Electronic Commerce Research and Applications 11 (2012) 560–569

Contents lists available at SciVerse ScienceDirect

Electronic Commerce Research and Applications

journal homepage: www.elsevier.com/locate/ecra

Measuring the coverage and redundancy of information search services on e-commerce platforms

Baojun Ma, Qiang Wei [*]

School of Economics and Management, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Available online 20 September 2012

Keywords:
Information coverage
Information redundancy
Information search
Information structure

ABSTRACT

Today's widespread e-commerce applications pose a new challenge to information search services. They must extract a useful small set of search or recommendation results from a larger set that preserves information diversity. This paper proposes a novel metric setting to measure two important aspects of information diversity, information coverage and information redundancy. In addition to content coverage, we consider another important measure of information coverage called *structure coverage*, and model it using information entropy. This approach can better convey the information coverage of the extracted small set with respect to the original large set. The proposed metrics are effective and have various useful properties, which are demonstrated by theoretical and experimental analysis. We also designed a calculation method that shows good computational efficiency. Finally, we conducted an experiment using real data from online customer reviews to further emphasize the effectiveness of the proposed metrics.

1. Introduction

Electronic commerce (e-commerce) is the process of buying, selling, or exchanging products, services, and information via computer networks, including the Internet (Turban et al. 2010). In response to the recent rapid growth of e-commerce, many online e-commerce platforms have developed Web-based information search service systems to process the enormous number of transactions that occur via the Internet (Chen et al. 2009, Vuylsteke et al. 2010). With the development of e-commerce and online shopping, Internet consumers face a dizzying array of product choices and consequently suffer from information overload (Brynjolfsson et al. 2003, De et al. 2010). To help consumers find the products or information they actually need, which is usually a small subset of the overall relevant information, nearly all Internet retailers have started to provide and plan to continue to invest in improving information search service technologies (e.g., user search functions or recommendation systems) on their platforms (Mulpuru 2008).

Past research has demonstrated that information search services, if used properly, can significantly enhance consumers' shopping experiences by reducing search costs (Jie et al. 2006, Kumar and Lang 2007, Sen et al. 2006) and can yield significantly higher revenues and profits for Internet companies (De et al. 2010, Kuruzovich et al. 2010, Siwicki 2007). In addition, consumers usually use search engines or recommender systems to obtain and compare product information and other consumers' reviews and opinions about products from different e-commerce platforms (Kumar and Lang 2007, Sen et al. 2006).

Because usually only a small set of search results or recommended results can be browsed by customers (e.g., the top-ranked relevant links, the first several pages of online reviews, or the first several recommended products) (Liu 2007, Silverstein et al. 1999, Spink et al. 2002a,b, 2001), the information quality of the small set is of great importance to both information search service providers and customers. Ranking criteria, such as PageRank value, hotness, freshness, number of comments or visits, or helpfulness score, are widely used by many online information search service systems to help customers efficiently receive a small but useful set of information without browsing all the search results (Boldi et al. 2009, Fox et al. 2005, Ghose and Ipeirotis 2011, Mudambi and Schuff 2010).

Nevertheless, information search services on e-commerce platforms face new challenges because the top-ranked results on the first several pages cannot effectively and sufficiently reflect the diversity of all the retrieved results. In many cases, customers are increasingly concerned with information diversity: they prefer to have a representative view of all results satisfying the search criteria rather than a small set of top-ranked results (De et al. 2010, Sen et al. 2006). For example, a customer may browse all online product reviews before making a purchase decision, but the first several pages of the most timely, most commented on, or highest helpfulness-ranked reviews cannot reflect the diversity of all reviews (Tsaparas et al. 2011). In online shopping, though search-bots or recommender systems could easily present customers with hundreds of products or services satisfying the basic query criteria

[*] Corresponding author. Tel.: +86 10 62789824.
E-mail address: weiq@sem.tsinghua.edu.cn (Q. Wei).

1567-4223/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.elerap.2012.09.001

they provide, it would still be a challenging task for customers to summarize or compare all relevant information before selecting the appropriate product or service, especially for cold-start buyers (Adomavicius and Tuzhilin 2005, Fleder and Hosanagar 2009).

Therefore, only a small set of the results would be browsed due to cost and time limitations; thus, customers may be concerned about the extent of the information that the small set conveys compared to the entire results – the information diversity. Moreover, when users conduct searches on the web, the search intent and goals hidden behind search behaviors, such as navigational, resourceful, transactional, and informational searches (Broder 2002, Rose and Levinson 2004), may also be diverse. Essentially, a customer, driven by an information need, submits a product or service query, which is often ambiguous (Spärck-Jones et al. 2007). For instance, a query may not express a clearly defined category (e.g., 'jaguar'), or it may represent a genuine need for broader coverage of a clearly defined category (e.g., 'jaguar car brand'). In the first case, the query is open to different interpretations (e.g., a type of animal, a car brand, a type of cocktail or an operating system) (Agrawal et al. 2009, Chen and Karger 2006, Radlinski and Dumais 2006), whereas in the second case, the customer might be interested in different aspects or subtopics related to the query (e.g., models, prices, history of the company, etc.) (Carterette and Chandar 2009, He et al. 2011, Santos et al. 2010a). In these cases, without explicit or implicit customer feedback or usage history, information search service systems need to provide customers with more diverse information from the entire search results – information diversity (Boyce 1982, Goffman 1964, He et al. 2011).

Consequently, information diversity is one of the most important goals for online information search services (Lathia et al. 2010, Tsaparas et al. 2011, Vargas et al. 2011). Diversity possesses two major aspects, information coverage and information redundancy (Clarke et al. 2008, Hurley and Zhang 2011, Santos et al. 2011, Tsaparas et al. 2011, Xu and Yin 2008). Information coverage is a measure of the extent that a small extracted set reflects the information load of the set of all search results. For instance, given an original search results set $D = \{a,a,a,b,b,b,b,b,b\}$, if a customer can only browse a small set with three results, then, intuitively, $D_1 = \{a,b,b\}$ is better than $D_2 = \{b,b,b\}$ because D_1 conveys more contents of D than D_2 does. That is, the content coverage of D_1 is higher than D_2. Moreover, if two objects, d and d', are similar to each other, then d' can be regarded as partially covering the information content of d and vice versa. Clearly, similarity-based information coverage is widespread in online information search services (e.g., similar reviews, similar products or services, and similar search results), and many existing works also incorporate similarities (Hurley and Zhang 2011, Vargas et al. 2011, Zhang et al. 2005).

Furthermore, in addition to content coverage, another aspect of coverage, structure coverage, is important and worth investigating, especially when one is concerned with information diversity. *Structure coverage* is a measure of to what extent a small set can cover the information structure of the original set. Using the same example above, consider another extracted set $D_3 = \{a,a,b\}$. With respect to D, both D_1 and D_3 have the same content coverage (i.e., both a and b are covered), but customers may prefer D_1 over D_3 because D_1 conveys more appropriate diversity than D_3 since the percentage distribution (i.e., the information structure) in D_1 is consistent with that of D, whereas that of D_3 is not. D_3 could possibly cause a biased perception of the entire set of search results. Thus, structure coverage reflects a meaningful aspect of information coverage.

For instance, when browsing online reviews, the contents (e.g., opinions and sentiments) can significantly affect customers' purchasing behaviors, but the information structure (e.g., the contrast of positive and negative reviews) may play a more important role (Tsaparas et al. 2011). For online shopping recommender systems, the constitution structure of recommended products or services may also exercise a great influence on the customers (Vargas et al. 2011, Zhang and Hurley 2008). However, little effort is focused on information structure when designing and measuring information search services on e-commerce platforms. Therefore, this paper intends to provide an in-depth investigation of information coverage that is concerned with both content and structure.

