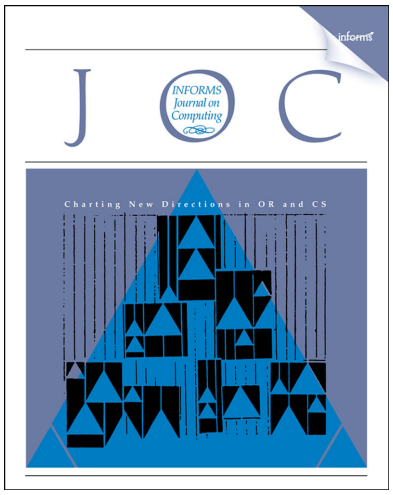


INFORMS Journal on Computing

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Content and Structure Coverage: Extracting a Diverse Information Subset

Baojun Ma, Qiang Wei, Guoqing Chen, Jin Zhang, Xunhua Guo

To cite this article:
Baojun Ma, Qiang Wei, Guoqing Chen, Jin Zhang, Xunhua Guo (2017) Content and Structure Coverage: Extracting a Diverse Information Subset. INFORMS Journal on Computing 29(4):660-675. https://doi.org/10.1287/ijoc.2017.0753

http://pubsonline.informs.org/journal/ijoc/

INFORMS JOURNAL ON COMPUTING
Vol. 29, No. 4, Fall 2017, pp. 660–675
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

Content and Structure Coverage: Extracting a Diverse Information Subset

Baojun Ma,[a] Qiang Wei,[b,*] Guoqing Chen,[b] Jin Zhang,[c] Xunhua Guo[b]

[a] School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China; [b] Research Center for Contemporary Management, School of Economics and Management, Tsinghua University, Beijing 100084, China; [c] School of Business, Renmin University of China, Beijing 100872, China

[*] Corresponding author

Contact: mabaojun@bupt.edu.cn, http://orcid.org/0000-0002-2274-3089 (BM); weiq@sem.tsinghua.edu.cn (QW); chengq@sem.tsinghua.edu.cn (GC); zhangjin@rbs.ruc.edu.cn (JZ); guoxh@sem.tsinghua.edu.cn (XG)

Received: July 9, 2014
Revised: February 2, 2016; September 26, 2016
Accepted: February 7, 2017
Published Online: August 11, 2017

https://doi.org/10.1287/ijoc.2017.0753

Abstract. Recent years have witnessed a rapid increase in online data volume and the growing challenge of information overload for web use and applications. Thus, information diversity is of great importance to both information service providers and users of search services. Based on a diversity evaluation measure (namely, information coverage), a heuristic method—*FastCov*$_{C+S}$*-Select*—with corresponding algorithms is designed on the greedy submodular idea. First, we devise the *Cov*$_{C+S}$*-Select* algorithm, which possesses the characteristic of asymptotic optimality, to optimize information coverage using a strategy in the spirit of simulated annealing. To accelerate the efficiency of *Cov*$_{C+S}$*-Select*, its fast approximation (i.e., *FastCov*$_{C+S}$*-Select*) is then developed through a heuristic strategy to downsize the solution space with the properties of information coverage. Furthermore, ample experiments have been conducted to show the effectiveness, efficiency, and parameter robustness of the proposed method, along with comparative analyses revealing the performance's advantages over other related methods.

History: Accepted by Ram Ramesh, Area Editor for Knowledge Management and Machine Learning.

Funding: Baojun Ma is supported by the National Natural Science Foundation of China (NSFC) [71402007]. Qiang Wei is supported by NSFC [71372044]. Guoqing Chen and Xunhua Guo are partly supported by the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities [12JJD630001] and NSFC [71110107027/71490724/71572092].

Supplemental Material: The online appendix is available at https://doi.org/10.1287/ijoc.2017.0753.

Keywords: information diversity • information content coverage • information structure coverage • submodularity • fast approximation

1. Introduction

With the rapid increase in online data volume, web users usually suffer from information overload and consequently face huge cost: to find the information they want or are interested in (Brynjolfsson et al. 2003, De et al. 2010). Although the typical information search service technologies (e.g., keyword search in search engines, user search functions or recommendation systems on e-commerce platforms, etc.) have been developed to help web users find information or products they need or are interested in (Mulpuru 2008), usually the volume of the result set could still be huge. For instance, Google could provide millions of links matching the keywords a user inputs, and an online shopping platform could recommend hundreds of products satisfying a query; etc. In many cases, from all such search results that satisfy users' search criteria, only a small set of results can actually be exhibited to or browsed by users (Liu 2011; Silverstein et al. 1999; Spink et al. 2002, 2001). Hence, the information quality of the small set is of great importance to both information search service providers and users.

The information quality of the small set could be measured based on various Metrics—e.g., freshness, hotness, number of visits or comments, PageRank value, helpfulness, etc.—to generate a ranking list of results accordingly. However, in light of the whole picture of all the original search results, these metrics and respective search methods often lack diversity consideration to effectively and efficiently reflect the overall picture (De et al. 2010, Ma and Wei 2012, Sen et al. 2006). The reason for such a consideration stems from the fact that in many applications, users may be quite interested in how the small set represents the original one in an overall manner from different aspects (e.g., features or attributes and characteristics regarding the information/content of the search results). Examples include when consumers read product reviews to find diversified opinions on features (e.g., price, quality, and location) with sentimental polarity (e.g., positive or negative) in online shopping; search for all-sided documents for overview of an online wiki system; browse representative blog articles on a travel website to obtain an overall idea about attractive spots

for travel planning; and so on. Apparently, because of the limitation in users' browsing time and cognitive cost as well as in device functionality (e.g., small screen size of mobile phones), consumers end up with either a small number (e.g., the top 10 or 20 results) or they stop after browsing one to two web pages. In this regard, how the information of the original set is covered by the small set in light of diversity is an important issue for both academia and practitioners, giving rise to various research efforts (Agrawal et al. 2009, Carbonell and Goldstein 1998, Carterette and Chandar 2009, Chen and Karger 2006, He et al. 2011, Santos et al. 2010, Wang and Zhu 2009). Note that most of the diversity-oriented methods in these studies are generally based on a greedy approximation strategy targeting the diversity that mainly relates to content. Nevertheless, as will be investigated in detail, diversity can be viewed as a problem of coverage nature in two ways: one is referred to as content coverage for similar textual/topic elements; the other is referred to as structure coverage for difference distributions.

For instance, given 1,000 online reviews of a product on an e-shopper, there are 600 positive reviews, 300 negative reviews, and 100 neutral reviews. If a user only browses 10 reviews, the problem is how to extract the appropriate subset that covers not only the sufficient content information of the 1,000 reviews, but also the structural information (e.g., similar distribution on different opinions/sentiments). Here, if 10 positive reviews were extracted, such a subset might make sense, since the majority of the content would be covered, which, notably, could often be observed in practice; but this does not sound intuitively appealing to many users. Instead, it is preferable to find a subset composed of six positive, three negative, and one neutral review, which as a whole seems to not only cover the sufficient content but deliver similar opinion distribution. However, though this viewpoint of diversity in content coverage and structure coverage is considered meaningful and important, little literature has addressed related issues with effective and efficient methods.

Meanwhile, clustering-based methods are somewhat effective for extracting a small and diverse set (Han and Pei 2011, Liu 2011). Clustering is an unsupervised grouping of a set of objects into clusters such that the objects within each cluster are similar to each other and the objects in different clusters are dissimilar to each other (Aliguliyev 2009a, b; Carpineto et al. 2009a, b; Grabmeier and Rudolph 2002). Thus, clustering on the original search result set and then extracting centroid for each cluster could form a diverse result set, because each centroid could cover the maximum information of its corresponding cluster. However, the clustering-based methods have not explicitly taken structure coverage into account, and the sizes of different clusters are somewhat ignored in extracting centroids, which may lead to unsatisfactory distributions.

In our previous work (Ma and Wei 2012), we proposed a diversity-based metric, called information coverage, to evaluate the diversity of information in search results. The metric possesses useful properties, high comprehensibility, and effectiveness. Information coverage is composed of two perspectives, i.e., content coverage and structure coverage, in which the former reflects the overall information content (e.g., literal content, features, opinions, sentiments, etc.) coverage of the result set with respect to the original set, and the latter measures the information structure closeness/consistency between the two sets. Thus, optimizing the total information coverage could also be regarded as a solution to pursuing search result diversification.

Based on these discussions, we posit that the strategy to optimize the total information coverage metric should lead to obtaining a small set with desirable information diversity with respect to the original set. Specifically, three research questions are to be discussed in this paper:

RQ-1: How should an extraction method be designed to obtain a diverse result set considering and information coverage metric?

RQ-2: How are the effectiveness, efficiency, and parameter robustness of the proposed method?

RQ-3: How does the proposed method perform compared with other methods in light of diversity?

The remainder of this paper is organized as follows. Section 2 discusses related work on diversity-oriented evaluation metrics and extraction methods, followed by Section 3, which introduces our proposed diversity-oriented extraction method. In Section 4, evaluation experiments on effectiveness, efficiency, and parameter robustness are discussed. To further justify the outperformance of the proposed method, Section 5 shows comparative experiments with other methods. Finally, Section 6 concludes the work and highlights future research.

2. Related Work

This section will review the work on coverage-based evaluation metrics and diversity-oriented extraction methods.

2.1. Related Metrics

Several evaluation metrics about coverage exist. Pan et al. (2005) proposed a "coverage" metric to evaluate the information coverage of the result set with respect to the original set given predefined class labels, i.e., $C(R)/|C|$, where $C(R)$ represents the number of unique class labels in the result set and $|C|$ is the total number of class labels predefined in the original set. Zhai et al. (2003) designed a *S-recall@K* metric to evaluate

the information content coverage under the context of predefined subtopic retrieval, which calculates the percentage of the number of subtopics in all documents with respect to the number of the subtopics, which could be regarded as generally similar to the previous one. In addition, Zhuang et al. (2008) put forward two "representative metrics" to reflect the information of extracted blog sets with respect to the original blog sets in the context of blog extraction. Nevertheless, the above-mentioned metrics do not consider information structure.

