Searching for Communities: a Facebook Way

Viet Ha-Thuc
vhathuc@fb.com
Facebook Inc.
Menlo Park, CA

Srinath Aaleti
srinatha@fb.com
Facebook Inc.
Menlo Park, CA

Rongda Zhu
rongdazhu@fb.com
Facebook Inc.
Menlo Park, CA

Nade Sritanyaratana
nade@fb.com
Facebook Inc.
Menlo Park, CA

Corey Chen
coreychen@fb.com
Facebook Inc.
Menlo Park, CA

ABSTRACT

Giving people the power to build community is central to Facebook's mission. Technically, searching for communities poses very different challenges compared to the standard IR problems. First, there is a vocabulary mismatch problem since most of the content of the communities is private. Second, the common labeling strategies based on human ratings and clicks do not work well due to limited public content available to third-party raters and users at search time. Finally, community search has a dual objective of satisfying searchers and growing the number of active communities. While A/B testing is a well known approach for assessing the former, it is an open question on how to measure progress on the latter. This talk discusses these challenges in depth and describes our solution.

KEYWORDS

Privacy, embeddings, causal effect, counterfactual, explainability

ACM Reference Format:
Viet Ha-Thuc, Srinath Aaleti, Rongda Zhu, Nade Sritanyaratana, and Corey Chen. 2019. Searching for Communities: a Facebook Way. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3331184.3331426

1 INTRODUCTION

Connecting people with communities has been one of the top priorities of Facebook in the past few years. Among different channels, search is a natural and effective way for users on Facebook to discover and connect with relevant communities. Figure 1 shows how users can use the global Search (left) and vertical Search (right) to find relevant groups (communities). In the former, different vertical search engines (e.g., groups, events, people and videos) retrieve and rank results within their verticals and then send the top results to a federation system which blends these results. The focus of this talk is on how to build such a vertical search for Facebook groups.

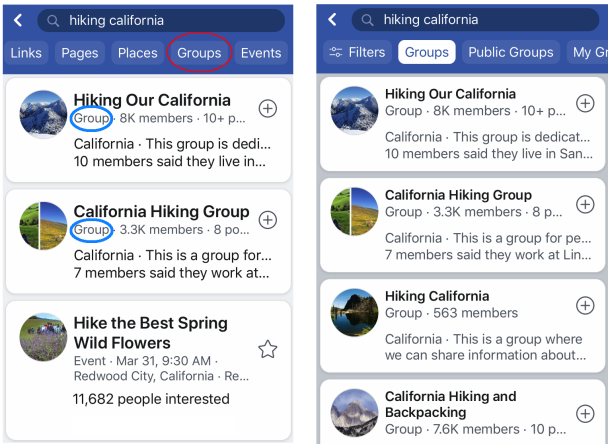

Figure 1: Users can discover groups on Facebook Global Search (left) or select Groups tab on the top-right corner to switch to Group Vertical Search (right)

Compared to traditional search engines, e.g., Web search, searching for communities has unique challenges. First, many groups on Facebook are closed groups, meaning that the group content cannot be used to match with non-member's queries. Second, it is challenging for a third-party rater to accurately judge groups due to their privacy, and user clicks have low correlation with user long-term satisfaction since only limited content is available at click time. Finally, as a part of a bigger ecosystem, the objective of groups search is not only searcher satisfaction but also community inventory growth on Facebook, which in turn will improve the overall user experience with more communities to connect with. While A/B testing is a well-known technique to measure the former, it is an open question of how to measure the causal impact of search improvements on the inventory (the document side).

2 PRIVACY AND VOCABULARY MISMATCH

Many communities on Facebook are closed groups. This means only public information, such as their titles, descriptions and admin tags like group topics are available to non-member searchers. The actual posts inside are not visible to non-members until they join the community; thus, post content cannot be used to match with queries at search time. This creates a vocabulary mismatch problem between queries and the limited text on the document side.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331426