In addition to information coverage, information redundancy is another important aspect of measuring information diversity and can frequently be observed in web-based information search services. Research on large-scale e-commerce data has showed that redundant information in query results can remarkably reduce customers' satisfaction with information search services such as web search (Bernstein and Zobel 2005), product search and recommendation (Adomavicius and Tuzhilin 2005, Herlocker et al. 1999, Zhang et al. 2002), and online review ranking (Duan et al. 2008, Ghose and Ipeirotis 2011, Koh et al. 2010, Mudambi and Schuff 2010). For instance, given a new extracted set $D_4 = \{a,a,b,b,b,b\}$, though the information coverage, including both content and structure, is the same as that of D_1, D_4 is obviously more redundant because of the duplicated elements and could result in decreased customer satisfaction. Moreover, as with information coverage, information redundancy can also be extended using a similarity-based framework (Carbonell and Goldstien 1998, Vargas et al. 2011, Zhang et al. 2002), for example, if d is similar to d', then d' is regarded as partially redundant with respect to d and vice versa.

Therefore, diversity is an important quality for information search services on e-commerce platforms, where information coverage, including content coverage and structure coverage, as well as information redundancy needs to be considered. More explicitly, for an information search service on an e-commerce platform, the small set of diverse results should have high information coverage and low information redundancy with respect to the original data set (Clarke et al. 2008, Santos et al. 2010a). Recently, various efforts have been made to present users with highly diverse information using different methods (Agrawal et al. 2009, Allan and Raghavan 2002, Carbonell and Goldstien 1998, Carterette and Chandar 2009, Chen and Karger 2006, Gollapudi and Sharma 2009, He et al. 2011, Radlinski and Dumais 2006, Radlinski et al. 2008, Rafiei et al. 2010, Santos et al. 2010a,b; Spärck-Jones et al. 2007, Vee et al. 2008, Wang and Zhu 2009, Xu and Yin 2008, Yue and Joachims 2008, Zhai and Lafferty 2006, Zhai et al. 2003, 2005), most of which are based on a greedy approximation strategy to make tradeoffs between the relevance and diversity of search results (Santos et al. 2010a). However, though important, formulating good evaluation metrics for diversity remains a challenging task, and, in particular, structure coverage has not been treated yet; this state of affairs motivates our efforts. Hence, we present a novel metric setting to evaluate information diversity in terms of both coverage and redundancy. It is noteworthy that structure coverage is evaluated using information entropy and considered with content coverage in a combined manner.

The remainder of this article is organized as follows. Section 2 overviews related work on existing metrics, including their limitations. Section 3 proposes a novel metric setting for evaluating diversity using coverage and redundancy measures, where coverage is defined in terms of both content and structure. Section 4 introduces the procedure for calculating the corresponding metrics and presents an illustrative example to demonstrate the advantages of the proposed metrics. An experimental analysis and evaluation of the proposed metrics using real e-commerce data is discussed in Section 5. Section 6 concludes.

2. Related work

There are a number of evaluation metrics, such as *recall*, *precision* (Kraft and Bookstein 1978), *MAP* (Buckley and Voorthees

2000), *bpref* (Buckley and Voorthees 2004) and *nDCG* (Järvelin and Kekäläinen 2002), that are commonly used to assess the quality of search results and recommended results using binary relevance judgments or grade relevance judgments. These metrics are generally based on the probabilistic ranking principle (Robertson 1977) that assumes that the relevance of a search result is independent of the others, which thus cannot be utilized to evaluate information diversity considering the relations among search results (Clarke et al. 2008).

A variety of attempts to formulate related metrics that consider the two major aspects of diversity (i.e., information coverage and information redundancy) using different methods have been made. For information coverage, Pan et al. (2005) proposed a metric that divides the number of distinct classes in the subset by the total number of the classes using predefined class labels. Zhai et al. (2003) offered a similar coverage measure for subtopic retrieval applications. These two metrics suffer from the limitation that it is impossible to calculate the metric values without predetermined class labels, which makes the metric results quite sensitive to the clustering and classification processing. Moreover, the method of simply treating similar objects in the same class equally leads to inaccurate information coverage ratios. Zhuang et al. (2008) proposed two representativeness metrics to quantify the information coverage of blog profiling. However, both of the metrics do not satisfy reflexivity, meaning that the metrics would not consider a blog set to cover the information contained in the set itself, which is counterintuitive. Moreover, one of the two measures also faces the same problem of sensitivity to clustering results.

There are three main types of metrics proposed to evaluate the redundancy of one object with respect to a set, maximum similarity based (Carbonell and Goldstien 1998), minimum similarity based (Pan et al. 2005), and average similarity based (Zhuang et al. 2008), each of which uses a different approach for measuring an object with respect to a set. However, how to measure the overall redundancy rate inside a set has not been thoroughly investigated. If each of the above three strategies is directly extended, it may cause some distortion. First, the direct extension of the maximum or minimum strategy to reflect the redundancy of a set could be easily affected by a single outlier. Second, even if the average strategy were adopted, the metric value would still be inconsistent with intuition in certain special cases. For instance, given an n-size set with n_a objects a and $n - n_a$ objects b, if n is large, the intuitively perceived redundancy should be close to 100%, whereas the value of the redundancy metric using the average strategy would approach zero. Thus, it is considered necessary and meaningful to design an information redundancy metric to measure redundancy by taking into account the objects inside a set.

In addition to the above efforts at measuring diversity by considering coverage and redundancy, there are two attempts at introducing aggregated metrics to evaluate the overall diversity, *α-nDCG@k* (Clarke et al. 2008) and *precision-IA@k* (Agrawal et al. 2009, Clarke et al. 2010). Briefly, the *α-nDCG@k* metric adopted the *nDCG* measure (Järvelin and Kekäläinen 2002) to address both coverage and redundancy simultaneously by using manual judgments of search results. The parameter α denotes the probability that an assessor makes a relevance judgment error. The *precision-IA@k* metric used the concept of intent-aware precision, which extends the traditional notion of precision to account for the possible aspects underlying a query and their relative importance. However, these two metrics have some limitations. First, these two metrics can only be applied to ranked results, which makes them inapplicable to many real-world applications. Second, both metrics require manual relevance judgments, which is usually costly and unrealistic. Furthermore, human assessors are known to be inconstant with their judgments (Voorhees 1998). Relevance judgments inferred from implicit user feedbacks may be inaccurate (Agichtein et al. 2006a,b; Dupret et al. 2007, Joachims et al. 2005). Although the *α-nDCG@k* metric introduces the parameter α to handle the error rate, it is still too costly and impractical to require human assessment each time a user wants to evaluate the coverage of a list of search results. Third, for both metrics, predefined taxonomies or aspects for the query are needed to help calculate the values of metrics, but the taxonomies may be obsolete, and the metric values may be sensitive to the assumed taxonomies. More importantly, it is very costly or even impossible to maintain timely taxonomies in the rapidly updating information environments of e-commerce platforms.

3. A novel metric setting for information diversity

Information diversity can be measured using a metric setting that combines coverage and redundancy, where coverage includes two aspects, content and structure. Without loss of generality, consider an original set of objects $D = \{d_1, d_2, \ldots, d_n\}$, where n is normally large and d_i, with $i = 1, 2, \ldots, n$, denotes an object, such as webpage content, a document, a product or service description, or a product review, satisfying the basic query or search criteria customers provide. Due to search cost and time limitations, customers often only browse a small set D', where $D' = \{d'_1, d'_2, \ldots, d'_m\} \subseteq D$, and usually $m \ll n$. Thus, D' can partially cover the information content of D when the objects in D' are identical to or similar to the objects in D.