In recent years, several integrated diversity metrics have also been proposed with the development of diversity-oriented methods. For example, *α-nDCG@k* was proposed by Clarke et al. (2008) to test the novelty and diversity in information retrieval, which adopted the idea of *nDCG* (Järvelin and Kekäläinen 2002) and considered the redundant probabilities of documents to be extracted with respect to those already chosen. Agrawal et al. (2009) suggested using the intent aware metric—*nDCG-IA@k*—to take the importance distribution of query aspects into account based on *nDCG*. *Precision-IA@k* has been used to perform diversity evaluation in the TREC2009 Conference (Clarke et al. 2010), which considered multitopic and multi-subtopic problems. Apart from the lack of intuitiveness, they also show limitations in evaluating the diversity of an unranked list of documents, which, however, is regarded necessary when considering the whole picture of the original set. Furthermore, these metrics do not consider information structure.

In Ma and Wei (2012), a metric to evaluate diversity was proposed from a combined perspective of information content coverage and information structure coverage. Given an original set D of n documents and an extracted small set D' with size= k to reflect the information content coverage of D' with respect to D—i.e., $Cov_C(D', D)$—the mathematical average operation is used to aggregate the content coverage degree of all documents in D, which is illustrated in Equation (1). Clearly, $Cov_C(D', D)$ captures the content coverage from the perspective of similarity, which is rooted in a general observation that if document a is similar to document b (e.g., literally, on opinions, or on topics, etc.), then a could be deemed to cover the information of b to some extent. Moreover, $Cov_C(D', D)$ possesses some useful properties. First, it is in the range $[0, 1]$ and is reflexive. Second, it is monotonic, i.e., if $D'' \subseteq D'$, then $Cov_C(D'', D) \le Cov_C(D', D)$. Third, $k/n \le Cov_C(D', D) \le 1$.

In the meantime, information structure coverage, i.e., $Cov_S(D', D)$, is modeled with information entropy as a part of the total information coverage metric, which can appropriately measure the information distribution in D' with respect to that in D. Each document, i.e., d'_j, $j = 1, 2, \ldots, k$, in D' could be treated as an implicit class label, resulting in k natural classes. Then each d in D could be assigned into the class with a label m, where $m = \arg\max_{j=1,2,\ldots,k}(sim(d'_j, d))$ with $m = 1, 2, \ldots, k$. Thus, the n documents in D could be assigned into k classes, denoted as $D_1, D_2, \ldots, D_k$, respectively. If document d is assigned to D_j, then d belongs to D_j with $sim(d'_j, d)$, reflecting the extent of information of d covered/loaded in D_j, where d'_j is the natural label of D_j. Thus, the total "information load" in D_j, i.e., the cardinality with respect to the pieces of information covered/loaded in D_j to reflect the information in the original set D, is $\sum_{d \in D_j} sim(d'_j, d)$, denoted as n_j^v, and the total information load in D is $\sum_j n_j^v$, denoted as n^v (Dubois and Prade 1985, Herrera and Martínez 2000, Ralescu 1995). Thereafter, $Cov_S(D', D)$ can be calculated in the spirit of information entropy, as illustrated in Equation (2). Moreover, $Cov_S(D', D)$ also exhibits some useful properties. First, it is in the range $(0, 1]$ and is reflexive. Second, if the information load in D could be conveyed with the equivalent distribution into D', then D' preserves the best information structure. Third, if the information load in D could be assigned in a manner of more proximate distribution into D', $Cov_S(D', D)$ would be higher, which is important for designing better strategies for extracting a subset with higher coverage. Meanwhile, the proposed metrics can be interpreted intuitively and also possess better properties than existing metrics. For the above reasons, a combined information coverage metric, i.e., $Cov(D', D)$ or Cov in short, as shown in Equation (3), will be used as the key metric in the paper.

$$Cov_C(D', D) = \frac{\sum_{d \in D} \max_{d' \in D'}\{sim(d', d)\}}{|D|} \tag{1}$$

$$Cov_S(D', D) = \begin{cases} 1 & \text{if } k = 1 \\ -\dfrac{1}{\log_2 k} \displaystyle\sum_{j=1}^{k} \frac{n_j^v}{n^v} \cdot \log_2\left(\frac{n_j^v}{n^v}\right) & \\ & \text{if } k > 1 \end{cases} \tag{2}$$

$$Cov(D', D) = \begin{cases} Cov_C(D', D) = \dfrac{1}{n} \displaystyle\sum_{d \in D} sim(d'_1, d) & \\ \quad \text{if } k = 1 & \\ Cov_C(D', D) \times Cov_S(D', D) & \\ \quad = \dfrac{1}{n} \displaystyle\sum_{d \in D} \max_{d' \in D'}\{sim(d', d)\} & \\ \quad \times \left\{ -\dfrac{1}{\log_2 k} \displaystyle\sum_{j=1}^{k} \frac{n_j^v}{n^v} \cdot \log_2\left(\frac{n_j^v}{n^v}\right) \right\} & \\ \quad \text{if } k > 1. \end{cases} \tag{3}$$

It should be noted that the above metrics are based on a pairwise similarity metric (i.e., $sim(d', d)$). For different contexts, different similarity metrics should be

carefully selected. For instance, considering only literal content for clustering, the Cosine similarity metric is a popular option (Liu 2011). If feature information is to be incorporated, Euclidean distance or Kullback-Leibner divergence-type metrics will be more suitable (Baeza-Yates and Ribeiro-Neto 1999, Liu 2011, Manning et al. 2008). Furthermore, when sentiments or topics are considered in measuring the similarities of online reviews, some sentiment analysis methods with global topic modeling (Li et al. 2010) could be integrated. Therefore, the similarity metric does play a significant role and should be carefully selected before calculating the information coverage values.

2.2. Extraction Methods

For diversity-oriented extraction, there are two main streams: search result diversification (SRD) methods and clustering-based methods. SRD methods can be categorized as either implicit or explicit, depending on how they account for the different aspects underlying a query (Santos et al. 2010).

Implicit SRD methods assume that similar documents cover similar aspects of a query and should be denoted in the final results. Among the implicit SRD methods, the maximal marginal relevance (MMR) strategy proposed by Carbonell and Goldstein (1998) is the most typical. The general idea of MMR is to trade off document similarity with respect to the query and to document dissimilarity with respect to the already selected documents. Subsequent implementations of this idea include the method of Zhai and Lafferty (2006) to model relevance and novelty within a risk-minimization framework by using highly divergent language models. Wang and Zhu (2009) used the portfolio theory in finance to diversify document ranking: two documents were compared based on the correlation of their relevance scores. There are also researchers who targeted this problem using greedy algorithms from the perspective of recommendation diversity. For instance, Qin and Zhu (2013) combined entropy regularizer possessing good properties of monotonicity and submodularity with a modular rating set function to capture the notion of diversity. In addition, Krishnan and Goldberg (2015) proposed a minimum conductance dissimilarity cut (MCDC) algorithm by solving a graph-partition problem on a weighted dissimilarity graph.

In contrast, explicit search result diversification methods explicitly model the aspects underlying a query. For instance, Agrawal et al. (2009) employed a classification taxonomy over queries and documents to represent query aspects (called IA-Select). The method iteratively promotes documents that share a high number of classes with a query, while demoting those with classes already well represented in the ranking. Similarly, Carterette and Chandar (2009) proposed a probabilistic method (i.e., FM-LDA) to maximize the coverage of the retrieved documents with respect to the aspects of a query; they did this by modeling these aspects as topics identified from the top-ranked documents using latent Dirichlet allocation (LDA) (Blei et al. 2003). Recently, Santos et al. (2011) introduced the xQuAD probabilistic framework for search result diversification, which explicitly represents different query aspects as "subqueries." They defined a diversification objective based on the estimated relevance of documents to multiple subqueries, as well as on the relative importance of each subquery in light of the initial query.

As discussed in the Introduction, clustering-based methods possess the relatively consistent objectives with diversity. The popular and commonly used clustering methods, such as K-means (MacQueen 1967, Zhao and Karypis 2004) and agglomerative hierarchical clustering (i.e., AHC) (Fung et al. 2003, Malik et al. 2010), with cluster-centroid extraction, could provide diverse subsets to some extent. Moreover, unlike the above clustering methods, He et al. (2011) proposed a framework to extract diverse subsets (called RR) based on query-specific clustering as well as cluster ranking. In their method, a latent Dirichlet allocation model was used for clustering the query aspects and documents, in which clusters were ranked based on their relevance probabilities with respect to query aspects.

It is worth noting that, though many efforts have been made in diversity extraction from different angles, existing methods take little aspect of structure coverage into consideration, while structure coverage is regarded as meaningful and important for web users.

3. A Heuristic Method for Extracting Diverse Subset

In this section, a heuristic method is introduced to answer RQ-1, based on the idea of optimizing the *Cov* metric of the extracted subsets. Using the previous example of 1,000 online reviews, denoted as D, suppose only sentiment polarity of content is to be considered and there are three extracted sets, each with 10 reviews, i.e., A = {10 positive}, B ={4 positive, 4 negative, 2 neutral}, and C = {6 positive, 3 negative, 1 neutral}. Intuitively, B and C are likely regarded preferable to A since each covers richer content (i.e., higher content coverage) than A in consideration of different polarities of reviews. Next, though both B and C each cover three sentimental polarities, C is usually regarded as preferable to B in light of conformation to the polarity distribution of the original set. Based on Equations (1)–(3), it could be calculated that $Cov_C(A,D) = 60\%$, $Cov_C(B,D) = 100\%$, $Cov_C(C,D) = 100\%$, $Cov_S(A,D) = 100\%$, $Cov_S(B,D) = 96.2\%$, and $Cov_S(C,D) = 100\%$; thus $Cov(A,D) = 60\%$, $Cov(B,D) = 96.2\%$ and $Cov(C,D) = 100\%$, which intuitively reflects the rationality of the metric. Thereafter,

given an original set of documents, the objective is to find a subset by maximizing *Cov*.

Driven by the abovementioned idea, a two-step optimization strategy could be devised: (1) content coverage maximization along with an algorithm called Cov_C-*Select*, and (2) total information coverage (i.e., both content and structure) maximization along with an algorithm called Cov_{C+S}-*Select*, which will be discussed in Sections 3.1 and 3.2. Moreover, to further optimize the efficiency of the computation, a heuristic method, called *FastCov*$_{C+S}$-*Select*, incorporating a fast approximation strategy on step 2, is introduced in Section 3.3.

