Treating Each Intent Equally? The Equilibrium of IA-Select

Yingying Wu
Department of Mathematics
The University of Texas at Austin
yw@math.utexas.edu

Yiqun Liu
DCST, Tsinghua University
Beijing, China
yiqunliu@tsinghua.edu.cn

Ke Zhou
University of Nottingham
Nottingham, U.K.
ke.zhou@nottingham.ac.uk

Xiaochuan Wang
DCST, Tsinghua University
Beijing, China
wxc@sogou-inc.com

Min Zhang
DCST, Tsinghua University
Beijing, China
z-m@tsinghua.edu.cn

Shaoping Ma
DCST, Tsinghua University
Beijing, China
msp@mail.tsinghua.edu.cn

ABSTRACT

Diversifying search results to satisfy as many users' intentions as possible is NP-hard. Some research employs a pruned exhaustive search, and some uses a greedy approach. However, the objective function of the result diversification problem adopts the cascade assumption which assumes users' information needs will drop once their subtopic search intents are satisfied. As a result, the intent distribution of diversified results deviates from the actual distribution of user intentions, and each subtopic tends to be chosen equally. This phenomenon is unreasonable, especially when the original distribution of user intent is unbalanced. In this paper, we present empirical evidence of the diversification equilibrium by showing that the standard deviations of subtopic distribution approach zero.

KEYWORDS

Diversified search; subtopic retrieval; heuristic algorithms

ACM Reference Format:

Yingying Wu, Yiqun Liu, Ke Zhou, Xiaochuan Wang, Min Zhang, and Shaoping Ma. 2018. Treating Each Intent Equally? The Equilibrium of IA-Select. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23-27, 2018, Lyon, France*, Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3184558.3186955>

1 INTRODUCTION

Every query can be considered ambiguous due to the inherently limited representation of generically complex information needs. For example, by conducting a user study with 50 participants recruited via social networks to survey search intents for a specific query “apple,” we find that a breakdown of the potential intents is as follows: 19 users intend to search for “devices from Apple” (38%), 17 users intend to search for a “fruit apple” (30%), 12 users intend to search for “creative arts” (24%), three users intend to search for “Apple facility locations” (6%), and only one user intends to search for “Apple services” (2%), as shown in Fig. 1(a). Therefore, search result diversification has been proposed to produce rankings to satisfy the multiple potential information needs. The result

diversification problem is a reduction from the maximum coverage problem and is hence NP-hard. Chen uses an exhaustive pruning strategy to obtain the optimal diversification [3]. The objective function of the diversification problem is submodular, which allows a $(1 - 1/e)$ -approximation [5]; therefore, a greedy approximation called IA-Select is proposed [1]. The algorithm attempts to maximize the probability that a user finds at least one useful result within k results. However, the underlying cascade assumption asserts that the gains from results reduce once a satisfying result from the same category is selected. This assumption tends to over-penalize results in chosen categories; as a consequence, IA-Select switches between categories to maximize the probability of users with different intentions finding at least one relevant document. Thus, the ratio of results selected from each intent converges quickly. Capannini [2] directly addresses the distribution of subtopics, but does not consider the sub-intent weights in the diversified ranking.

2 EQUILIBRIUM OF IA-SELECT ALGORITHM

In this section, we present challenges faced by the result diversification problem. We begin by an exposition of the IA-Select algorithm [1]. We adopt canonical assumptions of diversification problems: 1) the existence of a taxonomy of information, 2) user intents, queries, and retrieved results categorized according to this taxonomy, 3) completed a priori categories belonging to a query, and 4) independent probabilities of results satisfying users, which implies that the probability that the set of results fails to satisfy a general user is equal to the product of the probability of each result in the set that fails to satisfy. Given a query q , a collection of user intents $C = \{c_1, \dots, c_n\}$, the distribution of each intent $U(c_i)$, the quality value of each result d of user intent c_i is denoted $V(d|c_i)$, and the set of retrieved results $\mathcal{D} = \{D_1, \dots, D_n\}$, where D_i is the list of initial results with respect to intent c_i , the objective of the diversification problem is to maximize the probability that an average user finds at least one satisfying result within the top k results. That is, we seek a set of results $S^k \subseteq \mathcal{D}$ that maximizes

$$\mathcal{G}(S^k) = \sum_{c_i \in C} U(c_i) \left(1 - \prod_{d \in S^k} (1 - V(d|c_i)) \right). \quad (1)$$

Denote $U(c_i|S^k)$ as the probability that the query q belongs to intent c_i , given that all documents S^k fail to satisfy a user, and $C(d) \subset C$ as the set of intents a document d belongs to. IA-Select

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '18 Companion, April 23-27, 2018, Lyon, France

© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.

ACM ISBN 978-1-4503-5640-4/18/04.