Because similarity among search results (e.g., web content, product descriptions, or reviews) cannot be ignored by information search services on e-commerce platforms, for two objects d and d', d, $d' \in D$, their similarity, denoted as $sim(d, d')$, could be measured and calculated using existing methods (Aliguliyev 2009b, Atlam et al. 2000, Chehreghani et al. 2009, Huang 2008). The evaluation metrics discussed below will incorporate the similarities between search results.

3.1. Information coverage

Given two objects $d \in D$ and $d' \in D'$, d covers the content of d' with degree $sim(d, d')$ and vice versa. Therefore, given an extracted set D' and an object $d \in D$, we can define the degree to which D' covers the content of d as the maximum similarity degree, $\max_{j=1,2,\ldots,m}(sim(d'_j, d))$, where $d'_j \in D'$, which means that the degree a set covers the content of an object d is determined by the object in the set most similar to d. Obviously, D' is able to cover all the information inherent in d if d appears in D'. If not, without further knowledge, $\max_{j=1,2,\ldots,m}(sim(d'_j, d))$ is the best approximation and can be regarded as the lower bound of the real or ideal information content coverage.

To reflect the overall content coverage of D' with respect to D, the mathematical average operator can be used to aggregate the content coverage degree of all objects in D. Therefore, the degree to which D' covers the content of D, denoted as $Cov_C(D', D)$, can be constructed as follows:

$$Cov_C(D', D) = \frac{\sum_{d \in D} \max\limits_{j=1,2,\ldots,m} (sim(d'_j, d))}{n} \qquad (1)$$

$Cov_C(D', D)$ possesses some useful properties. First, it is in the range $[0,1]$ and is reflexive, so $Cov_C(D, D) = 100\%$, and monotonic, so if $D'' \subseteq D'$ then $Cov_C(D'', D) \leqslant Cov_C(D', D)$. Second, $m/n \leqslant Cov_C(D', D) \leqslant 1$ because at least m objects in D appear in D'.

Nevertheless, as discussed in the previous sections, $Cov_C(D', D)$ is good only at measuring information coverage with respect to content, but it ignores the information structure. For an example, consider an original set $D = \{a, a, b, b, b, b, b, b\}$ and two extracted sets $D_1 = \{a, b, b, b\}$ and $D_2 = \{a, a, b, b\}$ and suppose $sim(a, b) = 0$ for simplicity. The content coverage degrees of both D_1 and D_2 with respect to D are 100% because both D_1 and D_2 contain the objects a and b; however, D_1 conveys more suitable information coverage than D_2 because it is more representative of the structure of D.

To incorporate information structure into our analysis, the widely used information entropy (Shannon 1948) is used. Consider a set D with n objects classified into m distinct groups, where the number of objects in each group is n_j, with $j = 1,2,\ldots,m$ and $n_1 + n_2 + \cdots + n_m = n$. Then, the information entropy can be calculated as $-\frac{1}{\log\ m}\sum_{j=1}^{m}\frac{n_j}{n}\log\left(\frac{n_j}{n}\right)$ and used to express the distribution of information load, which is the information structure of D.

In the previous example, the reason that D_1 better captures the information structure of D than D_2 does is that the entropy of $D_1(-1/\log(2) * (1/4 * \log(1/4) + 3/4 * \log(3/4)) = 81\%)$ is closer to that of D (81%) than that of D_2 (65%). Therefore, D_1 fully covers the information structure of D, whereas D_2 does not.

However, the prerequisite of the above entropy calculation is to perform classification operations on D and D', respectively, which not only requires high computational complexity but also results in high sensitivity to the classification procedures, as discussed in Section 2. Therefore, we simplify the calculation using an assignment operation. From the perspective of information structure, the ideally extracted set should be D, because every object has been equally represented. Thus, each object in D could be treated as a default class label, resulting in n classes with 1 object in each, and the ideal entropy is 100%. Given an extracted set D', each object d in D could be assigned to the corresponding d', if $d = d'$ in a crisp sense. If there exist multiple class labels in D' equal to d, then d could be assigned randomly or uniformly to one of them. Then, the entropy of D' could be calculated without classifying D and D'. For the same example, the entropy of D_1 is 100% (i.e., $-1/\log(4) * (2/8 * \log(2/8) + 2/8 * \log(2/8) + 2/8 * \log(2/8) + 2/8 * \log(2/8))$ and the entropy of D_2 is 91% (i.e., $-1/\log(4) * (1/8 * \log(1/8) + 1/8 * \log(1/8) + 3/8 * \log(3/8) + 3/8 * \log(3/8))$. Clearly, D_1 is better than D_2 at conveying the information structure of D because the entropy of D_1 is higher than that of D_2. Then, given a set D and an extracted set D', the entropy of D' calculated using the assignment operation could be used to denote the structure coverage of D' with respect to D.

In real applications, because usually the similarity of any two search results is neither 0 nor 1, but rather in the range (0,1), the above assignment operation should be extended accordingly. Each object in D' could be treated as a class label, resulting in m classes in D'. Then, each d in D could be assigned into the class with label k, where $k = \arg\max_{j=1,2,\ldots,m}(sim(d'_j, d))$ with $1 \leqslant k \leqslant m$. After the assignment, the n objects in D could be assigned into m classes, denoted as $D_1, D_2, \ldots, D_m$, respectively. If an object d is assigned to D_j, then d belongs to D_j with $sim(d'_j, d)$, where d'_j is the label of D_j. The total information load of all objects in D_j is $\sum_{d\in D_j} sim(d'_j, d)$, denoted as n^v_j. Overall, for m sets, the total information load of all objects is $\sum_{j=1,2,\ldots,m} n^v_j$, denoted as n^v.

Based on the above assignment operation, the information entropy could also be extended to measure the structure coverage of D' with respect to D. In this regard, the information structure coverage of D' with respect to D, denoted as $Cov_S(D', D)$, could be constructed as follows:

$$Cov_S(D', D) = -\frac{1}{\log m}\sum_{j=1}^{m}\frac{n^v_j}{n^v}\log\left(\frac{n^v_j}{n^v}\right) \tag{2}$$

$Cov_S(D', D)$ exhibits several interesting properties. First, it is in the range (0,1] and possesses reflexivity, i.e., $Cov_S(D', D) = 100\%$. Second, if $n^v_{j_1} = n^v_{j_2}$ for any $j_1, j_2 = 1,2,\ldots,m$, then $Cov_s(D', D) = 100\%$, which means that if the information load in D could be conveyed uniformly into D', then D' preserves the best information structure of D. Furthermore, without further knowledge of the information structure of D, if the information load in D could be more uniformly assigned into D', $Cov_s(D', D)$ would be higher. This property is also important for designing better methods for extracting sets with higher coverage. Thereafter, the proposed $Cov_s(D', D)$ could be used to measure the structure coverage of D' with respect to D. Furthermore, we can aggregate the structure coverage metric with the content coverage metric to construct a so-called information coverage metric, denoted as $Cov(D', D)$, as described in Definition 1.