3.1. Content Coverage Maximization

The content coverage maximization problem could be formulated as follows:

Definition 1 (*maxContCov*(k))**.** Given an original set $D = \{d_1, d_2, \ldots, d_n\}$, the similarity between any two documents in D (i.e., $sim(d_i, d_j)$) and an integer k (i.e., $1 < k < n$), the content coverage maximization problem (i.e., *maxContCov*(k)) is to find a subset of documents $D'_C \subseteq D$ with $|D'_C| = k$ such that

$$Cov_C(D'_C, D) = \max_{D' \subseteq D, |D'|=k} \{Cov_C(D', D)\} = \max_{D' \subseteq D, |D'|=k} \left\{ \frac{\sum_{d \in D} \max_{d' \in D'} \{sim(d', d)\}}{n} \right\}. \quad (4)$$

Notably, but not surprisingly, by mapping it into a classical NP-hard problem named Max Coverage (Hochba 1997), it could be observed that the desired objective of *maxContCov*(k) is also an NP-hard problem. Fortunately, not all is lost. The objective function *maxContCov*(k) possesses a desirable property—submodularity (Nemhauser et al. 1978)—that allows a greedy strategy to be used that solves the problem quite well.

Submodularity. *Given a finite ground set N, a set function $f: 2^N \mapsto \mathbb{R}$ is submodular if and only if for all sets $S, T \subseteq N$, such that $S \subseteq T$, and $d \in N \backslash T$, $f(S + d) - f(S) \geq f(T + d) - f(T)$.*

Intuitively, a submodular function satisfies the economic principle of diminishing marginal returns; i.e., the marginal benefit of adding a document to a larger collection is less than that of adding it to a smaller collection. It can be proved that the function of information content coverage is a submodular function (Proposition 1); the proof is given in the online appendix.

Proposition 1. *The function of information content coverage $Cov_C(D', D)$ is submodular.*

Nemhauser et al. (1978) has pointed out that even adopting the greedy strategy to optimize the submodular function could lead to bounded error. Specifically, for a submodular set function f, let S^* be the optimal set of k elements that maximizes f and let S' be the k-element set constructed by greedily selecting one element at a time that gives the largest marginal increase to f, then $f(S') \geq (1 - 1/e) f(S^*)$.

Figure 1. Cov_C-*Select* Algorithm

Algorithm 1: Cov_C-*Select*
Input: $D = \{d_1, d_2, \ldots, d_n\}$, k, $sim(d_i, d_j)$
Output: set of k documents D'
1. $D' = \emptyset$, $D_{Candidate} = D$;
2. **while** $|D'| < k$ **do**
3. **for** $d \in D_{Candidate}$ **do**
4. $Cov_C(D' \cup d, D) \leftarrow \sum_{d_i \in D} \max_{d' \in D' \cup d} \{sim(d', d_i)\}$;
5. **end for**
6. $d^* \leftarrow \arg\max_d \{Cov_C(D' \cup d, D)\}$;
7. $D' \leftarrow D' \cup d^*$;
8. $D_{Candidate} \leftarrow D_{Candidate} \backslash \{d^*\}$;
9. **end while**
10. **return** D'

Based on the above property and Proposition 1, a naïve greedy algorithm called Cov_C-*Select* can be proposed for computing a solution to *maxContCov*(k) as shown in Figure 1. In the algorithm, given the original set $D = \{d_1, d_2, \ldots, d_n\}$, let $D' = \emptyset$ and D initially be the candidate set $D_{Candidate}$ (line 1). Each extraction step derives a result d^* in $D_{Candidate}$ into set D', which makes the current value of $Cov_C(D' \cup d^*, D)$ maximum and thus contributes the largest marginal value of information content coverage (lines 4–7). Since the objective *maxContCov*(k) is submodular and the algorithm Cov_C-*Select* is in a greedy manner, Proposition 2 can be derived.

Proposition 2. *The Cov_C-Select algorithm is a $(1 - 1/e)$-approximation algorithm for maxContCov(k).*

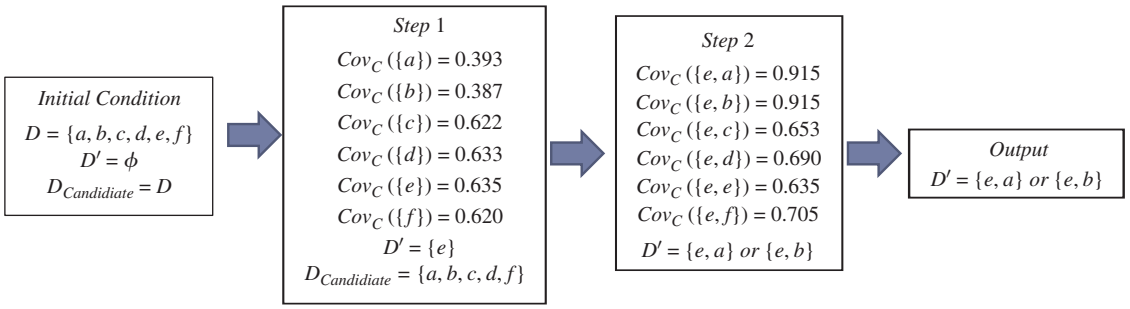
Moreover, given n as the size of original set and k as the size of the result set, it could be easily inferred that the computation complexity of the Cov_C-*Select* algorithm is $O(k^2 n^2)$. To further clarify the extraction procedure of Cov_C-*Select*, see Example 1.

Example 1. Given a set D with six documents, i.e., $D = \{a, b, c, d, e, f\}$ and the similarity matrix of the six documents in Figure 2, then the extraction procedure to retrieve a small subset of two diverse documents by using Cov_C-*Select* is as shown in Figure 3. It could be found that the extracted results are the same as those derived with exhaustive strategy.

Figure 2. Similarity Matrix of Documents in Example 1

	a	b	c	d	e	f
a	1	0.95	0.03	0.05	0.12	0.21
b	0.95	1	0.13	0.08	0.15	0.01
c	0.03	0.13	1	0.87	0.92	0.78
d	0.05	0.08	0.87	1	0.85	0.95
e	0.12	0.15	0.92	0.85	1	0.77
f	0.21	0.01	0.78	0.95	0.77	1

Figure 3. (Color online) The Extraction Procedure Using Cov_C-*Select*



3.2. Total Information Coverage Maximization

In this section, we aim to optimize the total information coverage on the basis of content coverage maximization. First, Definition 2 describes the total information coverage maximization problem (i.e., *maxCov*(k)) based on the definition of information coverage metric.

Definition 2 (*maxCov*(k))**.** Given an original set $D = \{d_1, d_2, \ldots, d_n\}$, the similarity between any two documents in D (i.e., $sim(d_i, d_j)$), and an integer k ($1 < k < n$), the total information coverage maximization problem (i.e., *maxCov*(k)) is to find a subset of documents $D'_* \subseteq D$ with $|D'_*| = k$ such that

$$\begin{aligned} Cov(D'_*, D) &= \max_{D' \subseteq D, |D'|=k} \{Cov_C(D', D) \times Cov_S(D', D)\} \\ &= \max_{D' \subseteq D, |D'|=k} \left\{ \frac{\sum_{d \in D} \max_{d' \in D'} \{sim(d', d)\}}{n} \right. \\ &\quad \left. \times \left\{ -\frac{1}{\log_2 k} \sum_{j=1}^{k} \frac{n_j^v}{n^v} \cdot \log_2 \left(\frac{n_j^v}{n^v} \right) \right\} \right\}. \quad (5) \end{aligned}$$

Similarly, it can be obviously observed that *maxCov*(k) problem is also an NP-hard problem. However, it could be proved that the total information coverage metric $Cov(D', D)$ is not a submodular function (Proposition 3), which is illustrated in the online appendix.

Proposition 3. *The function of information coverage metric $Cov(D', D)$ is not submodular.*

Working from Proposition 3, it is hard to apply the simple greedy strategy like Cov_C-*Select* to guarantee the *Cov* value of the resultant subset to be or be very close to global optimum. Instead, the idea of simulated annealing is adopted, which is complex but effective.

Simulated annealing is a compact and robust technique that provides excellent solutions to single and multiple objective optimization problems with a substantial reduction in computation time (Suman and Kumar 2006). It is a kind of stochastic search algorithm based on Monte-Carlo iterations (Metropolis et al. 1953), which is inspired by heating and controlled cooling of a material, and then is independently described and applied to solve combinatorial optimization problems (Kirkpatrick et al. 1983, Černú 1985). This technique normally begins from a very high initial cooling temperature and tries to "climb hill" to avoid local convergence and reach global optimization by making use of stochastic search strategy as the temperature drops. Simulated annealing has been proven to possess the characteristic of asymptotic optimality (Cormen et al. 2009, Granville et al. 1994) and thus has been widely applied in engineering, including production scheduling, control engineering, machine learning, neural network, signal processing, etc.

In our study, by treating the output of Cov_C-*Select* as the initial solution and applying the stochastic search strategy used in simulated annealing, we propose a heuristic algorithm named Cov_{C+S}-*Select* to solve the *maxCov*(k) problem, whose pseudocode is illustrated in Figure 4. In Cov_{C+S}-*Select*, unlike traditional simulated annealing method, a memory state variable is introduced to avoid the problem of missing the best solution at the time of certain iteration due to the implementation step of acceptance probabilities.

Given the original set D of size n, the extraction size k, the initial solution (state) D_0 from the output of Cov_C-*Select*, the similarity value between any two documents $sim(d_i, d_j)$ as well as the initial cooling temperature T_0 and final temperature $T_{\min}$, Cov_{C+S}-*Select* tries to extract a set containing k documents with maximum value of information coverage. In the initialization stage, the memorial variable $D_{\max}$ is introduced to record the best solution with the highest value of information coverage at the time after each iteration (line 1). The iteration does not terminate until the current temperature T drops to the minimum temperature $T_{\min}$. In the algorithm of Cov_{C+S}-*Select* as shown in Figure 4, each iteration procedure is composed of four steps as follows.

