ denotes the probability that document i satisfies user intent c given query q. In a sense, V captures how well document i covers topic c. Analogously, p expresses the popularity of the different interpretations (topics) of query q. Note that in this example of coverage-based diversity, covering more common interpretations (topics) of a query is considered more important than covering less popular ones. The

overall goal is to maximize the probability that the average user finds at least one useful result in the selected set. This is expressed by the following objective: Given a query q, the set of documents $\mathcal{I}$, and an integer k, find $S \subset \mathcal{I}$, $|S| = k$, with the maximum

$$\sum_c P(c|q)(1 - \prod_{i \in S}(1 - V(i|q, c))) \quad (5)$$

In the second example, the authors consider S as an ordered list of documents.[10] Both documents and queries are modeled as sets of *information nuggets*. A document i is considered relevant to a given query q if i contains at least one of the nuggets of q. Given an ordered list S of documents for q, the probability that the m-th document is diverse from the $m-1$ preceding ones is equal to the probability that document m contains a nugget that cannot be found in the previous $m-1$ documents.

In the final example, diversity is seen as a means to counteract the "tyranny of the majority." The authors consider diversity in the context of news and opinions aggregators, where news articles are ordered by the number of users who have voted for them.[41] Three measures of diversity are proposed, based on how many users are covered by the presented articles. Intuitively, a user is covered by an ordered list S of articles, if S contains articles voted for by the user. *Inclusion* diversity measures the proportion of users who have at least one voted-for item in S. *Alienation* diversity counts, for each user u, the position of u's first voted-for item in S, and outputs the sum of these counts for all users. Finally, *proportional representation* considers that users and articles are partitioned into groups, and asks that groups are proportionally covered by the articles in S. This is achieved through a divergence scoring function that is minimized when the result set S has votes from different groups in proportion to the representation of the group in the voter population.

Hybrid distance-based and coverage-based measures

The distinction between distance-based and coverage-based approaches is not absolute. In some cases, it may be possible to express coverage with an appropriately defined pairwise distance function. Furthermore, coverage-based and distance-based approaches may be combined.

An example that combines distance and coverage is *r-DisC diversity*.[44,45] Instead of specifying the required size k for the set S, r-DisC takes as input a real number r, $r \geq 0$, called *radius*. Given a pairwise distance measure d between the elements of $\mathcal{I}$, the diverse set S is formed such that (i) (coverage condition) for each $i \in \mathcal{I}$, there is a $j \in S$, such that $d(i, j) \leq r$ and (ii) (dissimilarity condition) for each pair $i, j \in S$ with $i \neq j$, it holds that $d(i, j) > r$. The first condition ensures that all items in $\mathcal{I}$ are represented by at least one similar item in S, and the second condition ensures that the items in S are dissimilar to each other. By decreasing the value of r, we get larger sets of items with less diversity; whereas by increasing the value of r, we get sets that are smaller and more diverse.

Novelty-based measures

Novelty-based diversity measures define diversity with respect to the elements seen in the past (Refs.[9,15,46]). Given the past, we select elements that are diverse (distance based or coverage based) from these past elements. The main goal of novelty-based diversity in search and recommendation is to reduce *redundancy*.

In novelty-based approaches, the elements are often selected one at a time. Let us denote with $div(S, i)$ the diversity of an element i with respect to a set S. Similar to the definition of diversity of a set S, $div(S, i)$ may be distance based (defined as either the average or the minimum distance of i from the elements in S) or coverage based.

A common formulation of a novelty-based approach is as follows. Given a measure of diversity div between an element and a set, a set $\mathcal{I}$ of n elements, and a set p of previously selected items, select the element e such that

$$e = \underset{i \in \mathcal{I} \setminus P}{argmax}\ div(P, i) \quad (6)$$

An early example of novelty-based diversity used in IR is maximal marginal relevance (MMR).[9] Given a query q and a set p of already seen documents, *MMR* selects a document d based on the utility of d, expressed as its similarity with query q, and the diversity between d and the previously seen documents in p, expressed as the negative of its similarity with p. The utility and diversity scores of d are combined in a weighted manner, with λ ($0 \leq \lambda \leq 1$).

$$d = \underset{i \notin P}{\text{argmax}}\{\lambda sim(i, q) - (1 - \lambda) \max_{j \in P}\{sim(i, j)\}\} \quad (7)$$

Sometimes, novelty is related with *popularity*. For instance, one can approximate novelty by selecting the most unpopular item, since this is an item that the

user has probably not seen in the past.[15] A special type of novelty is *serendipity*,[47] where the goal is to select the most unusual or surprising elements.