Definition 1. Given an original set D with n objects and an extracted set D' with m objects, where $D' \subseteq D$, the information coverage of D' with respect to D is defined as

$$Cov(D', D) = \begin{cases} Cov_C(D', D) = \frac{\sum_{d\in D} sim(d'_1, d))}{n} & if\ m = 1 \\ Cov_C(D', D) \times Cov_S(D', D) \quad = \frac{\sum_{d\in D}\max_{j=1,2,\ldots,m}(sim(d'_j, d))}{n} & \\ \qquad \times\left(-\frac{1}{\log m}\sum_{j=1}^{m}\frac{n^v_j}{n^v}\log\left(\frac{n^v_j}{n^v}\right)\right) & if\ m > 1 \end{cases} \tag{3}$$

Definition 1 reveals that, if $m = 1$ (there is only one extracted object), the structure coverage could be ignored, which is consistent with intuition. When $m > 1$, in addition to content coverage, structure coverage will also play an important role in a human's perception of information diversity.

Similarly, the information coverage metric also has some useful properties. First, $0 < Cov(D', D) \leqslant 1$ and $Cov(D, D) = 100\%$. Second, if any two objects in D are completely different, i.e., $sim(d_1, d_2) = 0$ for any $d_1, d_2 \in D$, then $Cov(D', D) = m/n$, which shows that the extracted set D' can only convey its own information load.

3.2. Information redundancy

In contrast to the information coverage metric, which is a binary relation on two sets, i.e., it compares D' to D, information redundancy considers a single set, such as D'. Given two search results d_1 and $d_2 \in D'$, d_1 is called redundant with respect to d_2 with degree of $sim(d_1, d_2)$ because clearly, part of the information about d_2 has been duplicated by d_1 and vice versa. This concept can be extended to measure the extent to which an object d_1 is redundant in D'. To avoid distortions caused by direct extensions, as discussed in Section 2, the degree to which d_1 is redundant in D', denoted as $Red(d_1, D')$, where $d_1 \in D'$, could be defined as follows:

$$Red(d_1, D') = 1 - \frac{1}{\sum_{d\in D'} sim(d_1, d)} \tag{4}$$

In Eq. (4), $\sum_{d\in D'} sim(d_1, d)$ represents the total amount of redundant information associated with d_1 in D'; thus, $\frac{1}{\sum_{d\in D'} sim(d_1, d)}$ could denote the proportion of d_1's information in $\sum_{d\in D'} sim(d_1, d)$. Therefore, $Red(d_1, D')$ is the proportion of other objects' information that is duplicated by d_1, which could be regarded as the degree of information redundancy caused by d_1 in D'. In turn, the degree of redundancy in the set D', denoted as $Red(D')$, could be defined via Definition 2.

Definition 2. Given a set D' containing m objects, $Red(D')$ is defined as follows:

$$Red(D') = \frac{\sum_{d_1\in D'} Red(d_1, D')}{m} = \frac{1}{m} \times \sum_{d_1\in D'}\left(1 - \frac{1}{\sum_{d\in D'} sim(d_1, d)}\right) \tag{5}$$

$Red(D')$ is used to measure the average information redundancy in D'. Moreover, the metric also possesses some useful properties. First, $0 \leqslant Red(D') < 1$ if D' is not empty. Second, if $m = 1$, that is, there is only one object in D', or any two objects in D' are mutually different, that is, for any $d_1, d_2 \in D'$, $sim(d_1, d_2) = 0$, then $Red(D') = 0$.

Third, if all the objects are the identical in D', then $Red(D') = 1 - 1/m$, meaning that there is only $1/m$ information load in D' that not redundant. Thus, the proposed $Red(D')$ metric also can be interpreted intuitively.

4. Calculation procedure and an illustrative example

As discussed in the previous sections, deriving the degrees of information coverage and information redundancy requires a multi-stage calculation procedure, which is illustrated in Fig. 1. First, given an original set D and an extracted set D' containing, such as search results, product descriptions, or reviews, the related data pre-processing operations, including document modeling and similarity calculation (line 1–5), should be performed. For online search or recommendation, objects (e.g., web documents, online reviews or product descriptions) are usually in the form of text; thus, word segmentation (line 1), stemming (line 2) and stop-word removal (line 3) operations are necessary. Therefore, the widely used vector space model (VSM) (Liu 2007, Salton et al. 1975) could be used to model each object as a multi-dimensional vector (line 4). Second, based on the derived vectors, the similarities between any two objects should also be calculated (line 5). For instance, the cosine measure is one of the most popular similarity measures applied to text documents (Baeza-Yates and Ribeiro-Neto 1999, Huang 2008, Korenius et al. 2007, Salton 1971). Third, based on the similarities among objects, the operation for assigning each object in D with the corresponding label in D', preserving the maximum similarity, could be conducted (line 6–15). Thereafter, $Cov_C(D',D)$ and $Cov_S(D',D)$ could be calculated, and the final $Cov(D',D)$ could also be calculated (line 16–18). Simultaneously, $Red(D')$ could also be calculated (line 19). More importantly, the procedure could be conducted without human intervention, which overcomes the existing limitations.

Generally, given a set D of size n and extracted set D' of size m, the computational complexities of calculating $Cov(D',D)$ and $Red(D')$ are $O(m \times n)$ and $O(m^2)$, respectively. Because typically $m \ll n$, the computational complexity of the calculations is low. To further illustrate the efficiency of the proposed calculation, a scalability experiment has also been conducted. Note that the calculation of information coverage, including assignment and other operations, is complicated, and n is usually large, whereas the calculation of information redundancy is straightforward and rapid because m is normally small. Hence, the scalability is only relevant to the information coverage calculation. The experimental environment is a PC with a quad-core CPU 2.50 GHz and 3.96 Gb RAM and running Microsoft Windows Server 2003 Standard Edition, and the code is written in the Java language.

In the scalability experiment, the synthetic data were generated using real online consumer reviews collected from Taobao.com (www.taoboa.com), the leading online shopping platform in China (Li et al. 2008a, You et al. 2011), because the online review function is one of the most important information search services on e-commerce platforms for which diversity is an important quality (Tsaparas et al. 2011). Normally, the size of D', i.e., m, is small, so we assume $m = 10$ because 10 is the most common number of search results presented on the first page, and vary the size of the original search results set D, i.e., n, from 1000 to 200,000, which is large enough to mimic real-world applications. To avoid possible biases caused by inherent characteristics of the data, for each n value, the average runtime for 1000 independent random

```
Inputs:
   $D = \{d_1, d_2, \ldots, d_n\}$, $D' = \{d'_1, d'_2, \ldots, d'_m\}$, $m << n$;
Outputs:
   $Cov(D', D)$, $Red(D')$;
1     Word segmentation for $D$ and $D'$;
2     Stemming for $D$ and $D'$;
3     Stop-word removal for $D$ and $D'$;
4     VSM model building for $D$ and $D'$;
5     Similarity calculation for the document vectors in $D$ and $D'$;
6     $D_j = \Phi$, $n^v_j = 0$ $(j = 0,1,\ldots m)$;
7     for $i = 1$ to $n$ do
8         $k = \arg\max_{j=1,2,\ldots,m}(sim(d'_j, d_i))$;
9         if there is more than one $k$ that satisfies the above condition then
10            randomly assign to $k$;
11        end if
12        $D_k = D_k \cup d_i$;
13        $n^v_k = n^v_k + sim(d'_k, d_i)$;
14    end for
15    $n^v = \sum_{j=1,2,\ldots,m} n^v_j$;
16    $Cov_C(D', D)$ = calculate$Cov_C(D', D)$;
17    $Cov_S(D', D)$ = calculate$Cov_S(D', D)$;
18    $Cov(D', D) = Cov_C(D', D) \times Cov_S(D', D)$;
19    $Red(D')$ = calculate$Red(D')$;
20    return $\{Cov(D', D), Red(D')\}$.
```

Fig. 1. The procedure for calculating $Cov(D',D)$ and $Red(D')$.

extractions, where for each extraction 10 reviews were randomly selected, is calculated and presented as the final runtime. The runtimes, illustrated in Fig. 2, show that first, the runtime is rapid, e.g., even in the case $n = 200{,}000$, the runtime is less than 0.5 s; second, the runtime curve is almost a linear function of n, which is consistent with the theoretical analysis. Thus, the algorithm is computationally efficient.