(1) *New solution generation from current state* (*lines* 3–16). The objective of Cov_{C+S}-*Select* lies in improving *Cov* by optimizing structure coverage on the basis of the subset with high content coverage. According to the properties of information structure coverage, if the information load in D could be assigned into D' with a more proximate distribution, implying that the cumulative similarities in each subset are more similar to each other, the value of $Cov_S(D', D)$ would be

Figure 4. *Cov_{C+S}-Select* Algorithm

Algorithm 2: *Cov_{C+S}-Select*
Input: D, k, $sim(d_i, d_j)$, T_0, $T_{\min}$
Output: set of k documents D'

1. $D_0 = Cov_C\text{-}select(k)$;
2. $D' = D_0$, $D_{\max} = D_0$, $T = T_0$, $N = 1$;
3. **while** $T > T_{\min}$ **do**
4. $Cov_0 \leftarrow Cov(D', D)$;
5. **for** $d' \in D'$ **do**
6. $n^v_{d'} \leftarrow$ calculateCumulativeSimi(d', D', D);
7. **end for**
8. $n^v_{avg} \leftarrow \sum_{d'} n^v_{d'}/k$;
9. $n^v_{\min} \leftarrow \min\{n^v_{d'}\}$, $\min = \arg\min_{d'}\{n^v_{d'}\}$;
10. **if** $n^v_{\min} = n^v_{avg}$ **do**
11. $\min = Random(1, k)$;
12. **end if**
13. $D_p = D \backslash D'$;
14. **for** $d \in D_p$ **do**
15. $Cov_d \leftarrow Cov(D' \backslash \{d'_{\min}\} + \{d\}, D)$;
16. **end for**
17. $Cov_{\max} \leftarrow \max\{Cov_d\}$, $\max = \arg\max_d\{Cov_d\}$;
18. $\Delta Cov = Cov_{\max} - Cov_0$;
19. **if** $\Delta Cov \geq 0$ **do**
20. $D' \leftarrow D' \backslash \{d'_{\min}\} + \{d_{\max}\}$;
21. **else do**
22. $acceptProb = e^{\Delta Cov/T}$;
23. $RandomProb = Random(0, 1)$;
24. **if** $acceptProb \geq RandomProb$ **do**
25. $D' \leftarrow D' \backslash \{d'_{\min}\} + \{d_{\max}\}$;
26. **end if**
27. **end else**
28. **end if**
29. $T \leftarrow T/\log(1 + N)$;
30. $N \leftarrow N + 1$;
31. **if** $Cov(D') > Cov(D_{\max})$ **do**
32. $D_{\max} \leftarrow D'$;
33. **end if**
34. **end while**
35. $D' \leftarrow D_{\max}$;
36. **return** D'

higher (Ma and Wei 2012). Hence, optimizing information structure coverage would be first finding the subset with the lowest value of cumulative similarity, i.e., $n^v_{\min}$ (line 8). If there exist multiple documents with identical cumulative similarity, randomly choose one for substitution operation (line 10). Next, calculate all *Cov* values by substituting the document with minimal value for those in potential set D_p (lines 12–15) and choose the document with the largest *Cov* value after substitution as a potential new solution (line 16).

(2) *Calculating the difference of Cov values between new solution and current status* (*line* 17). This step is straightforward but crucial, which involves calculating the value gain of *Cov* by using the new potential solution to substitute the current state (i.e., ΔCov). The value gain could be used to not only determine whether the new potential solution could be adopted directly, but also to play a decisive role in generating acceptance probability in step (3).

(3) *Judgment on whether to accept the new solution* (*lines* 18–27). This part is the core step incorporating the stochastic search strategy. In this step, if ΔCov calculated in step (2) is positive, meaning that the new potential solution would improve the total information coverage, then directly accept the new solution (lines 18–19). Otherwise, unlike the simple greedy strategy, the unimproved new solution would be adopted with certain "acceptance probability" (i.e., *acceptProb*) rather than being rejected (lines 20–26). Based on the idea of simulated annealing, the acceptance probability would be determined by both ΔCov and the current temperature T (line 21). Since the temperature drops quickly with the increase of the iteration number (line 28), the *acceptProb* tends to be smaller even if ΔCov remains invariant, implying that it is increasingly difficult to accept an unimproved solution with the iterations. This phenomenon is called "annealing procedure," which has been proven to converge to a global optimal solution with a probability of 100% (Cormen et al. 2009, Granville et al. 1994). It is worth noting that, even the value gain of information coverage ΔCov is 0; this algorithm would still prefer new solutions, which is also aimed at encouraging more search and avoiding falling into local optimum.

(4) *Updating relevant variables* (*lines* 28–32). After determining a new solution for current iteration, several iteration-related variables—i.e., current temperature (line 28), iteration number (line 29), the best solution so far (lines 30–32)—need to be updated accordingly.

At the end, the best solution with the highest *Cov* value is obtained as the final output, rather than the result of the last iteration (lines 34–35), which avoids the problem of missing the best solution at the time of certain iteration because of the implementation step of acceptance probabilities.

It can be seen that *Cov_{C+S}-Select* is an algorithm of typical simulated annealing nature. The convergence rate and error bounds have been comprehensively discussed with a definite conclusion in literature. Concretely, *Cov_{C+S}-Select* implemented the Metropolis criterion to generate the new potential solution, guaranteeing that the unimproved new solution would be adopted with certain probability. Because the states in the solution space with respect to the problem defined in Equation (5) are finite, the probability that the simulated annealing algorithm *Cov_{C+S}-Select* terminates with a global optimal solution approaches 100% (Cormen et al. 2009, Granville et al. 1994). Furthermore, *Cov_{C+S}-Select* possesses the ability of asymptotic convergence that has been proved by prior theoretical research (Cormen et al. 2009, Granville et al. 1994). In addition, to further quantitatively analyze *Cov_{C+S}-Select*, effectiveness experiments on the optimization performance of *Cov_{C+S}-Select* will be discussed in Section 4.3.

As previously mentioned, the computational complexity of *Cov_C-Select* is $O(k^2n^2)$. For *Cov_{C+S}-Select*, its magnitudes of outer loop (controlled by cooling temperature) is $O(T_0)$, and the maximum magnitude of

the inner loop is $O(kn^2)$. Hence, considering that the initial solution D_0 is usually the output of Cov_C-*Select*, the total computation complexity for Cov_{C+S}-*Select* is $O(k^2n^2) + O(T_0kn^2)$, in which T_0kn^2 is the main influence factor, since $k \ll n$ and $k \ll T_0$ generally. Although theoretically the final solutions of simulated annealing have nothing to do with the initial inputs (Granville et al. 1994, Kirkpatrick et al. 1983), in real implementation we still suggest adopting the output of Cov_C-*Select* as its initial input D_0. First, compared with Cov_{C+S}-*Select*, the computational complexity of Cov_C-*Select* is relatively small, which would hardly affect the total efficiency dramatically. Second, by combining the two algorithms, the number of iterations could be reduced as much as possible and thus the computational complexity degrades.

3.3. A Fast Approximate Heuristic Method—*FastCov*$_{C+S}$-*Select*

In the algorithm of Cov_{C+S}-*Select*, to find the solution as close as possible to global optimum, the initial cooling temperature T_0 would be set at a rather large value. Meanwhile, if the size of the original set is very large, the running speed would be significantly slowed, which is also a major concern for simulated annealing (Suman and Kumar 2006). To alleviate this problem, based on Cov_{C+S}-*Select*, hereby a fast approximate heuristic method, called *FastCov*$_{C+S}$-*Select*, is proposed to find the satisfactory result sets in a relatively short time utilizing the properties of *Cov*. The algorithm of the *FastCov*$_{C+S}$-*Select* is as shown in Figure 5.

The core idea of *FastCov*$_{C+S}$-*Select* lies in that it is more likely to obtain the extracted set with higher total information coverage in optimizing the structure coverage on the basis of a subset with significantly high content coverage. Proposition 2 guarantees that the output of Cov_C-*Select* would possess sufficiently high total information coverage. Thus, it is considered meaningful to conduct the iteration optimization in a potential set that Cov_C-*Select* outputs, where the size of the output is set relatively larger than the final extraction size k, rather than conducting the whole original set. Though the *FastCov*$_{C+S}$-*Select* method cannot assure theoretically that the extracted set is globally optimum or asymptotically optimal on total information coverage, the experimental results show desirable extraction effectiveness and excellent performances on efficiency. In essence, compared with Cov_{C+S}-*Select*, *FastCov*$_{C+S}$-*Select* only differentiates on initial input D_0 and potential search set D_p. Specifically, the initial input D_0 in *FastCov*$_{C+S}$-*Select* is a subset of $t \times k$ documents that Cov_C-*Select* extracts rather than of k documents, in which t is a small integer, such as 3, 5, 10, etc. In *FastCov*$_{C+S}$-*Select*, the initial solution is the top k documents of D_0 (line 1) and the potential search set is the remaining $(t-1)\times k$ documents in D_0 (line 13). Clearly, the computational complexity of *FastCov*$_{C+S}$-*Select* is

Figure 5. *FastCov*$_{C+S}$-*Select* Algorithm

Algorithm 3: *FastCov*$_{C+S}$-*Select*
Input: D, k, t, $sim(d_i, d_j)$, T_0, $T_{\min}$
Output: set of k documents D'

1. $D_0 = Cov_C\text{-}select(t \times k)$;
2. $D' = selectSubList(D_0, k)$, $D_{\max} = D'$, $T = T_0$, $N = 1$;
3. **while** $T > T_{\min}$ **do**
4. $Cov_0 \leftarrow Cov(D', D)$;
5. **for** $d' \in D'$ **do**
6. $n^v_{d'} \leftarrow$ calculateCumulativeSimi(d', D', D);
7. **end for**
8. $n^v_{avg} \leftarrow \sum_{d'} n^v_{d'}/k$;
9. $n^v_{\min} \leftarrow \min\{n^v_{d'}\}$, $\min = \arg\min_{d'}\{n^v_{d'}\}$;
10. **if** $n^v_{\min} = n^v_{avg}$ **do**
11. $\min = Random(1, k)$;
12. **end if**
13. $D_p = D_0 \backslash D'$;
14. **for** $d \in D_p$ **do**
15. $Cov_d \leftarrow Cov(D'\backslash\{d'_{\min}\} + \{d\}, D)$;
16. **end for**
17. $Cov_{\max} \leftarrow \max\{Cov_d\}$, $\max = \arg\max_d\{Cov_d\}$;
18. $\Delta Cov = Cov_{\max} - Cov_0$;
19. **if** $\Delta Cov \geq 0$ **do**
20. $D' \leftarrow D'\backslash\{d'_{\min}\} + \{d_{\max}\}$;
21. **else do**
22. $acceptProb = e^{\Delta Cov/T}$;
23. $RandomProb = Random(0, 1)$;
24. **if** $acceptProb \geq RandomProb$ **do**
25. $D' \leftarrow D'\backslash\{d'_{\min}\} + \{d_{\max}\}$;
26. **end if**
27. **end else**
28. **end if**
29. $T \leftarrow T/\log(1 + N)$;
30. $N \leftarrow N + 1$;
31. **if** $Cov(D') > Cov(D_{\max})$ **do**
32. $D_{\max} \leftarrow D'$;
33. **end if**
34. **end while**
35. $D' \leftarrow D_{\max}$;
36. **return** D'

$O(tk^2n^2) + O(T_0tk^2n)$, in which $O(T_0tk^2n)$ is the main influence factor. Since in general $t \times k \ll n$, *FastCov*$_{C+S}$-*Select* runs much faster than Cov_{C+S}-*Select* theoretically and shows faster convergence rate.