Algorithms

Several algorithms have been proposed for locating diverse elements. Often, the proposed algorithms exploit specific properties of the underlying diversity problem and/or application. However, we can distinguish two major groups of algorithms that are widely used throughout the literature, namely, greedy algorithms and interchange algorithms. Next, we present such approaches along with others found in the literature, such as methods exploiting graphs or optimization techniques.

In most cases, locating an optimal diverse subset is an NP-hard problem (e.g., S-model in Refs.[7,38, 40,42,48]). Therefore, the vast majority of algorithms used in the literature do not produce optimal solutions. However, in some special cases, such as the 1-dimensional p-dispersion problem[49] or the T-model in Ref.[40] the selection of diverse subsets can be solved optimally in polynomial time.

Greedy algorithms

Greedy approaches are the most commonly used for all diversity definitions, that is, distance based, coverage based, and novelty based, either when aiming at maximizing diversity alone or along with some notion of utility or ranking.

Algorithm 1 presents a generic greedy algorithm that employs two sets of elements, namely (i) the set $\mathcal{I}$ of all n available elements and (ii) the set S of selected (or diverse) elements. Elements are iteratively moved from $\mathcal{I}$ to S until $|S|=k$ and $|\mathcal{I}|=n-k$.

At first, S is initialized with some elements. A number of variations exist, such as selecting a random element, as in Algorithm 1, the pair of the most distant elements or the element maximizing utility. Then, elements are moved, one at a time, from $\mathcal{I}$ to S until k of them have been selected. The element that is moved each time is the element i that has the maximum $div(S,i)$ from S, where $div(S,i)$ is defined based on the model. Ties are usually broken arbitrarily or according to the elements' utility if such information is available.

Greedy algorithms have been used, for example, in Ref.[38] and in the S-model in Ref.[40] considering $div(S,i)$ as simply the minimum or average distance of i from S. In Ref.[7] a greedy algorithm is used for the coverage-based model that bases $div(S,i)$ on diversity and relevance, whereas in Ref.[41] a greedy algorithm is used for the coverage-based model that bases $div(S,i)$ on diversity and popularity. Greedy algorithms have been used for novelty-based models as well, as in Refs.[9,15] where diversity is combined with utility and ranking.

Algorithm 1 Generic Greedy Algorithm.

Require: $\mathcal{I}$, an integer k.
Ensure: A set S with the k most diverse items of $\mathcal{I}$.

1: $i \leftarrow$ random element in $\mathcal{I}$
2: $S \leftarrow \{i\}$
3: **while** $|S| < k$ **do**
4: $\quad i^* \leftarrow \text{argmax}_{i \in \mathcal{I}}\, div(S,i)$
5: $\quad S \leftarrow S \cup \{i^*\}$
6: **end while**
7: **return** S

Neighborhood algorithms. Neighborhood algorithms constitute a special case of greedy algorithms. Algorithm 2 presents a generic algorithm in this class that starts with a solution S initialized in some way and then iteratively adds elements to S until $|S|=k$. In this case, however, the elements to be considered at each iteration are limited based on the notion of the *r-neighborhood* of an element $i \in \mathcal{I}$, $N(i,\mathcal{I},r)$, defined as:

$$N(i,\mathcal{I},r)=\{j \in \mathcal{I} : d(i,j) \leq r\}$$

In a nutshell, at each iteration, elements that are inside the r-neighborhoods of any already selected elements (i.e., are "close" to the already selected elements) are disqualified from further consideration. Out of the remaining elements, one is selected to be added to S. Again, there are a number of variations; for example, we may select an element with (i) the highest average distance to the already selected elements (as in MaxSum) or (ii) the largest minimum of distances to the already selected elements (as in MaxMin). Note that, given a specific value of r, a solution S with $|S|=k$ may not exist.