To further show the superiority of our proposed metrics, an illustrative example for the crisp case, when partial similarity is not considered, is presented to illustrate the merits of the proposed metrics in an easy-to-understand manner. Moreover, several major coverage and redundancy metrics presented in related work, denoted as Cov_1 (Pan et al. 2005), Cov_2 (Zhai et al. 2003), Cov_3 (Zhuang et al. 2008), Cov_4 (Zhuang et al. 2008), Red_1 (Carbonell and Goldstien 1998), Red_2 (Pan et al. 2005), and Red_3 (Zhuang et al. 2008), are also calculated for comparison.

Example 1. Let an original set contains 100 objects a, 200 objects b, 300 objects c, as well as 400 objects d, with similarity defined in a crisp manner, and assume $a \neq b \neq c \neq d$. For the following 4 extracted sets, Set 1: $(1a, 2b, 3c, 4d)$, Set 2: $(10a, 20b, 30c, 40d)$, Set 3: $(1a, 1b, 1c, 1d)$, and Set 4: $(4a, 3b, 2c, 1d)$, the corresponding metric values could be calculated as shown in Table 1.

The results in Table 1 have some interesting implications. Regarding information coverage, first, Sets 1 and 2 offer ideal extraction solutions, so not only all distinct objects have been extracted but also the information structure has been perfectly preserved. Cov_3 shows poor results, so it is not an appropriate metric. Second, though Sets 3 and 4 have the same content coverage as Sets 1 and 2, they do not preserve the structure as well; this shortcoming can be identified by our Cov metric but not the Cov_1, Cov_2 or Cov_4 metrics. Third, when comparing Set 1 with Set 3, we see there is a difference in information coverage, which can be distinguished using the Cov metric but not with Cov_1, Cov_2 or Cov_4. Fourth, compared with Cov_3, the proposed Cov metric is more easily interpreted as well. Considering information redundancy, first, there clearly exists different redundancy among the four extracted sets, but Red_2 fails to make a distinction among them. Second, though Set 2 is 10 times the size of Set 1, the redundancy of Set 2 could not be 100%, so Red_1 does not perform well. Third, compared with Red_1, Red_2 and Red_3, our Red metric is more compatible with intuition because the Red metric value reflects the percentage of duplicate objects whereas the others do not. Thus, the proposed diversity metrics, Cov and Red, outperform the other metrics.

Table 1
Metric values for the four sets in Example 1.

Set ID	Information coverage					Information redundancy			
	Cov	Cov_1	Cov_2	Cov_3	Cov_4	Red	Red_1	Red_2	Red_3
Set 1	1	1	1	0.55	1	0.60	0.90	0	0.22
Set 2	1	1	1	0.55	1	0.96	1	0	0.30
Set 3	0.92	1	1	0.46	1	0	0	0	0
Set 4	0.80	1	1	0.37	1	0.60	0.90	0	0.22

5. An experiment using real data

To further demonstrate the effectiveness of our proposed metrics for measuring information diversity, an experiment using real data from online customer reviews was conducted. When searching online product reviews, a customer may wish to know the diversity of the existing customers' reviews before making a purchase decision. Typically, the number of customer reviews available is in the hundreds, and sometimes greater. Thus, there is usually very high redundancy, so it will be very useful to present the customer with a small set of appropriately diverse reviews characterized by high information coverage and low information redundancy (Tsaparas et al. 2011). Nevertheless, the most popular strategy is to present customer reviews chronologically, from newest to oldest, which we refer to as the Newest Strategy. Obviously, the Newest Strategy cannot guarantee high diversity. Another straightforward strategy to target diversity is to randomly select a small subset of all reviews; we refer to this as the Random Strategy. Its effectiveness at maintaining diversity also needs to be demonstrated.

Clustering methods are somewhat effective at extracting a small and diverse set (Han and Kamber 2006, Liu 2007). Clustering is an unsupervised grouping of a given 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. 2009, Grabmeier and Rudolph 2002, Jain et al. 1999). Thus, using the original set of search results and given a clustering method, m clusters or sets could be defined. In each set, a search result with the maximum similarity to other results in the cluster, referred to as the centroid, could be extracted as a representative result because the result could be regarded as covering the maximum information load of all the results in the cluster. Moreover, because the m results are from m clusters and the similarities among different clusters are generally low, the m results will be mutually less redundant. Therefore, the final set of

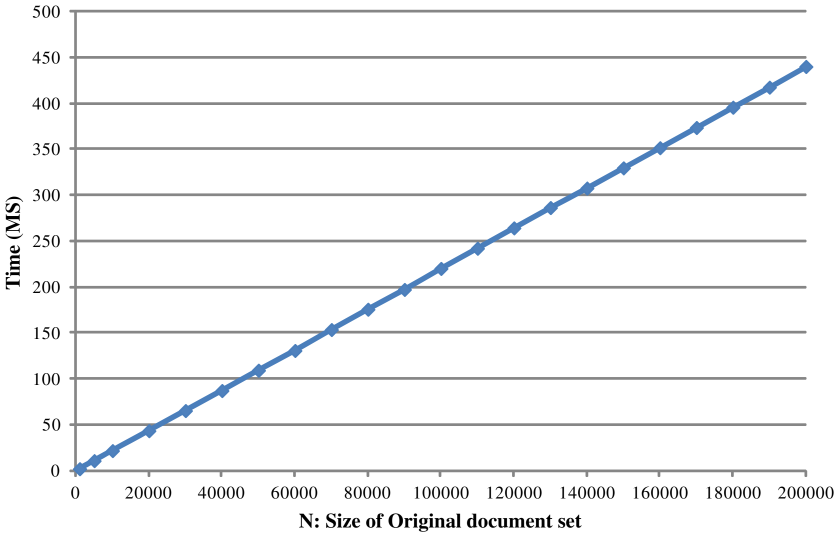

Fig. 2. Runtimes of $Cov(D', D)$ for $m = 10$.

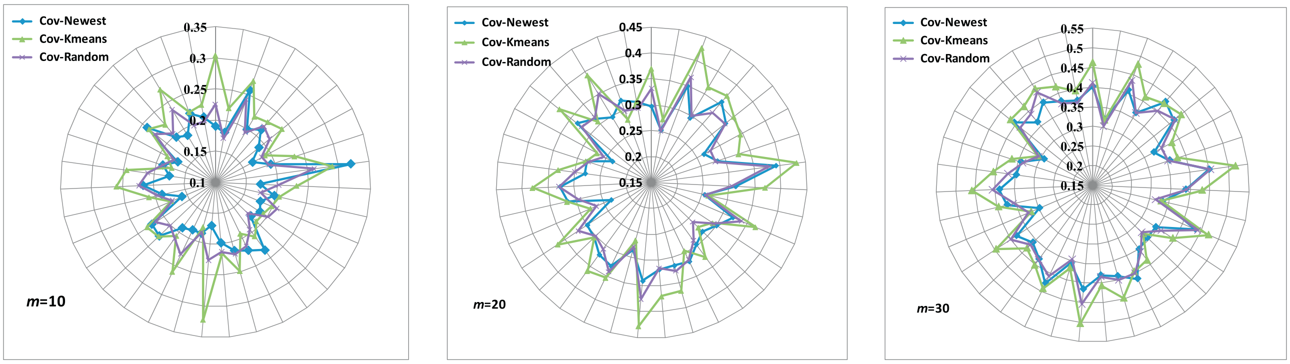

Fig. 3. Information coverage for three strategies applied to reviews of 35 products (m = 10, 20 and 30).