4. Evaluation Experiments on *FastCov*$_{C+S}$-*Select*

To answer the research question RQ-2—i.e., the effectiveness, efficiency and parameter robustness of the proposed *FastCov*$_{C+S}$-*Select*—a series of evaluation experiments was conducted to demonstrate the effectiveness, efficiency, and parameter robustness.

4.1. Experimental Descriptions and Setup

The setup and configurations of the three groups of experiments on, i.e., effectiveness, efficiency, and parameter robustness, are introduced. First, the effectiveness experiments aimed at testing the performance differences between the outputs generated by the proposed algorithms (i.e., Cov_C-*Select*, Cov_{C+S}-*Select*, and

FastCov$_{C+S}$-Select), as well as the globally optimal solutions, in which the brute-force enumeration strategy was used. It is worth mentioning that the size of the original data set could hardly be large when involving the enumeration strategy because of its computational inefficiency. Second, to make assessment of the extraction efficiency of the *Cov$_C$-Select*, *Cov$_{C+S}$-Select*, and *FastCov$_{C+S}$-Select* algorithms, several efficiency experiments were conducted with different sizes of data sets. Third, since the proposed *Cov$_{C+S}$-Select* and *FastCov$_{C+S}$-Select* algorithms incorporate the idea of simulated annealing with several controllable parameters, a series of experiments were conducted as well to test the robustness of the algorithms with respect to the parameters. Concretely, the parameter robustness experiments were conducted on the four key parameters: (1) initial cooling temperature T_0; (2) temperature cooling function; (3) initial input size ($t \times k$); and (4) initial input set.

4.2. Data Set and Evaluation Metrics

To comprehensively evaluate the performance of the proposed algorithms, the data set comprises Google search snippet results of 3,500 queries raised in KDD Cup 2005,[1] which is widely used in related studies. Specifically, the 3,500 queries were chosen as the search keywords in Google. Then all the snippets of each query were crawled from Google using Apache Lucene, HTML parser, and HTTP client packages and APIs, in which the number of snippets for each query is roughly 1,000. In total, the evaluation experiments were conducted on the data set of 3,500,000 (= 3,500 × 1,000) snippets. Moreover, since the data sets are mainly used for literal text mining, without explicitly specifying the similarity or distance metrics, the popular Cosine similarity metric is adopted for measuring the similarity in the following experiments.

In the experiments, the $Cov(D', D)$ in Equation (5) (*Cov* for short) was used to evaluate the information coverage of extracted subsets. Hence the average value of *Cov* of all the 3,500 Google queries, denoted as *avgCov*, was used as the evaluation metric.

4.3. Experimental Results

(1) *Effectiveness experiments.* In the effectiveness experiments, for each query, k results were extracted from original sets with the size of 50, where $k = 2, 3, 4$, and 5, respectively, due to the inefficiency of the enumeration method that was used to obtain the actual optimal solution. The results are shown in Table 1, listing the *avgCov* values as well as the average gaps (%) of *Cov* values between each proposed algorithm and the optimal solution (i.e., *avgGap*) over all 3,500 queries. It could be observed that the *avgGap* values of both *Cov$_{C+S}$-Select* and *FastCov$_{C+S}$-Select* were rather small, i.e., <1%, demonstrating the desirable effectiveness in terms of satisfactory error bounds of the proposed algorithms. Moreover, the performance of *FastCov$_{C+S}$-Select* was quite close to *Cov$_{C+S}$-Select*, meaning that *FastCov$_{C+S}$-Select* also found a desirable solution in a much faster time.

To further show the desirable effectiveness of *FastCov$_{C+S}$-Select* on large-scale data, we extracted k (k = 10, 20, 30, respectively) results from the whole original sets, with each having around 1,000 snippets over all the 3,500 queries. The results are illustrated in Table 2, which lists the *avgCov* values as well as *avgGap* values between *FastCov$_{C+S}$-Select* and *Cov$_{C+S}$-Select*. It is shown that the *avgGap* values were rather small (less than 1%) and stable, implying that on average *FastCov$_{C+S}$-Select* could find a desirable solution.

(2) *Efficiency experiments.* In the efficiency experiments, the size of original sets ranged from 1,000 to 10,000, which is a reasonable setting, according to online search services in practice nowadays, with extraction size k of 10 (i.e., the usual size of results on the first page); thus, the running time of the three proposed algorithms are as shown in Figure 6. Notably, the *Cov$_C$-Select* is quick. *FastCov$_{C+S}$-Select* is much faster than *Cov$_{C+S}$-Select* , and the running time of *FastCov$_{C+S}$-Select* is linear with n value, i.e., consistent with the theoretical analysis in Section 3.3.

(3) *Parameter robustness experiments.* Because *FastCov$_{C+S}$-Select* is more efficient than *Cov$_{C+S}$-Select* in terms of effectiveness, according to the previous experimental results, and the essential procedures of *Cov$_{C+S}$-Select* and *FastCov$_{C+S}$-Select* are the same, the parameter robustness experiments were conducted only on *FastCov$_{C+S}$-Select*.

- *On the initial cooling temperature T_0.* Assume that the initial cooling temperature T_0 is c times of original set size n (i.e., $T_0 = c \times n$), the results with different T_0

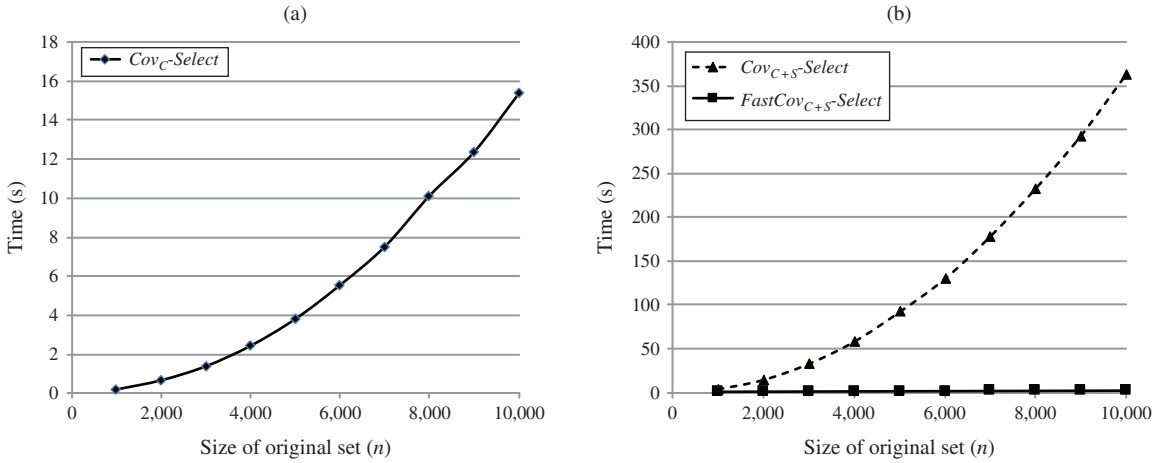
Table 1. *avgCov* and *avgGap* Values for Different Algorithms ($n = 50$)

	$k = 2$		$k = 3$		$k = 4$		$k = 5$	
Algorithms	*avgCov*	*avgGap* (%)	*avgCov*	*avgGap* (%)	*avgCov*	*avgGap* (%)	*avgCov*	*avgGap* (%)
Optimal	0.1591	0	0.2008	0	0.2357	0	0.2664	0
Cov$_C$-Select	0.1565	−1.40	0.1981	−1.28	0.2321	−1.36	0.2623	−1.43
Cov$_{C+S}$-Select	0.1583	−0.51	0.1999	−0.54	0.2338	−0.77	0.2642	−0.75
FastCov$_{C+S}$-Select	0.1581	−0.60	0.1993	−0.75	0.2333	−0.91	0.2636	−0.96

Table 2. *avgCov* and *avgGap* Values for Cov_{C+S}-*Select* and $FastCov_{C+S}$-*Select* (3,500 Queries)

	$k = 10$		$k = 20$		$k = 30$	
Algorithms	*avgCov*	*avgGap* (%)	*avgCov*	*avgGap* (%)	*avgCov*	*avgGap* (%)
Cov_{C+S}-*Select*	0.2599	0	0.3371	0	0.3902	0
$FastCov_{C+S}$-*Select*	0.2566	−0.9364	0.3329	−0.9816	0.3858	−0.9281

Figure 6. (Color online) Experimental Results for Efficiency of Three Proposed Algorithms ($k = 10$)



on *avgCov* of 3,500 queries are listed in Table 3. The results reveal that the initial cooling temperature had no significant influence on result diversity.

- *On the temperature cooling function.* Generally, the temperature cooling function is either linear or logarithmic in the denominator. Theoretically, by using a logarithmic cooling function, the cooling speed is much slower normally, leading to higher possibility of locating the optimum (Cormen et al. 2009). Nevertheless, the results in Table 4 reflect that logarithmic cooling function in $FastCov_{C+S}$-*Select* had no significant influence on results though sacrificing more search time.

- *On the initial input size* ($t \times k$). By changing the parameter t of initial input size ($t \times k$) from 2 to 18, the results in Figure 7 indicate that when t reaches a reasonable value, i.e., around 5, the effectiveness of $FastCov_{C+S}$-*Select* becomes stable, whereas the running time increases quickly as t rises.

- *On the initial input set.* As indicated previously, the Cov_C-*Select* output could be set as the input of $FastCov_{C+S}$-*Select*. The results in Table 5 reveal that, compared with randomly initialized values as input, the Cov_C-*Select* output indeed offered significant awarded marks on the final results for $FastCov_{C+S}$-*Select*.