Neighborhood algorithms have been mainly used with distance-based models, for example, in Ref.[38] A similar approach is also used in Ref.[45] where elements are selected in rounds, and each selected element leads to the disqualification of its neighborhood. However, in that case, r may be different for each element and period. The employed neighborhood algorithm implementation used spatial indexes (M-Trees) and pruning rules to speed up computation.

Algorithm 2 Generic Neighborhood Algorithm.

Require: A set of items $\mathcal{I}$, r.
Ensure: A set S with the most diverse items of $\mathcal{I}$.

$i \leftarrow$ random element in $\mathcal{I}$
$S \leftarrow \{i\}$
$\mathcal{I} \leftarrow \mathcal{I} \setminus \{j : j \in N(i, \mathcal{I}, r)\}$
while $|\mathcal{I}| > 0$ **do**
  $i \leftarrow$ select element in $\mathcal{I}$
  $\mathcal{I} \leftarrow \mathcal{I} \setminus \{j : j \in N(i, \mathcal{I}, r)\}$
  $S \leftarrow S \cup \{i\}$
end while
return S

Interchange algorithms

Another major category of diversity algorithms used for all types of diversity models are the interchange (or swap) algorithms, which are shown in Algorithm 3. Such algorithms are typically initialized with a random solution S of size k, and they then iteratively attempt to improve that solution by interchanging an element in S with an element in $\mathcal{I} \setminus S$. Again, different variations exist, such as selecting (i) the first located element that improves the diversity of S or (ii) the element that improves the diversity of S the most. Thus, interchange algorithms are, in essence, local search algorithms that aim at identifying a locally optimal solution.

For example, in Ref.[14] an interchange algorithm is used for a distance-based model that combines diversity with utility (in the form of relevance). S is initialized with the k most relevant elements. Then, at each iteration, the element of S that contributes the least to the diversity of S, that is, the one with the minimum $div(S, i)$, is interchanged with the most relevant element in $\mathcal{I} \setminus S$. Interchanges stop when there are no more elements in $\mathcal{I} \setminus S$ with higher relevance than a given threshold. As another example, Ref.[48] uses an interchange algorithm to achieve coverage-based diversity over structured data. The goal is to find a subset of attributes that can best differentiate between tuples, by starting with a random subset of attributes and iteratively interchanging one of the attributes with a better candidate.

Algorithm 3 Generic Interchange Algorithm.

Require: A set of items $\mathcal{I}$, an integer k.
Ensure: A set S with the k most diverse items of $\mathcal{I}$.

$S \leftarrow$ random solution
while changes are made to S **do**
  $i^* \leftarrow \text{argmin}_{i \in S} div(S \setminus \{i\}, i)$
  for all $z \in \mathcal{I} \setminus S$ **do**
    $S' \leftarrow \{S \setminus \{i^*\}\} \cup \{z\}$
    If $div(S') > div(S)$ **then**
      $S \leftarrow S'$;
    end if
  end for
end while
return S

Graph algorithms

A graph-based algorithm based on random walks, called GRASSHOPPER, is presented in Ref.[50] as an alternative to the novelty-based MMR model. GRASSHOPPER employs a weighted graph, which models both diversity and relevance of data items. Nodes are annotated with a weight representing their relevance, whereas edges are annotated with a weight representing the similarity of their adjacent nodes. This graph serves as the representation of states and transitions of an absorbing Markov chain. At each iteration, a random walk is performed. The walker either moves, with some probability, to a neighbor state according to similarity—represented by edge weights—or teleports to a random state according to relevance. At the end of the walk, the selected node is turned into an absorbing state and the walk is repeated.

Another graph-based approach is considered in Ref.[51] A graph, called Affinity Graph, is constructed, in which nodes correspond to elements and edges between nodes are weighted based on the affinity between elements, which is, in turn, based on their similarity but is defined in an asymmetrical way. Generally, groups of heavily connected nodes correspond to elements containing information on some specific topic, that is, a coverage-based model is used. A Markov chain is employed to initialize the algorithm, which later uses the graph to select k elements.