<https://doi.org/10.1145/3184558.3186955>

Figure 1: (a) Intent survey questionnaire and results. The ratio of images retrieved by IA-Select with quality values being (b) $1/r$ and (c) $e^{-r/N}$.

iteratively chooses one result d^* with the highest marginal utility

$$d^* = \operatorname{argmax}_{d \in \mathcal{D}} \sum_{c_i \in C(d)} U(c_i | S^k) V(d | c_i). \quad (2)$$

Initially, $S^0 = \emptyset$, and $U(c_i | \emptyset)$ is the surveyed intent distribution; for all $c_i \in C$, IA-Select updates $S^{k+1} = S^k \cup \{d^*\}$ and $U(c_i | S^k)$ by:

$$U(c_i | S^{k+1}) = \begin{cases} (1 - V(d^* | c_i)) U(c_i | S^k) & \text{if } c_i \in C(d^*) \\ U(c_i | S^k) & \text{otherwise} \end{cases}. \quad (3)$$

The IA-Select algorithm reduces the value of a sub-intent by a factor of $1 - V(d^* | c_i)$ every time a result is selected from this intent. Hence, the subtopic distribution of chosen results tends to converge to equilibrium, where each subtopic tends to be chosen an equal number of times. To empirically exhibit this scenario, we use intent data for the query "apple" from a user study whose questionnaire and response distribution are presented in Fig. 1(a). Furthermore, the quality value for a document d at position r_d^i in user intent c_i is simulated by $V_1(d | c_i) = \frac{1}{r_d^i}$ and $V_2(d | c_i) = e^{-r_d^i/50}$.

The distribution of subtopics as a function of the number of retrieved results is reported in Fig. 1, where the dashed lines represent the sub-intents collected from the user study. The dots in the figures represent the percentage of each subtopic when retrieving N search results for N from 1 to 200. The statistics suggest that as the number of results increases, the proportion of results in each sub-intent selected by the IA-Select algorithm tends to be equal. The average standard deviation of subtopic distribution $\mathcal{P}^N$ approaching zero implies the diversification equilibrium. Consider the set $\mathcal{P}^N = \{q_1^N, \dots, q_n^N\}$ where q_i^N is the proportion of results in user intent c_i in N retrieved results. Having the standard deviation of subtopic distribution approach zero is a necessary and sufficient condition for the diversification equilibrium.

Consider the subtopic distribution $\{0.38, 0.3, 0.24, 0.06, 0.02\}$ collected from the user survey, which admits a standard deviation of 0.139. Table 1 presents the mean standard deviation of $\mathcal{P}^N$ for N in an interval $[a, b]$ with quality scores listed in the previous section. The averaged standard deviations decrease over time as more results are retrieved. Such a tendency suggests that the ratio of results in each intent tends to be the same as more results are retrieved; therefore, the intent distribution deviates from the ground truth intent distribution.

Table 1: Average standard deviations of the distribution of subtopics in retrieved results (quality scores are $1/r$ and $e^{-r/N}$).

σ	[1, 10]	[11, 20]	[21, 30]	[31, 40]	[41, 50]	[91, 100]	[191, 200]
$1/r$	0.189	0.121	0.111	0.102	0.098	0.094	0.092
$e^{-r/N}$	0.115	0.032	0.023	0.018	0.015	0.011	0.009

3 CONCLUSION

Result diversification is an important research area and deserves much attention [1], but there is some existing work [4, 6] that shows diversified ranking lists are not always reasonable in practical search scenarios. In this study, we report an observation on the diversification equilibrium, which suggests a deviation of subtopic distribution of diversified results from the ground truth intent distribution. We conduct a user study to examine the subtopic distribution of retrieved results by IA-Select algorithm. Fig. 1(b)-(c) and Table 1 show that the number of results selected from each sub-intent tends to converge to an equilibrium point, which can be a potential pitfall for retrieval systems. This observation promotes research attention to the optimization of objective function, which potentially boosts the performances of existing result diversification algorithms.

4 ACKNOWLEDGEMENTS

The authors thank Andrew Blumberg and François Baccelli for discussions on the equilibrium convergence. This work is supported by the Natural Science Foundation of China (Grant No. 61622208, 61732008, 61532011) and the National Key Basic Research Program (2015CB358700). Wu was supported in part by NIH grant 5U54CA193313.

REFERENCES

- [1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Jeong. 2009. Diversifying search results. In *WSDM'09*. ACM, 5–14.
- [2] Gabriele Capannini, Franco Maria Nardini, Raffaele Perego, and Fabrizio Silvestri. 2011. Efficient diversification of web search results. *Proceedings of the VLDB Endowment* 4, 7 (2011), 451–459.
- [3] Fei Chen, Yiqun Liu, Jian Li, Min Zhang, and Shaoping Ma. 2014. A Pruning Algorithm for Optimal Diversified Search. In *WWW'14*. ACM, 237–238.
- [4] Harr Chen and David R Karger. 2006. Less is more: Probabilistic models for retrieving fewer relevant documents. In *SIGIR'06*. ACM, 429–436.
- [5] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functions - I. *Mathematical Programming* 14, 1 (1978), 265–294.
- [6] Rodrigo LT Santos, Craig Macdonald, and Iadh Ounis. 2012. On the role of novelty for search result diversification. *Information Retrieval* 15, 5 (2012), 478–502.