In sum, we have the findings from the above evaluation experiments. First, compared with the actual optimal results, both Cov_{C+S}-*Select* and $FastCov_{C+S}$-*Select* can achieve satisfactory results. Second, $FastCov_{C+S}$-*Select* is significantly more efficient than Cov_{C+S}-*Select*. Third, compared with a randomly selected initial set, the output of Cov_C-*Select* shows explicit advantage. Finally, parameter robustness experiments reveal that $FastCov_{C+S}$-*Select* is not sensitive to initial temperature or a temperature cooling function, and the initial input size as 5 performs cost-effectively. Therefore, in the following comparative experiments, $FastCov_{C+S}$-*Select* is implemented with the previous parameter configuration.

Table 3. Experimental Results on the Initial Cooling Temperature T_0 ($T_0 = c \times n$)

	avgCov				*avgCov*		
c	$k = 10$*	$k = 20$*	$k = 30$*	c	$k = 10$*	$k = 20$*	$k = 30$*
5	0.2566	0.3329	0.3858	20	0.2566	0.3329	0.3858
10	0.2566	0.3329	0.3858	25	0.2566	0.3329	0.3858
15	0.2566	0.3329	0.3858	30	0.2566	0.3329	0.3858

*No significant differences.

5. Comparative Experiments

Targeting on RQ-3—i.e., whether $FastCov_{C+S}$-*Select* is better than other methods in light of diversity—a series of comparative experiments were conducted to assess performances between $FastCov_{C+S}$-*Select* and other mainstream extraction methods on diversity.

Table 4. Experimental Results on the Temperature Cooling Function

k	*Cooling Function*	*avgCov**	*avgTime* (s)**	*Cooling Function*	*avgCov**	*avgTime* (s)**
10	Logarithmic	0.2566	0.219	Linear	0.2566	0.118
20	Logarithmic	0.3329	0.793	Linear	0.3329	0.429
30	Logarithmic	0.3858	1.362	Linear	0.3858	0.741

*No significant differences. **Significant differences statistically ($p < 0.001$).

Figure 7. (Color online) Experimental Results on the Initial Input Size of *FastCov*$_{C+S}$*-Select* ($k = 10$)

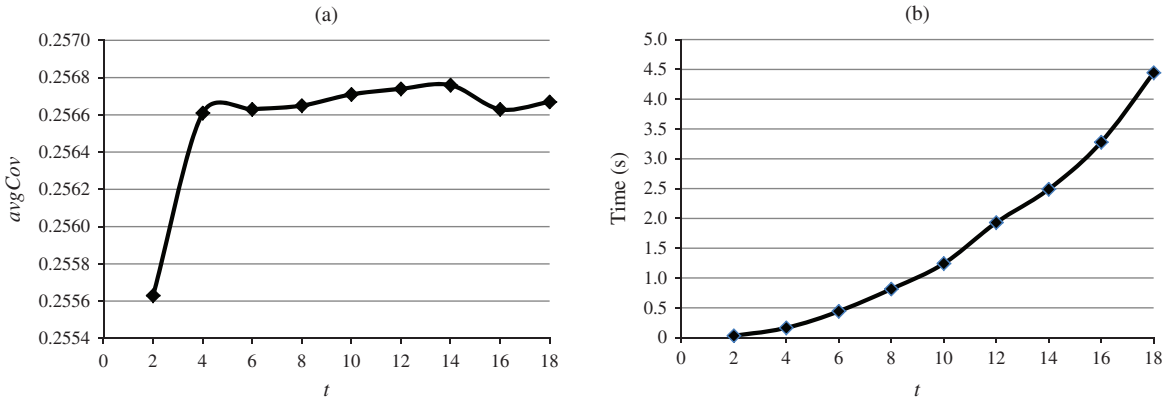


Table 5. Experimental Results on Different Initial Inputs

k	*avgCov** of *Cov*$_C$*-Select* as initial input	*avgCov** of random values as initial input
10	0.2566	0.2284
20	0.3329	0.2894
30	0.3858	0.3247

*Significant differences ($p < 0.001$) for both algorithms.

This section first introduces the benchmark methods used in comparative experiments as well as the parameter settings for these methods. Next, the performances of these methods were compared on different test collections from three distinct perspectives: information coverage perspective, external labelling perspective, and human evaluation perspective, where the latter two could be regarded as two types of external standards.

5.1. Benchmark Methods and Parameter Settings

In the comparative experiments, the *FastCov*$_{C+S}$*-Select*, together with other 11 benchmark diversity-oriented extraction methods, grouped into five categories, were compared (as listed in Table 6). The column "Query based" in Table 6 represents whether the methods directly utilize the relationship between search results and user queries.

In the experiments, similarly to the work in Carterette and Chandar (2009) and He et al. (2011), the LDA model (Blei et al. 2003) was utilized to explicitly model the query aspects or topics from all the original search results, in which the topic number was set to the extraction size k. By the training process, the LDA model could result in a document-topic probability matrix, in which each value represents the probability of a result belonging to a topic. For each new result, it is easy to obtain the document-topic distribution probability through the inference process.

In the naïve category, the baseline method directly extracted the top-k results from the original set without any post-processing techniques. The Random50 method applied a random strategy to extract k results from the original set. To avert possible biases caused by random sampling, the results of the Random50 were the mean values of 50 independent samplings.

For the implicit SRD category, in the MMR method (Carbonell and Goldstein 1998), the popular BM25 weighting (Robertson et al. 1994) was adopted as the relevance function between query and document. The parameter λrepresenting the tradeoff between the above two functions was determined by five-fold cross validation. In the Portfolio method (Wang and Zhu 2009), the expectation and standard deviation of relevance between document and query aspect were estimated based on the document-topic probability matrix generated by LDA modeling. In the MCDC method (Krishnan and Goldberg 2015), the similarities between documents were calculated to estimate the similarities between items.

For the methods in the explicit SRD category, three methods, i.e., IA-Select (Agrawal et al. 2009), FM-LDA (Carterette and Chandar 2009), and xQuAD (Santos et al. 2010), were selected. The core issue in their

Table 6. Benchmark Methods List

Category	ID	Method	Query based	References
Proposed	1	*FastCov$_{C+S}$-Select*	No	This paper
Naïve	2	Baseline	—	—
	3	Random50	No	—
Implicit SRD	4	MMR	Yes	Carbonell and Goldstein (1998)
	5	Portfolio	Yes	Wang and Zhu (2009)
	6	MCDC	No	Krishnan and Goldberg (2015)
Explicit SRD	7	IA-Select	Yes	Agrawal et al. (2009)
	8	FM-LDA	Yes	Carterette and Chandar (2009)
	9	xQuAD	Yes	Santos et al. (2010)
Clustering based	10	*K*-means based	No	Zhao and Karypis (2004)
	11	AHC based	No	Zhao and Karypis (2005)
	12	RR	Yes	He et al. (2011)

models and implementations lies in two folds: the query aspects distribution and document-topic relevance probability. The former was obtained by the LDA inference process, treating the query as a new document and the latter was directly found in a document-topic probability matrix generated by LDA modeling.

For the clustering-based category, the well-known CLUOTO package was used to implement *K*-means-based and AHC-based methods (Zhao and Karypis 2004, 2005). Specifically, the number of clusters was set as the extraction size k. Moreover, in CLUTO the criterion functions for *K*-means and AHC were Direct and UPGMA respectively. As for the RR method, LDA modeling were also used for estimating parameters, which is consistent to that in He et al. (2011).

5.2. Information Coverage on Google Search Results Data Set

The results on average information coverage (i.e., *avgCov*) of extracting 10, 20, and 30 results for 3,500 queries of the 12 methods are listed in Table 7. At the same time, statistical significance tests were conducted on the gap of *avgCov* values of each of the 11 methods and *FastCov$_{C+S}$-Select*. Clearly, *FastCov$_{C+S}$-Select* performed statistically significantly better than all the other 11 extraction methods on information coverage in different extraction sizes.

5.3. External Labelling on Document-Clustering Benchmark Data Sets

To further justify the outperformance of *FastCov$_{C+S}$-Select* over other methods, a series of experiments were conducted on 24 benchmark text data sets with external ground truth topic labels (as listed in Table 8), which are widely used to evaluate the performance of text clustering methods (Whissell and Clarke 2011; Xiong et al. 2009; Zhao and Karypis 2004, 2005; Zhong 2006). Compared with online reviews (e.g., where sentiments should be incorporated into similarity measuring), the benchmark data sets are mainly used for analyzing literal content. Thus, the vector space modeling on keywords as well as the Cosine similarity

Table 7. *avgCov* and Significance of *avgCov* Gap on Google Search Results Data Set (3,500 Queries)

			avgCov (sig. of gap with *FastCov$_{C+S}$-Select*)		
Category	ID	Method name	$k = 10$	$k = 20$	$k = 30$
Proposed	1	*FastCov$_{C+S}$-Select*	0.2566	0.3329	0.3858
Naïve	2	Baseline	0.0963$^{(***)}$	0.1458$^{(***)}$	0.1835$^{(***)}$
	3	Random50	0.1179$^{(***)}$	0.1857$^{(***)}$	0.2356$^{(***)}$
Implicit SRD	4	MMR	0.0799$^{(***)}$	0.1274$^{(***)}$	0.1651$^{(***)}$
	5	Portfolio	0.0976$^{(***)}$	0.1134$^{(***)}$	0.1352$^{(***)}$
	6	MCDC	0.0773$^{(***)}$	0.1006$^{(***)}$	0.1209$^{(***)}$
Explicit SRD	7	IA-Select	0.0783$^{(***)}$	0.1219$^{(***)}$	0.1561$^{(***)}$
	8	FM-LDA	0.0601$^{(***)}$	0.1046$^{(***)}$	0.1570$^{(***)}$
	9	xQuAD	0.0682$^{(***)}$	0.1093$^{(***)}$	0.1544$^{(***)}$
Clustering based	10	*K*-means based	0.1476$^{(***)}$	0.2211$^{(***)}$	0.2763$^{(***)}$
	11	AHC based	0.0585$^{(***)}$	0.0984$^{(***)}$	0.1354$^{(***)}$
	12	RR	0.0698$^{(***)}$	0.0861$^{(***)}$	0.0900$^{(***)}$

$^{***}p < 0.001$.