Optimization algorithms

In Ref.[52] the diversification problem is formulated as an optimization one under certain constraints. A distance-based model is employed in which feasible solutions are represented with binary vectors, and both the objective and the constraints are quadratic or linear expressions over the binary vector. Utility, in the form of relevance, is also considered. More specifically, three different problems are considered: (i) Maximize diversity under the constraint that the selected items exhibit relevance greater than some threshold, (ii) maximize relevance exhibited under the constraint that diversity is greater than some threshold, and (iii) express the trade-off between the two objectives explicitly in a linear combination by using a parameter α that represents the importance given to diversity in comparison to that given to relevance.

The authors referred to in Ref.[52] use algorithms for solving binary vector optimization problems. The first and third variations correspond to binary quadratic programming problems with linear constraints,

whereas the second variation is a binary linear programming problem with quadratic and linear constraints. Thus, known optimization algorithms are employed to search the solution space, where the solutions are generated by solving relaxations of the binary problems to problems with real-valued solutions.

Another approach for the distance-based model is used in Ref.[38] where dimensionality reduction is used to project elements in one dimension. Then, a solution is approximated via problem relaxation and quantization. However, the approximation does not come with any guarantees and also does not perform well empirically, and so it can only be used in a multiple-tries setting in practice.

In Ref.[40] the introduced coverage-based T-model is formed as an optimization problem in which the user sets specific demands on the coverage of the various topics. The optimization is to select the best subset of elements such that the user's demand is satisfied. The problem is shown to be equivalent to the *Exact k-item Knapsack Problem* (E-kKP) and, thus, can be solved by various algorithms, that is, a backtracking approach, dynamic programming, or linear programming, as described in Ref.[53]

Other algorithms

Next, we present some other algorithms found in the literature.

In Ref.[20] the authors address the problem of lack of diversity in rankings over structured datasets. Lack of diversity at top ranks is due to subtle interactions between item attributes and the ranking function. The authors develop a rank-aware clustering framework that groups together items that are comparable with respect to the ranking function, computes a concise and informative description for each group, and, finally, ranks items within each group. Diverse ranked results keep the user more engaged, as demonstrated with a large-scale user study in a dating application.

A number of approaches, mainly aiming at achieving diversity in database systems, are based on tree index structures. For example, in Ref.[39] a Dewey tree is employed. 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 some diversity ordering of the relation, and diverse tuples are retrieved by traversing this tree. This approach can be used only with a specific diversity function on structured data that defines the diversity ordering.

A more general approach that can be used with any diversity function is presented in Ref.[54] The spatial indexing properties of Cover trees are employed for selecting items that are dissimilar to each other. This approach can also be used for the *dynamic* diversification problem, where $\mathcal{I}$ is not static but insertions and deletions of elements are allowed and, thus, the diverse set needs to be refreshed to reflect such updates.

Finally, a hybrid greedy/interchange heuristic is used in Ref.[55] in the context of dynamic data as well. In this case, a diverse subset S is located by using a greedy approach and then its diversity is further improved by performing interchanges.

Related Concepts

The term *fairness* has recently been increasingly used in the context of big data, much more so than the term *diversity*. One usually expects that fairness to under-represented groups will lead to greater diversity. For this reason, the two concepts are related.

However, fairness has multiple interpretations. For example, a recent paper by Celis et al.[25] considered a trade-off that concerns diversity of an *input* dataset, rather than of the result of an algorithmic task. The authors proposed a sampling method that combines combinatorial diversity, which measures the entropy along a low-dimensional categorical attribute, and volume-based diversity, briefly discussed in Section 2. The authors refer to combinatorial diversity as "fairness," but this use of the term is nonstandard, since fairness is typically measured with respect to an outcome, rather than quantified over the input.