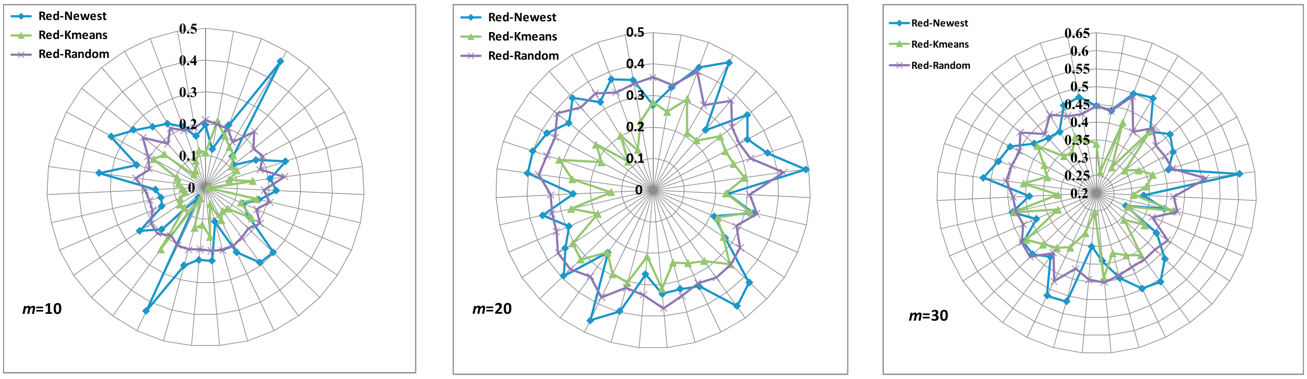

Fig. 4. Information redundancy for three strategies applied to reviews of 35 products (m = 10, 20 and 30).

Table 2
Statistical results for information coverage and information redundancy (m = 10, 20 and 30).

m		Information coverage			Information redundancy		
		Newest	K-means	Random	Newest	K-means	Random
10	Mean	0.2005	0.2269	0.2056	0.2177	0.1144	0.2005
	Stdev	0.0306	0.0359	0.0210	0.0876	0.0499	0.0189
20	Mean	0.3050	0.3370	0.3079	0.3494	0.2502	0.3454
	Stdev	0.0330	0.0488	0.0331	0.0704	0.0549	0.0272
30	Mean	0.3755	0.4083	0.3789	0.4446	0.3587	0.4381
	Stdev	0.0384	0.0509	0.0400	0.0611	0.0522	0.0256

m representative results from the m clusters could be obtained; this set not only covers the information of the original set well but also has low redundancy, so it may show high diversity when evaluated with the proposed metric.

There exist many types of clustering methods, including K-means based clustering methods, graph-based clustering methods, and agglomerative based clustering methods, among others (Aliguliyev 2009a, Carpineto et al. 2009, Grabmeier and Rudolph 2002, Han and Kamber 2006, Hastie et al. 2009, Jain et al. 1999, Liu 2007). In the following discussion, a well-known K-means clustering method, called the K-means Strategy (Han and Kamber 2006, Jain and Dubes 1988, MacQueen 1967), is used to extract a small set. The three strategies, Newest, Random and K-means, will be compared in the experiment using real data. To avert possible biases caused by random sampling, the result of the Random Strategy is the average value of 50 independent samples.

The real customer reviews data were collected from Taobao.com. Based on the categories provided by Taobao.com, 35 products, including sports and outdoors (6 products, 1650 reviews), clothes and shoes (8 products, 2009 reviews), electronics and computers (8 products, 2431 reviews), home, garden and tools (8 products, 2023 reviews) as well as health and beauty (5 products, 1284 reviews), were randomly selected. In total, 9397 reviews were collected; the maximum number of reviews for a product is 481, the minimum number of reviews for a product is 100, and the average number of reviews for a product is 268. The reviews were ranked in descending order using posting time.

In the experiment, considering that the size of the set of reviews a customer may browse is normally small, we set m = 10, 20 and 30. For the Newest Strategy, the newest m reviews were extracted; for the Random Strategy, m reviews were randomly extracted in each sampling; and, for the K-means Strategy, m clusters were grouped, and for each cluster a representative centroid was selected to form an extracted set.

The reviews were first pre-processed with an open-source Chinese word segmentation package called "Paoding's Knives" (code.google.com/p/paoding), and the corresponding VSM and similarity calculation was conducted using Java and Apache Lucene (lucene.apache.org/). Moreover, the well-known CLUTO software package (glaros.dtc.umn.edu/gkhome/views/cluto) was selected to perform the K-means clustering (Chen et al. 2010, Li et al. 2008b, Malik et al. 2010, Zhao and Karypis 2004, 2005, Zhong and Ghosh 2005). The experimental environment is the same as that in the scalability experiment.

Figs. 3 and 4 show the information coverage and information redundancy for the three strategies applied to the reviews of 35 products. Tables 2 and 3 show the statistical results.

The results presented in Figs. 3 and 4 as well as Tables 2 and 3 have some important implications. First, the K-means strategy outperforms the other two strategies both in terms of information coverage and information redundancy – it has significantly higher

Table 3
Paired *t*-test results for information coverage and information redundancy (m = 10, 20 and 30).

m	Metric	Hypothesis	*t*-Test value	*p*-Value
10	Coverage	*K*-means > Newest	4.073	2.627E−4***
		K-means > Random	5.26	7.894E−6***
		Random > Newest	1.223	0.230
	Redundancy	*K*-means < Newest	5.578	3.048E−6***
		K-means < Random	10.362	4.653E−12***
		Random < Newest	1.138	0.263
20	Coverage	*K*-means > Newest	5.578	3.042E−6***
		K-means > Random	6.839	7.151E−8***
		Random > Newest	0.945	0.352
	Redundancy	*K*-means < Newest	7.945	2.955E−9***
		K-means < Random	10.388	4.357E−12***
		Random < Newest	0.368	0.715
30	Coverage	*K*-means > Newest	7.449	1.215E−8***
		K-means > Random	10.336	4.976E−12***
		Random > Newest	1.159	0.255
	Redundancy	*K*-means < Newest	7.093	3.409E−8***
		K-means < Random	9.635	2.999E−11***
		Random < Newest	0.730	0.470

*** $p < 0.001$.

information coverage and lower information redundancy. This implies that the *K*-means clustering method could effectively extract diverse reviews, which is consistent with previous analysis. Second, the Newest Strategy performed poorly in terms of both information coverage and information redundancy, which also indicates that, though the Newest Strategy can provide fresh reviews, it is insufficient at maintaining diversity, which is a key goal for this category of information service. Third, the Random Strategy also performed poorly. Though intuitively it should be better at maintaining diversity because of the random sampling procedure, there is actually no significant difference between the Random and Newest Strategies, which indicates that diversity could not be easily maintained using the Random Strategy. Furthermore, as m is increased, both of the two metrics for the three strategies increase because a larger small set can cover more information load and generate more redundancy; these results are also consistent with the theoretical analysis. Thus, the proposed metrics could effectively measure the differences in diversity among the different strategies.