Table 8. Summary of the 24 Benchmark Document-Clustering Data Sets

Data set	Source	No. of classes	No. of docs	Data set	Source	No. of classes	No. of docs
cacmcisi	SMART	2	4,663	Ohscal	Reuters-21578	10	11,162
classic	SMART	4	7,094	re0	Reuters-21578	13	1,504
cranmed	SMART	2	2,431	re1	TREC	25	1,657
fbis	TREC	17	2,463	Reviews	TREC	5	4,069
hitech	TREC	6	2,301	Sports	TREC	7	8,580
k1a	WebACE	20	2,340	tr11	TREC	9	414
k1b	WebACE	6	2,340	tr12	TREC	8	313
la1	TREC	6	3,204	tr23	TREC	6	204
la12	TREC	6	6,279	tr31	TREC	7	927
la2	TREC	6	3,075	tr41	TREC	10	878
mm	TREC	2	2,521	tr45	WebACE	10	690
new3	20 NewsGroups	44	9,558	Wap	Reuters-21578	20	1,560

measuring is widely adopted to help derive the similarity values between documents, because of the effectiveness (Baeza-Yates and Ribeiro-Neto 1999, Liu 2011, Manning et al. 2008). We used the same similarity measuring process in the experiments.

With the external topic labels, if an extracted set could appropriately cover the information of the original set, it means that the topic label distribution in the extracted set should be close to or consistent with that in the original set. Therefore, targeting the topic label distribution closeness, the well-known Kullback-Leibler (KL) divergence measurement (Kullback and Leibler 1951), which is to evaluate the distance between two distributions, e.g., P and Q, can be used. Clearly, it could be regarded as a kind of external metric for measuring information coverage in this context of data with external labels. Concretely, the Jensen-Shannon (JS) divergence (Endres and Schindelin 2003) metric, which is a symmetric version of KL divergence measurement, was adopted as shown in Equation (6). It indicates that the smaller the *JS* value is, the higher consistency two distributions possess, meaning that the extracted set can better cover the original set.

$$JS(P\|Q) = \frac{1}{2}KL(P\|M) + \frac{1}{2}KL(Q\|M), \quad \text{where } M = \frac{1}{2}(P+Q) \tag{6}$$

For each method, a set of k documents was extracted ($k = 2, 6, 10, 20, 30$), and the *JS* value of the extracted set with respect to the original set was calculated. Since the 24 data sets did not come from the same data population, the Wilcoxon signed-rank test was conducted to examine the *JS* value differences between $FastCov_{C+S}$-*Select* and each of the other methods. The results are listed in Table 9, where the positive Z value that represents $FastCov_{C+S}$-*Select* performs better on distribution consistency than the other method, and the value in the parentheses shows its significance. Obviously, on the external metric of information coverage, i.e., *JS* divergence, the $FastCov_{C+S}$-*Select* method significantly outperformed other methods on different extraction sizes. It should be noted that experiments on more varied extraction sizes have also been conducted, showing

Table 9. Wilcoxon Signed-Rank Test Results for the Hypothesis of $JS(FastCov_{C+S}\text{-}select) < JS(M_i)$

			Z value (sig.)				
Category	ID	Method (M_i)	$k = 2$	$k = 6$	$k = 10$	$k = 20$	$k = 30$
Naïve	2	Baseline	$3.702^{(***)}$	$4.200^{(***)}$	$4.286^{(***)}$	$3.886^{(***)}$	$4.286^{(***)}$
	3	Random50	$2.728^{(**)}$	$2.657^{(**)}$	$4.200^{(***)}$	$4.257^{(***)}$	$4.286^{(***)}$
Implicit SRD	4	MMR	$3.702^{(***)}$	$4.257^{(***)}$	$4.086^{(***)}$	$4.286^{(***)}$	$4.286^{(***)}$
	5	Portfolio	$3.620^{(***)}$	$2.657^{(**)}$	$3.111^{(**)}$	$4.000^{(***)}$	$4.057^{(***)}$
	6	MCDC	$3.702^{(***)}$	$4.257^{(***)}$	$4.286^{(***)}$	$4.257^{(***)}$	$4.286^{(***)}$
Explicit SRD	7	IA-Select	$2.857^{(**)}$	$4.200^{(***)}$	$4.200^{(***)}$	$4.286^{(***)}$	$4.286^{(***)}$
	8	FM-LDA	$2.728^{(**)}$	$4.257^{(***)}$	$4.286^{(***)}$	$4.286^{(***)}$	$4.286^{(***)}$
	9	xQuAD	$3.702^{(***)}$	$4.286^{(***)}$	$4.114^{(***)}$	$4.286^{(***)}$	$4.286^{(***)}$
Clustering based	10	K-means based	$2.539^{(*)}$	$2.403^{(*)}$	$2.400^{(*)}$	$3.072^{(**)}$	$2.342^{(*)}$
	11	AHC based	$3.285^{(**)}$	$4.229^{(***)}$	$3.771^{(***)}$	$4.000^{(***)}$	$3.886^{(***)}$
	12	RR	$2.763^{(**)}$	$4.229^{(***)}$	$3.467^{(***)}$	$3.886^{(***)}$	$3.771^{(***)}$

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

similar results. Due to space limitations, the results are not presented here.

5.4. Human Evaluation on Online Consumer Reviews

To further justify the effectiveness of the *FastCov*$_{C+S}$*-select* method, a human evaluation study was conducted. Human evaluators were recruited to directly assess the information coverage of differently extracted sets with respect to the original sets. Obviously, human assessment could be regarded as another kind of external standard.

Considering the capability of human evaluators and effectiveness of evaluation, the evaluation process was carefully elaborated. First, in addition to *FastCov*$_{C+S}$*-select*, four nonnaïve methods were selected for comparison, where three methods (i.e., Portfolio, FM-LDA, and *K*-means based) were selected from their corresponding categories with good performances, and one method (i.e., MCDC) was from a 2015 effort. Second, from one of the largest online restaurant review platforms (www.dianping.com), four groups of reviews (size = 30 for each group) were crawled with diversified content on well-recognized features (i.e., taste, environment, service, and price) and various sentiment distributions, where the sentiment distributions on (positive:negative) of the four groups were (20%:80%), (40%:60%), (60%:40%), and (80%:20%). Third, the Euclidean distance was used to calculate the similarity based on the features and polarity between any two reviews, as widely used in online review analysis (Shi and Liang 2015, Tsur et al. 2010). Fourth, with each method, five reviews were extracted from each group of 30 reviews, i.e., totally 4×5 extracted review sets. Fifth, for each group of reviews, 27 evaluators were recruited, i.e., totally $4 \times 27 = 108$ evaluators. The evaluation was designed and executed in line with the general practice in design science research (Gregor and Hevner 2013). The evaluators were recruited from two high-prestige universities in China. All the evaluators were experienced with online shopping and familiar with online product reviews. The evaluation work was conducted in a university lab. Each evaluator was asked to assess the information coverage of an extracted set with respect to the corresponding group of 30 original reviews. The evaluators were randomly assigned to one of the extracted review sets and they were not aware of the extraction methods. They were not allowed to discuss the experiment with others during the evaluation work. In the evaluation, the evaluators were first asked to carefully read the complete set of 30 original reviews. Subsequently, they were required to read the five reviews in the extracted set. After that, they reported their judgments about the levels of coverage using a five-point Likert score sheet (i.e., 1 represents zero coverage and 5 represents full coverage). The coverage levels were evaluated on five aspects (overall, taste, environment, service, and price), based on the extent to which the information conveyed in the original set was covered in the extracted set, including sentiments and proportions of positive/negative opinions on each aspect. For comparison, a paired *t*-test was conducted on the average score gap between *FastCov*$_{C+S}$*-select* and each of the other methods, with results as shown in Table 10.

Table 10. Paired *t*-Test Results for the Hypothesis of *AvgScore*(*FastCov*$_{C+S}$*-select*) > *AvgScore*(M_i)

Method name (M_i)	*t* value (sig.)				
	Overall	Taste	Environment	Services	Price
Portfolio	$11.26^{(***)}$	$7.29^{(***)}$	$5.45^{(***)}$	$8.84^{(***)}$	$7.06^{(***)}$
MCDC	$8.30^{(***)}$	$8.31^{(***)}$	$2.13^{(*)}$	$7.00^{(***)}$	$3.65^{(***)}$
FM-LDA	$10.33^{(***)}$	$5.42^{(***)}$	$5.20^{(***)}$	$9.33^{(***)}$	$3.25^{(**)}$
K-means based	$9.33^{(***)}$	$8.14^{(***)}$	−0.99 (0.323)	$4.18^{(***)}$	$5.09^{(***)}$

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Table 10 explicitly shows that the proposed *FastCov*$_{C+S}$*-select* method significantly outperformed other methods on all aspects but one (i.e., environment) where the difference was not statistically significant from the *K-means*-based method.

Table 11. Paired *t*-Test Results for the Hypothesis of *AvgScore*(*JS*) > *AvgScore*(EU)

	Overall	Taste	Environment	Services	Price
t value (sig.)	$3.104^{(**)}$	$3.373^{(**)}$	$3.803^{(***)}$	$2.611^{(*)}$	$2.335^{(*)}$

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Since the selection of similarity metric may significantly impact the performance of the proposed method, an additional human evaluation experiment, in which 80 human evaluators were recruited, was conducted on the same online reviews data. Concretely, Euclidean similarity on literal content and JS-divergence similarity measuring with sentiment analysis (Li et al. 2010), denoted as Topic-Sentiment similarity, were employed and the comparative result are as shown in Table 11. The results reveal that, in the context of analyzing online reviews, where the topics and sentiments do matter, a more appropriate similarity metric, i.e., Topic-Sentiment similarity, significantly outperforms the Euclidean Similarity (i.e., only considering literal content), which further indicates that the similarity metric should be carefully selected in different contexts.