The more common notions of algorithmic fairness have a problem setup that is similar to diversity. For example, when selecting job applicants to interview, or college applicants to admit, both qualifications of the candidates and fair treatment of members of legally protected groups must be considered alongside diversity. A useful dichotomy is between individual fairness—a requirement that similar individuals be treated similarly—and group fairness, also known as statistical parity, a requirement that demographics of those receiving a particular positive (or negative) *outcome*, for example, a positive or negative classification, are identical to the demographics of the population as a whole.[26]

In what follows, we will compare and contrast the statistical parity interpretation of fairness with diversity, since, like diversity, statistical parity is stated as a property of a collection of items. Let us assume that

membership in a protected group is represented by a so-called *sensitive attribute*. Further, to start, let us assume that this attribute is binary, using gender as an example and hiring as the selection task.

Statistical parity requires that the distribution of values of gender in the result of the task be the same as its distribution in the input population. Assuming that there is an approximately equal number of men and women in the input, a fair algorithm will produce an approximately equal number of men and women in the result. This outcome is the same as what most diversity models would produce.

It is, however, not always the case that fairness and diversity objectives are in agreement. Suppose that women are under-represented in the input, for example, that among the job applicants 10% are women and 90% are men. Although a fair algorithm will preserve the same gender ratio in the result, as is seen in the input, a distance-based diversity objective may require that one half of the individuals in the output belong to each gender.

Determining conditions under which fairness leads to diversity, and diversity leads to fairness, is an open research question that warrants further investigation. Casting these two objectives as part of the same framework may allow to optimize them simultaneously, leading to solutions that meet both objectives better and that are more efficient to compute.

Conclusions

In this article, we argued that diversity is an important aspect of responsible data analysis and use. We described recent empirical studies that consider the diversity of information in the social sphere, and gave an overview of existing diversity models and algorithms, with a focus on diversity of elements in the output of an algorithm.

We focused our attention primarily on diversity in a selection task. To a limited extent, we also considered diversity in the closely related ranking task. Diversity can be important in many other contexts. Further consideration is required to identify these contexts and to expand on the relevant technical work.

An important direction for future work is developing a holistic treatment of diversity through different stages of the data management and analysis life-cycle—from data cleaning, integration, and preprocessing, through selection and ranking, to result interpretation. An aspect of such a framework is support for enforcing diversity incrementally through *individual independent choices*, rather than as a constraint on the *set of final results*. Interestingly, because diversity is measured over a set, the addition or deletion of a single element to the input could radically alter the composition of the output. An important line of work, thus, concerns enabling incremental maintenance of diversity properties of a result, under monotonicity guarantees, and reasoning about stability of the selected set in response to incremental changes in the input.

Acknowledgments

This work was supported in part by the National Science Foundation Grants No. 1464327, 1539856, and 1250880, and by the US-Israel Binational Science Foundation Grant No. 2014391.

Author Disclosure Statement

No competing financial interests exist.

References

1. Simpson EH. Measurement of diversity. Nature. 1949;163:688.
2. Page SE. The difference: How the power of diversity creates better groups, firms, schools, and societies. Princeton, NJ: Princeton University Press, 2007.
3. Surowiecki J. The wisdom of the crowds. New York: Random House, Inc., 2005.
4. Barocas S, Selbst AD. Big data's disparate impact. California Law Review 2016;104:671–732.
5. Crawford K. Artificial intelligence's white guy problem. The New York Times, June 26, 2016.
6. Lerman J. Big data and its exclusions. Stanford Law Review Online, 2013;66. Available at: https://www.stanfordlawreview.org/online/privacy-and-big-data-big-data-and-its-exclusions (accessed June 7, 2017).
7. Agrawal R, Gollapudi S, Halverson A, Ieong S. Diversifying search results. In: Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9–11, 2009, pp. 5–14.
8. Capannini G, Nardini FM, Perego R, Silvestri F. Efficient diversification of web search results. PVLDB 2011;4:451–459.
9. Carbonell JG, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: SIGIR'98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, August 24–28 1998, pp. 335–336.
10. Clarke CLA, Kolla M, Cormack GV, et al. Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20–24, 2008. pp. 659–666.
11. Dang V, Croft WB. Diversity by proportionality: An election-based approach to search result diversification. In: The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR'12, Portland, OR, USA, August 12–16, 2012. pp. 65–74.
12. Kaminskas M, Bridge D. Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. ACM Trans Interact Intell Syst. 2016;7:2:1–2:42.
13. Vargas S, Castells P. Rank and relevance in novelty and diversity metrics for recommender systems. In: Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, October 23–27, 2011, pp. 109–116.
14. Yu C, Lakshmanan LVS, Amer-Yahia S. It takes variety to make a world: Diversification in recommender systems. In: EDBT 2009, 12th International Conference on Extending Database Technology, Saint Petersburg, Russia, March 24–26, 2009, pp. 368–378.