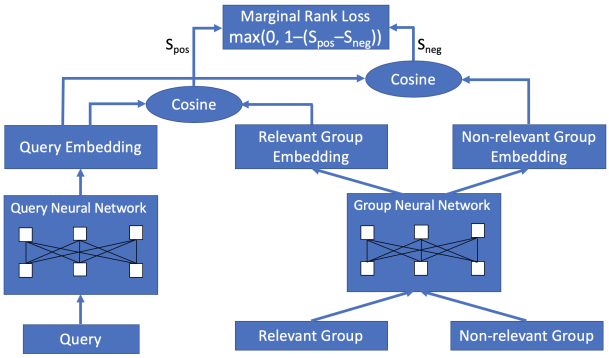

Figure 2: Learning Group and Term Embeddings

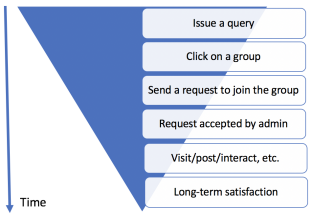

Figure 3: Engagement Funnel

To alleviate this, we learn an *embedding* [2] for each group from its public information and embeddings for the terms in the vocabulary on the same space. To ensure fast retrieval at runtime, for each group we index the most semantically similar terms based on their embeddings. Figure 2 shows the system learning the group and query embeddings. Query embeddings are pooled from the embeddings of query terms and n-grams. For each query, relevant and non-relevant groups are inferred from click logs. The objective is to jointly learn all of the embeddings such that the similarity between a query and its relevant group is maximized and the similarity between a query and its non-relevant group is minimized.

3 OPTIMIZING LONG-TERM SATISFACTION

While inferring labeled data for the ranking function from clicks or human labels is a common approach in literature [1], it does not work well for community search. Given that many communities are closed, it is challenging for a third-party rater to accurately judge their relevance. At the same time, clicks have low correlation with long-term satisfaction also because of limited content visible to the searcher at the click time and a long funnel of user engagement (Figure 3) from searching and clicking on a result to long-term satisfaction. Long-term satisfaction is determined by activities and social interactions with the communities over a long time (i.e., months) and satisfaction surveys. Our goal is to learn a ranking function optimizing this long-term engagement.

Since long-term satisfaction takes a long time to determine, directly inferring labels from this causes practical issues. For instance, when we make a change on the retrieval stage (e.g., query expansion) which alters the retrieved document distribution, it takes

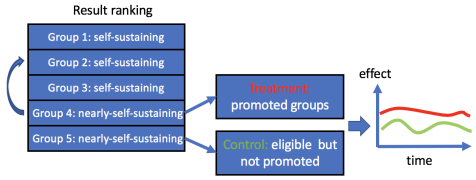

Figure 4: Counterfactual Framework to Evaluate

months to collect new training data to re-train the ranker. To overcome this, we learn a few indicators available *shortly after* a user joins a community (Step 4 on the funnel), e.g., time spent in the community over the first few days, that are highly correlated with long-term satisfaction. These indicators are then used in aggregate in the multi-objective function to optimize the ranker for.

4 CAUSAL EFFECTS ON DOCUMENTS

Our study indicates that social communities typically require a certain number of active members to stay active and grow over time by themselves. These communities are called *self-sustaining* groups. Unlike the traditional IR setting, our search system is a part of a bigger ecosystem, and its objective is not only to optimize searcher satisfaction but also to grow the number of self-sustaining groups. As a simple strategy, given a ranking optimized for searcher satisfaction, non-self-sustaining groups that are relevant to the query and just below the *self-sustaining threshold* are promoted into the top K. The idea is to give the *nearly* self-sustaining groups some extra exposure so that they will attract enough active members and grow by themselves. However, this leads to a challenge of how to evaluate the number self-sustaining groups that the strategy gains.

To evaluate the causal effect on the groups, we propose a counterfactual framework. Instead of promoting all eligible documents, we *deterministically* promote half of them, e.g., only groups with *even* ID. As demonstrated in Figure 4, both groups 4 and 5 are eligible for promotion, but only group 4 gets promoted. The even and odd promotion-eligible groups are logged in treatment and control, respectively. Since we use a deterministic selection strategy, no group can appear in both treatment and control, even though it might be eligible for promotion in many sessions. The difference in terms of the number of self-sustaining groups in treatment and control after an experiment period is completely attributable to the promotion. In a general setting, given a baseline and a new ranking function aiming to promote a target set of documents, the counterfactual framework can be used to measure the causal effect of the ranking change on the documents.

5 CONCLUSIONS

In this paper, we present the unique challenges of building Facebook community search engine. We also lay out our approach addressing these challenges while still respecting all existing privacy policies as well as discuss some ongoing efforts.

REFERENCES

[1] Viet Ha-Thuc and Shakti Sinha. 2016. Learning to Rank Personalized Search Results in Professional Networks. In *Proceedings of the 39th ACM SIGIR*. 461–462.
[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013).