6. Conclusions and Future Work

With the information overload on Internet, it is very helpful for both information search service providers

and users to extract a small set of search or recommendation results that possess high diversity (i.e., high content coverage and high structure coverage). This paper has investigated how to build an extraction method to obtain a diverse result set when considering information coverage metrics from a combined perspective of content and structure. More specifically, we have proposed a heuristic algorithm Cov_{C+S}-*Select* by applying the strategy of simulated annealing on the greedy submodular idea of Cov_C-*Select*. Based on these, a fast approximation method called *FastCov*$_{C+S}$-*Select* has been further proposed, aimed to extract diverse results in an effective, efficient, and robust manner, which has been demonstrated by evaluation experiments. On top of that, we have conducted a comprehensive and systematic investigation of 11 major diversity extraction methods in comparison with *FastCov*$_{C+S}$-*Select* from three perspectives, namely, information coverage, external labeling, and human evaluation. The comparison experiments further assert the superiority of the proposed method.

Future work can focus on two aspects: (1) further investigate the notion of information redundancy and its relationship with other metrics from their respective angles of interest and (2) explore and design an extended approach to targeting the coverage and redundancy as optimization goals.

Acknowledgments

The authors thank the editors and reviewers for their valuable comments to substantially improve this paper.

Endnotes

[1] KDD CUP 2005: http://www.sigkdd.org/kdd-cup-2005-internet-user-search-query-categorization.

References

Agrawal R, Gollapudi S, Halverson A, Ieong S (2009) Diversifying search results. *Proc. 2nd ACM Internat. Conf. Web Search Data Mining, Barcelona, Spain*, 5–14.

Aliguliyev RM (2009a) Clustering of document collection—A weighting approach. *Expert Syst. Appl.* 36(4):7904–7916.

Aliguliyev RM (2009b) Performance evaluation of density-based clustering methods. *Inform. Sci.* 179(20):3583–3602.

Baeza-Yates R, Ribeiro-Neto B (1999) *Modern Information Retrieval* (Addison-Wesley, New York).

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J. Mach. Learn. Res.* 3(Jan):993–1022.

Brynjolfsson E, Hu Y, Smith MD (2003) Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Sci.* 49(11):1580–1596.

Carbonell J, Goldstein J (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proc. 21st Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval, Melbourne, Australia*, 335–336.

Carpineto C, Mizzaro S, Romano G, Snidero M (2009a) Mobile information retrieval with search results clustering: Prototypes and evaluations. *J. Amer. Soc. Inform. Sci. Tech.* 60(5):877–895.

Carpineto C, Osinski S, Romano G, Weiss D (2009b) A survey of Web clustering engines. *ACM Comput. Surv.* 41(3):1–38.

Carterette B, Chandar P (2009) Probabilistic models of ranking novel documents for faceted topic retrieval. *Proc. 18th ACM Conf. Inform. Knowl. Management, Hong Kong, China*, 1287–1296.

Černú V (1985) Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *J. Optim. Theory Appl.* 45(1):41–51.

Chen H, Karger DR (2006) Less is more: Probabilistic models for retrieving fewer relevant documents. *Proc. 29th Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval, Seattle*, 429–436.

Clarke CLA, Craswell N, Soboroff I (2010) Overview of the TREC 2009 web track. *Proc. 18th Conf. Text Retrieval, Gaithersburg, MD.*

Clarke CLA, Kolla M, Cormack GV, Vechtomova O, Ashkan A, Büttcher S, MacKinnon I (2008) Novelty and diversity in information retrieval evaluation. *Proc. 31st Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval, Singapore*, 659–666.

Cormen TH, Leiserson CE, Rivest RL, Stein C (2009) *Introduction To Algorithms*, 3rd ed. (MIT Press, Cambridge, MA).

De P, Hu Y, Rahman MS (2010) Technology usage and online sales: An empirical study. *Management Sci.* 56(11):1930–1945.

Dubois D, Prade H (1985) Fuzzy cardinality and the modeling of imprecise quantification. *Fuzzy Sets Syst.* 16(3):199–230.

Endres DM, Schindelin JE (2003) A new metric for probability distributions. *IEEE Trans. Inform. Theory* 49(6):1858–1860.

Fung BCM, Wang K, Ester M (2003) Hierarchical document clustering using frequent itemsets. *Proc. SIAM Internat. Conf. Data Mining* (2003), *San Francisco, CA*, 59–70.

Grabmeier J, Rudolph A (2002) Techniques of cluster algorithms in data mining. *Data Min. Knowl. Discov.* 6(4):303–360.

Granville V, Krivanek M, Rasson JP (1994) Simulated annealing: A proof of convergence. *IEEE Trans. Pattern Anal. Machine Intell.* 16(6):652–656.

Gregor S, Hevner AR (2013) Positioning and presenting design science research for maximum impact. *MIS Quart.* 37(2):337–355.

Han J, Pei MK (2011) *Data Mining: Concepts and Techniques*, 3rd ed. (Morgan Kaufmann, New York).

He J, Meij E, Rijke Md (2011) Result diversification based on query-specific cluster ranking. *J. Amer. Soc. Inform. Sci. Tech.* 62(3): 550–571.

Herrera F, Martínez L (2000) A 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Trans. Fuzzy Syst.* 8(6):746–752.

Hochba DS (1997) Approximation algorithms for NP-hard problems. *ACM SIGACT News* 28(2):40–52.

Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inform. Syst.* 20(4):422–446.

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Sci.* 220(4598):671–680.

Krishnan S, Goldberg K (2015) The minimum conductance dissimilarity cut (MCDC) algorithm to increase novelty and diversity of recommendations. Working paper, Industrial Engineering and Operations Research Department, University of California, Berkeley, CA.

Kullback S, Leibler RA (1951) On information and sufficiency. *Ann. Math. Stat.* 22(1):79–86.

Li F, Huang M, Zhu X (2010) Sentiment analysis with global topics and local dependency. *Proc. 24th AAAI Conf. Artificial Intelligence* (*AAAI* 2010), *Atlanta*, 1371–1376.

Liu B (2011) *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, 2nd ed. (Springer-Verlag, Berlin).

Ma B, Wei Q (2012) Measuring the coverage and redundancy of information search services on e-commerce platforms. *Electron. Commer. Res. Appl.* 11(6):560–569.

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. Le Cam LM, Neyman J, eds. *Proc. 5th Berkeley Symposium Math. Stat. Probab.* (University of California Press, Berkeley, CA), 281–297.

Malik H, Kender J, Fradkin D, Moerchen F (2010) Hierarchical document clustering using local patterns. *Data Mining Knowl. Disc.* 21(1):153–185.

Manning CD, Raghavan P, Schütze H (2008) *Introduction to Information Retrieval* (Cambridge University Press, Cambridge, UK).

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21(6):1087–1092.

Mulpuru S (2008) The state of retailing online 2008: Merchandising and web optimization report. Report, Forrester Research, Cambridge, MA. Accessed April 12, 2014, https://www.forrester.com/report/The+State+Of+Retailing+Online+2008+Merchandising+And+Web+Optimization+Report/-/E-RES46187.

Nemhauser GL, Wolsey LA, Fisher ML (1978) An analysis of approximations for maximizing submodular set functions. *Math. Program.* 14(1):265–294.

Pan F, Wang W, Tung AKH, Yang J (2005) Finding representative set from massive data. *Proc. 5th IEEE Internat. Conf. Data Mining, Houston, TX*, 338–345.

Qin L, Zhu X (2013) Promoting diversity in recommendation by entropy regularizer. *Proc. 23rd Internat. Joint Conf. Artificial Intelligence, Beijing, China*, 2698–2704.

Ralescu D (1995) Cardinality, quantifiers, and the aggregation of fuzzy criteria. *Fuzzy Sets Systems* 69(3):355–365.

Robertson SE, Walker S, Jones S, Hancock-Beaulieu M, Gatford M (1994) Okapi at TREC-3. *Proc. 3rd Conf. Text Retrieval, Gaithersburg, MD*, 109–126.

Santos R, Macdonald C, Ounis I (2010) Selectively diversifying web search results. *Proc. 19th ACM Internat. Conf. Inform. Knowl. Management, Toronto, ON, Canada*, 1179–1188.

Santos R, Macdonald C, Ounis I (2011) Intent-aware search result diversification. *Proc. 34th Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval, Beijing, China*, 595–604.

Sen R, King RC, Shaw MJ (2006) Buyers' choice of online search strategy and its managerial implications. *J. Management Inform. Syst.* 23(1):211–238.

Shi X, Liang X (2015) Resolving inconsistent ratings and reviews on commercial webs based on support vector machines. *Proc. 12th Internat. Conf. Service Syst. Service Management, Guangzhou, China*, 1–6.

Silverstein C, Marais H, Henzinger M, Moricz M (1999) Analysis of a very large web search engine query log. *SIGIR Forum* 33(1):6–12.

Spink A, Jansen BJ, Wolfram D, Saracevic T (2002) From e-sex to e-commerce: Web search changes. *IEEE Comput.* 35(3): 107–109.

Spink A, Wolfram D, Jansen MBJ, Saracevic T (2001) Searching the web: The public and their queries. *J. Amer. Soc. Inform. Sci. Tech.* 52(3):226–234.

Suman B, Kumar P (2006) A survey of simulated annealing as a tool for single and multiobjective optimization. *J. Oper. Res. Soc.* 57(10):1143–1160.

Tsur O, Davidov D, Rappoport A (2010) ICWSM-A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. *Proc. 4th Internat. AAAI Conf. Weblogs Social Media, Washington, DC*, 162–169.

Wang J, Zhu J (2009) Portfolio theory of information retrieval. *Proc. 32nd Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval, Boston*, 115–122.

Whissell J, Clarke C (2011) Improving document clustering using Okapi BM25 feature weighting. *Inform. Retrieval* 14(5):466–487.

Xiong H, Wu J, Chen J (2009) K-means clustering versus validation measures: A data-distribution perspective. *IEEE Trans. Syst. Man Cybernetics* 39(2):318–331.

Zhai C, Lafferty J (2006) A risk minimization framework for information retrieval. *Inform. Processing Management* 42(1):31–55.

Zhai C, Cohen WW, Lafferty J (2003) Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. *Proc. 32nd Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval, Toronto, Canada*, 10–17.

Zhao Y, Karypis G (2004) Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learn.* 55(3):311–331.

Zhao Y, Karypis G (2005) Hierarchical clustering algorithms for document data sets. *Data Mining Knowl. Discov.* 10(2):141–168.

Zhong S (2006) Semi-supervised model-based document clustering: A comparative study. *Machine Learn.* 65(1):3–29.

Zhuang J, Hoi SCH, Sun A (2008) On profiling blogs with representative entries. *Proc. 2nd Workshop Analytics Noisy Unstructured Text Data, Singapore*, 55–62.