15. Ziegler CN, McNee SM, Konstan JA, Lausen G. Improving recommendation lists through topic diversification. In: Proceedings of the 14th international conference on World Wide Web, www 2005, Chiba, Japan, May 10–14, 2005, pp. 22–32.
16. Diaz-Uda A, Medina C, Schill B. Diversity's new frontier: Diversity of thought and the future of the workforce. 2013. Available online at http://dupress.deloitte.com/dup-us-en/topics/talent/diversitys-new-frontier.html (last accessed April 25, 2017).
17. Google Official Blog. Getting to work on diversity at Google. 2014. Available online at https://googleblog.blogspot.co.il/2014/05/getting-to-work-on-diversity-at-google.html (last accessed April 25, 2017).
18. Dobbin F, Kalev A. Why diversity programs fail. 2016. Available online at https://hbr.org/2016/07/why-diversity-programs-fail (last accessed April 25, 2017).
19. Rezvani S. Five trends driving workplace diversity in 2015. 2015. Available online at www.forbes.com/sites/work-in-progress/2015/02/03/20768/#58cc73dd34c91 (last accessed April 25, 2017).
20. Stoyanovich J, Amer-Yahia S, Milo T. Making interval-based clustering rank-aware. In: EDBT 2011, 14th International Conference on Extending Database Technology, Uppsala, Sweden, March 21–24, 2011, pp. 437–448.
21. Bakshy E, Messing S, Adamic LA. Exposure to ideologically diverse news and opinion on facebook. Science. 2015;348:1130–1132.
22. Kulesza A, Taskar B. Determinantal point processes for machine learning. Found Trends Mach Learn 2012;5:123–286.
23. Anari N, Gharan SO, Rezaei A. Monte carlo markov chain algorithms for sampling strongly rayleigh distributions and determinantal point processes. In: Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, June 23–26, 2016. pp. 103–115.
24. Deshpande A, Rademacher L. Efficient volume sampling for row/column subset selection. In: 51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23–26, 2010. Las Vegas, Nevada, pp. 329–338.
25. Celis LE, Deshpande A, Kathuria T, Vishnoi NK. How to be fair and diverse? CoRR, abs/1610.07183, 2016.
26. Dwork C, Hardt M, Pitassi T, et al. Fairness through awareness. In: Innovations in Theoretical Computer Science 2012, Cambridge, MA, January 8–10, 2012. pp. 214–226.
27. Romei A, Ruggieri S. A multidisciplinary survey on discrimination analysis. Knowl Eng Rev. 2014;29:582–638.
28. Zliobaite I. A survey on measuring indirect discrimination in machine learning. CoRR, abs/1511.00148, 2015.
29. Loehr A. 4 ways HR analytics can improve workplace diversity. 2015. Available online at www.cornerstoneondemand.com/rework/4-ways-hr-analytics-can-improve-workplace-diversity (last accessed April 25, 2017).
30. Zuckerman E. How diverse is your social network? how diverse should it be? 2011. Available online at www.ethanzuckerman.com/blog/2011/06/14/how-diverse-is-your-social-network-how-diverse-should-it-be (last accessed April 25, 2017).
31. Marlow C. Maintained relationships on facebook. 2009. Available online at www.facebook.com/note.php?note_id=55257228858 &ref=mf%20(2009) (last accessed April 25, 2017).
32. Nikolov D, Oliveira DFM, Flammini A, Menczer F. Measuring online social bubbles. PeerJ CompSci. 2015;1:e38.
33. Weng L, Menczer F. Topicality and impact in social media: Diverse messages, focused messengers. PLoS One 2015;10:e0118410.
34. Goodman EP, Powles J. Facebook and Google: Most powerful and secretive empires we've ever known. The Guardian, September 28, 2016.