6. Conclusion

Because of information overload on e-commerce platforms, extracting a small set of search or recommendation results that possesses high diversity (high information coverage and low information redundancy), is very helpful to both information search service providers and customers. We proposed a novel metric setting for measuring information coverage in terms of two important aspects, content coverage and structure coverage, and information redundancy. The proposed metric setting for quantifying diversity has the following merits: First, the degree to which the information structure is preserved is considered and modeled using information entropy as part of the information coverage metric; this approach can appropriately measure the information distribution of the extracted set with respect to the original set. Second, the calculation of the metrics can be implemented easily, effectively, and efficiently and does not require human intervention. Third, the proposed metrics can be interpreted intuitively and also possess better properties than existing metrics; the superior performance has been demonstrated by both a scalability experiment and an illustrative example. Finally, an experiment using customer reviews data from a real online shopping platform further supports the effectiveness of the proposed metrics.

Our future work will focus on two aspects. First, using the proposed metrics, an efficient heuristic method to extract results with high diversity will be designed. Second, further empirical validation will be conducted using Western e-commerce platforms to investigate the potential impact of cultural differences on the results.

Acknowledgments

The work was partially supported by the National Natural Science Foundation of China (70890083/71072015/71110107027) and the Tsinghua University Initiative Scientific Research Program (20101081741).

References

Adomavicius, G., and Tuzhilin, A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17, 6, 2005, 734–749.

Agichtein, E., Brill, E., and Dumais, S. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, Seattle, Washington, USA, ACM, 2006a, 19–26.

Agichtein, E., Brill, E., Dumais, S., and Ragno, R. Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA, ACM, 2006b, 3–10.

Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, Barcelona, Spain, ACM, 2009, 5–14.

Aliguliyev, R. M. Clustering of document collection – a weighting approach. *Expert Systems with Applications*, 36, 4, 2009, 7904–7916.

Aliguliyev, R. M. Performance evaluation of density-based clustering methods. *Information Sciences*, 179, 20, 2009, 3583–3602.

Allan, J., and Raghavan, H. Using part-of-speech patterns to reduce query ambiguity. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland, 2002, ACM, 307–314.

Atlam, E.-S., Fuketa, M., Morita, K., and Aoe, J.-i. Similarity measurement using term negative weight and its application to word similarity. *Information Processing & Management*, 36, 2000, 717–736. 5.

Baeza-Yates, R., and Ribeiro-Neto, B. *Modern Information Retrieval*. Addison Wesley Longman Publishing Co. Inc., New York, NY, 1999.

Bernstein, Y., and Zobel, J. Redundant documents and search effectiveness. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, Bremen, Germany, ACM, 2005, 736–743.

Boldi, P., Santini, M., and Vigna, S. PageRank: functional dependencies. *ACM Transactions on Information Systems*, 27, 4, 2009, 1–23.

Boyce, B. Beyond topicality: a two stage view of relevance and the retrieval process. *Information Processing & Management*, 18, 3, 1982, 105–109.

Broder, A. A taxonomy of web search. *SIGIR Forum*, 36, 2, 2002, 3–10.

Brynjolfsson, E., Hu, Y., and Smith, M. D. Consumer surplus in the digital economy: estimating the value of increased product variety at online booksellers. *Management Science*, 49, 11, 2003, 1580–1596.

Buckley, C., and Voorhees, E. M. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece, ACM, 2000, 33–40.

Buckley, C., and Voorhees, E. M. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, United Kingdom, ACM, 2004, 25–32.

Carbonell, J., and Goldstein, J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, ACM, 1998, 335–336.

Carpineto, C., Osinski, S., Romano, G., and Weiss, D. A survey of Web clustering engines. *ACM Computing Surveys*, 41, 3, 2009, 1–38.

Carterette, B., and Chandar, P. Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management*, Hong Kong, China, ACM, 2009, 1287–1296.

Chehreghani, M. H., Abolhassani, H., and Chehreghani, M. H. Density link-based methods for clustering web pages. *Decision Support Systems*, 47, 4, 2009, 374–382.

Chen, C.-L., Tseng, F. S. C., and Liang, T. An integration of WordNet and fuzzy association rule mining for multi-label document clustering. *Data & Knowledge Engineering*, 69, 11, 2010, 1208–1226.

Chen, H., and Karger, D. R. Less is more: Probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA, ACM, 2006, 429–436.

Chen, Y.-L., Kuo, M.-H., Wu, S.-Y., and Tang, K. Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. *Electronic Commerce Research and Applications*, 8, 5, 2009, 241–251.

Clarke, C. L. A., Craswell, N., and Soboroff, I. Overview of the TREC 2009 Web Track. Paper presented at the 18th Text REtrieval conference (TREC2009), Gaithersburg, MD, 2010.

Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Singapore, Singapore, ACM, 2008, 659–666.

De, P., Hu, Y., and Rahman, M. S. Technology usage and online sales: an empirical study. *Management Science*, 56, 11, 2010, 1930–1945.

Duan, W., Gu, B., and Whinston, A. B. Do online reviews matter? – an empirical investigation of panel data. *Decision Support Systems*, 45, 4, 2008, 1007–1016.

Dupret, G., Murdock, V., and Piwowarski, B. Web search engine evaluation using click-through data and a user model. In *Proceedings of the Workshop on Query Log Analysis (WWW)*, Banff, Canada, 2007.

Fleder, D., and Hosanagar, K. Blockbuster culture's next rise or fall: the impact of recommender systems on sales diversity. *Management Science*, 55, 5, 2009, 697–712.

Fox, S., Karnawat, K., Mydland, M., Dumais, S., and White, T. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23, 2, 2005, 147–168.

Ghose, A., and Ipeirotis, P. G. Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23, 10, 2011, 1498–1512.

Goffman, W. A searching procedure for information retrieval. *Information Storage and Retrieval*, 2, 2, 1964, 73–78.

Gollapudi, S., and Sharma, A. An axiomatic approach for result diversification. In *Proceedings of the 18th International Conference on World Wide Web*, Madrid, Spain, ACM, 2009, 381–390.

Grabmeier, J., and Rudolph, A. Techniques of cluster algorithms in data mining. *Data Mining and Knowledge Discovery*, 6, 4, 2002, 303–360.

Han, J., and Kamber, M. *Data Mining: Concepts and Techniques*, 2nd edition. Morgan Kaufmann Publishers, San Francisco, CA, 2006.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition. Springer-Verlag, New York, NY, 2009.

He, J., Meij, E., and Rijke, M. d. Result diversification based on query-specific cluster ranking. *Journal of the American Society for Information Science and Technology*, 62, 2011, 550–571. 3.

Herlocker, J. L., Konstan, J. A., Borchers, A., and Riedl, J. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, California, United States, ACM, 1999, 230–237.

Huang, A. Similarity measures for text document clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC 2008)*, Christchurch, New Zealand, 2008, 49–56.

Hurley, N., and Zhang, M. Novelty and diversity in top-N recommendation – analysis and evaluation. *ACM Transactions on Internet Technology*, 10, 4, 2011, 1–30.

Järvelin, K., and Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20, 4, 2002, 422–446.

Jain, A. K., and Dubes, R. C. *Algorithms for Clustering Data*. Prentice Hall, 1988.

Jain, A. K., Murty, M. N., and Flynn, P. J. Data clustering: a review. *ACM Computing Surveys*, 31, 3, 1999, 264–323.

