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