35. Drosou M, Pitoura E. Search result diversification. SIGMOD Record 2010;39:41–47.
36. Gollapudi S, Sharma A. An axiomatic approach for result diversification. In: Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20–24, 2009, pp. 381–390.
37. Adomavicius G, Kwon YO. Improving aggregate recommendation diversity using ranking-based techniques. IEEE Trans Knowl Data Eng. 2012;24:896–911.
38. Erkut E, Ülküsal Y, Yeniçerioglu O. A comparison of p-dispersion heuristics. Comput Oper Res. 1994;21:1103–1113.
39. Vee E, Srivastava U, Shanmugasundaram J, et al. Efficient computation of diverse query results. In: Proceedings of the 24th International Conference on Data Engineering, ICDE 2008, April 7–12, 2008. Cancún, México, pp. 228–236.
40. Wu T, Chen L, Hui P, et al. Hear the whole story: Towards the diversity of opinion in crowdsourcing markets. PVLDB 2015;8:485–496.
41. Munson SA, Zhou DX, Resnick P. Sidelines: An algorithm for increasing diversity in news and opinion aggregators. In: Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, May 17–20, 2009, pp. 130–137.
42. Yang Z, Fu AWC, Liu R. Diversified top-k subgraph querying in a large graph. In: Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, June 26–July 1, 2016, pp. 1167–1182.
43. Santos RLT, Macdonald C, Ounis I. Exploiting query reformulations for web search result diversification. In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, April 26–30, 2010, pp. 881–890.
44. Drosou M, Pitoura E. DisC diversity: Result diversification based on dissimilarity and coverage. PVLDB 2012;6:13–24.
45. Drosou M, Pitoura E. Multiple radii DisC diversity: Result diversification based on dissimilarity and coverage. ACM Trans Database Syst. 2015;40:4.
46. Lathia N, Hailes S, Capra L, Amatriain X. Temporal diversity in recommender systems. In: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19–23, 2010. pp. 210–217.
47. Herlocker JL, Konstan JA, Terveen LG, Riedl J. Evaluating collaborative filtering recommender systems. ACM Trans Inf Syst. 2004;22:5–53.
48. Liu Z, Sun P, Chen Y. Structured search result differentiation. PVLDB 2009;2:313–324.
49. Wang DW, Kuo YS. A study on two geometric location problems. Inf Process Lett. 1988;28:281–286.
50. Zhu X, Goldberg AB, Gael JV, Andrzejewski D. Improving diversity in ranking using absorbing random walks. In: Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, April 22–27, 2007. Rochester, New York, pp. 97–104.
51. Zhang B, Li H, Liu Y, et al. Improving web search results using affinity graph. In: SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15–19, 2005, pp. 504–511.
52. Zhang M, Hurley N. Avoiding monotony: Improving the diversity of recommendation lists. In: Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys 2008, Lausanne, Switzerland, October 23–25, 2008. pp. 123–130.
53. Caprara A, Kellerer H, Pferschy U, Pisinger D. Approximation algorithms for knapsack problems with cardinality constraints. Eur J Operat Res 2000;123:333–345.
54. Drosou M, Pitoura E. Diverse set selection over dynamic data. IEEE Trans Knowl Data Eng. 2014;26:1102–1116.
55. Drosou M, Pitoura E. Diversity over continuous data. IEEE Data Eng Bull. 2009;32:49–56.

Cite this article as: Drosou M, Jagadish HV, Pitoura E, Stoyanovich J (2017) Diversity in big data: a review. *Big Data* 5:2, 73–84, DOI: 10.1089/big.2016.0054.

Abbreviations Used

E-kKP = Exact k-item Knapsack Problem
IR = information retrieval
MMR = maximal marginal relevance