Jie, Z., Xiao, F., and Liu Sheng, O. R. Online consumer search depth: theories and new findings. *Journal of Management Information Systems*, 23, 2006, 71–95. 3.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, ACM, 2005, 154–161.

Koh, N. S., Hu, N., and Clemons, E. K. Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications*, 9, 5, 2010, 374–385.

Korenius, T., Laurikkala, J., and Juhola, M. On principal component analysis, cosine and Euclidean measures in information retrieval. *Information Sciences*, 177, 22, 2007, 4893–4905.

Kraft, D. H., and Bookstein, A. Evaluation of information retrieval systems: a decision theory approach. *Journal of the American Society for Information Science*, 29, 1, 1978, 31–40.

Kumar, N., and Lang, K. R. Do search terms matter for online consumers? The interplay between search engine query specification and topical organization. *Decision Support Systems*, 44, 1, 2007, 159–174.

Kuruzovich, J., Viswanathan, S., and Agarwal, R. Seller search and market outcomes in online auctions. *Management Science*, 56, 10, 2010, 1702–1717.

Lathia, N., Hailes, S., Capra, L., and Amatriain, X. Temporal diversity in recommender systems. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Geneva, Switzerland, ACM, 2010, 210–217.

Li, D., Li, J., and Lin, Z. Online consumer-to-consumer market in China – a comparative study of Taobao and eBay. *Electronic Commerce Research and Applications*, 7, 1, 2008, 55–67.

Li, Y., Chung, S. M., and Holt, J. D. Text document clustering based on frequent word meaning sequences. *Data & Knowledge Engineering*, 64, 1, 2008, 381–404.

Liu, B. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, Berlin, Heidelberg, 2007.

MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, 281–297.

Malik, H., Kender, J., Fradkin, D., and Moerchen, F. Hierarchical document clustering using local patterns. *Data Mining and Knowledge Discovery*, 21, 1, 2010, 153–185.

Mudambi, S. M., and Schuff, D. What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly*, 34, 1, 2010, 185–200.

Mulpuru, S. *The State of Retailing Online 2008: Merchandising and Web Optimization Report*. Forrester Research, Cambridge, MA, 2008.

Pan, F., Wang, W., Tung, A. K. H., and Yang, J. Finding representative set from massive data. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, IEEE Computer Society, 2005, 338–345.

Radlinski, F., and Dumais, S. Improving personalized web search using result diversification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA, ACM, 2006, 691–692.

Radlinski, F., Kleinberg, R., and Joachims, T. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, ACM, 2008, 784–791.

Rafiei, D., Bharat, K., and Shukla, A. Diversifying web search results. In *Proceedings of the 19th International Conference on World Wide Web*, Raleigh, North Carolina, USA, ACM, 2010, 781–790.

Robertson, S. E. The probability ranking principle in IR. *Journal of Documentation*, 33, 4, 1977, 294–304.

Rose, D. E., and Levinson, D. Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web*, New York, NY, USA, ACM, 2004, 13–19.

Salton, G. *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice-Hall Inc., 1971.

Salton, G., Wong, A., and Yang, C. S. A vector space model for automatic indexing. *Communication of the ACM*, 18, 11, 1975, 613–620.

Santos, R., Macdonald, C., and Ounis, L. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web*, Raleigh, North Carolina, USA, ACM, 2010a, 881–890.

Santos, R. L. T., Macdonald, C., and Ounis, I. Selectively diversifying web search results. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, Toronto, ON, Canada, ACM, 2010b, 1179–1188.

Santos, R. L. T., Macdonald, C., and Ounis, I. On the suitability of diversity metrics for learning-to-rank for diversity. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Beijing, China, ACM, 2011, 1185–1186.

Sen, R., King, R. C., and Shaw, M. J. Buyers' choice of online search strategy and its managerial implications. *Journal of Management Information Systems*, 23, 1, 2006, 211–238.

Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal*, 27, 1948, 379–423. 623–656.

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

Siwicki, B. The internet retailer hot 100 retail web sites. Available at http://www.internetretailer.com/article.asp?id=24575, 2007.

Spärck-Jones, K., Robertson, S. E., and Sanderson, M. Ambiguous requests: implications for retrieval tests, systems and theories. *SIGIR Forum*, 41, 2, 2007, 8–17.

Spink, A., Jansen, B. J., Wolfram, D., and Saracevic, T. From e-sex to e-commerce. Web search changes. *IEEE Computer*, 35, 3, 2002, 107–109.

Spink, A., Ozmutlu, S., Ozmutlu, H. C., and Jansen, B. J. U.S. versus European web searching trends. *SIGIR Forum*, 36, 2, 2002, 32–38.

Spink, A., Wolfram, D., Jansen, M. B. J., and Saracevic, T. Searching the web: the public and their queries. *Journal of the American Society for Information Science and Technology*, 52, 3, 2001, 226–234.

Tsaparas, P., Ntoulas, A., and Terzi, E. Selecting a comprehensive set of reviews. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, California, USA, ACM, 2011, 168–176.

Turban, E., Lee, J. K., King, D., Liang, T. P., and Turban, D. *Electronic Commerce: A Managerial Perspective*, 6th edition. Prentice Hall, 2010.

Vargas, S., Castells, P., and Vallet, D. Intent-oriented diversity in recommender systems. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Beijing, China, 2011, ACM, 1211–1212.

Vee, E., Srivastava, U., Shanmugasundaram, J., Bhat, P., and Yahia, S. A. Efficient computation of diverse query results. In *Proceedings of IEEE 24th International Conference on Data Engineering (ICDE 2008)*, 2008, 228–236.

Voorhees, E. M. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, ACM, 1998, 315–323.

Vuylsteke, A., Wen, Z., Baesens, B., and Poelmans, J. Consumers' search for information on the internet: how and why China differs from Western Europe. *Journal of Interactive Marketing*, 24, 4, 2010, 309–331.

Wang, J., and Zhu, J. Portfolio theory of information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Boston, MA, USA, ACM, 2009, 115–122.

Xu, Y., and Yin, H. Novelty and topicality in interactive information retrieval. *Journal of the American Society for Information Science and Technology*, 59, 2, 2008, 201–215.

You, W., Liu, L., Xia, M., and Lv, C. Reputation inflation detection in a Chinese C2C market. *Electronic Commerce Research and Applications*, 10, 5, 2011, 510–519.

Yue, Y., and Joachims, T. Predicting diverse subsets using structural SVMs. In *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, ACM, 2008, 1224–1231.

Zhai, C., and Lafferty, J. A risk minimization framework for information retrieval. *Information Processing & Management*, 42, 1, 2006, 31–55.

Zhai, C. X., Cohen, W. W., and Lafferty, J. Beyond independent relevance. methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, ACM, 2003, 10–17.

Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., Chen, Z., and Ma, W.-Y. Improving web search results using affinity graph. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, ACM, 2005, 504–511.

Zhang, M., and Hurley, N. Avoiding monotony: improving the diversity of recommendation lists. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, Lausanne, Switzerland, ACM, 2008, 123–130.

Zhang, Y., Callan, J., and Minka, T. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland, ACM, 2002, 81–88.

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

Zhao, Y., and Karypis, G. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10, 2, 2005, 141–168.

Zhong, S., and Ghosh, J. Generative model-based document clustering: a comparative study. *Knowledge and Information Systems*, 8, 3, 2005, 374–384.

Zhuang, J., Hoi, S. C. H., and Sun, A.. *On profiling blogs with representative entries. Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*. ACM, Singapore, 2008. 55–62.