Treating Each Intent Equally? The Equilibrium of IA-Select

Yingying Wu
Department of Mathematics
The University of Texas at Austin
ywu@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
Sogou Incorporation
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' intents as possible is NP-hard. There have been a plethora of studies on the result diversification problem; some employ a pruned exhaustive search, and some use the greedy algorithm. 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 of this assumption, the intent distribution of diversified results deviates from the actual distribution of user intentions until each subtopic is chosen equally. Such a selection is unreasonable, especially when the original distribution of user intent is unbalanced. In this paper, we prove that having the standard deviation of subtopic distribution approach zero is a necessary and sufficient condition for the diversification equilibrium and provides empirical evidence for the equilibrium.

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 *Proceedings of WWW'18,* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.475/123_4

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 intent for a specific query "apple," we find that a breakdown of the potential intent is as follows: 19 users intend to search for "devices from Apple" (38%), 17 users intend a"fruit apple" (30%), 12 users intend "creative arts" (24%), 3 users intend "Apple facility locations" (6%), and only 1 user intends to search for "Apple services" (2%), as shown in Fig. 1(a). Therefore, given the ambiguous information need, search result diversification has been proposed to produce rankings to satisfy the multiple potential information needs. The result diversification problem a reduction from the maximum coverage problem hence is NP-hard. Chen uses a pruning strategy to obtain the optimal diversification [3]; the pruned algorithm obtains optimal results, but it is not efficient enough. The submodularity of the objective function of the diversification problem allows a $(1-1/e)$-approximation [4]; hence 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 hence the ratio of results selected from each intent converges quickly. Capannini [2] directly addressed the distribution of subtopics in the diversified list, but does not consider the weight of each sub-intent in the diversified ranking.

WWW'18, April 2018, Lyon, France

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

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 assumptions that diversification problems canonically make: a taxonomy of information exists; user intent, queries, and retrieved results are categorized according to this taxonomy; the a priori categories belonging to a query are complete; and the conditional probabilities of the two results satisfying users are independent, which implies that the probability that the set of results all fail to satisfy is equal to the product of the probability of each result in the set fail to satisfy. Given a query q, a collection of user intents $C = \{c_1, \ldots, c_n\}$, 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 $\mathcal{D} = \{D_1, \ldots, D_n\}$, where D_i is the list of initial results with respect to intent c_i, the objective of diversification problem is to maximize the probability that an average user finds at least one satisfying result within the top k results. That is, to find 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 the user, and

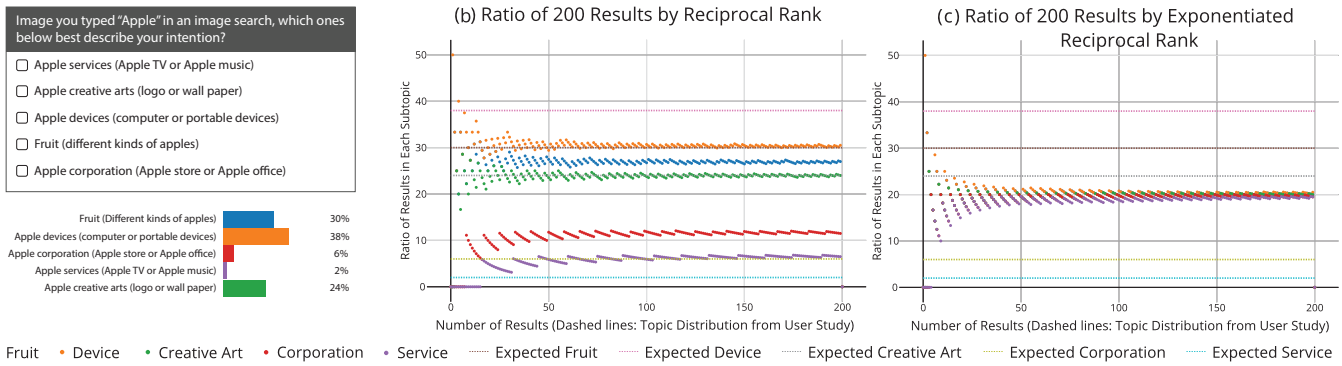

Figure 1: (a) Intent survey questionnaire and results. The ratio of images retrieved by IA-Select with quality values being (b) $1/r$ and (c) $e^{-r/N}$.

$C(d) \subset C$ is the set of intent a document d belongs to, IA-Select iteratively chooses one result d^* with the highest marginal utility

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

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

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

Due to the nature of IA-Select, the value of a sub-intent reduces 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 for an equal number of times. To empirically exhibit this scenario, we exploit 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 subtopic intention 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 $\mathcal{P}^N = \{q_1^N, \ldots, q_n^N\}$ where q_i^N is the proportion of results in user intent c_i in N retrieved results. The following proposition allows exhibiting the equilibrium in terms of the standard deviation.

Proposition 2.1. *$q_i^N = q_j^N$ for all $c_i, c_j \in C$ if and only if the standard deviation of $\mathcal{P}^N$ is zero.*

Proof. By the complete knowledge assumption, q_i^Ns sum up to 1. Since there are n subtopics, the mean value μ of $\mathcal{P}^N$ is $1/n$. If the standard deviation of $\mathcal{P}^N$ is zero, then

$$\frac{\sum_{i=0}^{n} \left(q_i^N - \mu\right)^2}{n} = 0 \Leftrightarrow \forall i,\ q_i^N = \mu. \tag{4}$$

Therefore, the ratio q_i^N for each intent is $1/n$. The other direction is obtained from the definition of a standard deviation. □

For example, the sub-intent distribution $\{0.38, 0.3, 0.24, 0.06, 0.02\}$ obtained from the user survey admits a standard deviation of 0.139. Table 1 presents the standard deviation of $\mathcal{P}^N$ averaged over $[a, b]$ with quality scores listed in the previous section. The averaged standard deviations decrease over time and approach zero 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 selected; hence 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

In this study, we report an observation on the diversification equilibrium, which suggests a deviation of subtopic distribution of diversified results from the actual user intention, hence can be a potential pitfall for result diversification problem. We conduct a user study to examine the subtopic distribution of retrieved result by IA-Select. Fig. 1(b)-(c) and Table 1 show that the number of results selected from each user intent tends to converge to an equilibrium point. This observation promotes research attention to the optimization of objective function, which potentially boosts the performances of existing result diversification algorithms in general.

4 ACKNOWLEDGEMENTS

The authors thank Andrew Blumberg and François Baccelli for discussions on the equilibrium convergence. This work is supported by Natural Science Foundation of China (Grant No. 61622208, 61732008, 61532011) and 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 Ieong. 2009. Diversifying search results. In *WSDM*. 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*. ACM, 237–238.
[4] 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.