Fast Query Evaluation for Ad Retrieval

Ye Chen
Microsoft Corporation
1065 La Avenida, Mountain View, CA 94043
yec@microsoft.com

Mitali Gupta
Microsoft Corporation
1065 La Avenida, Mountain View, CA 94043
mitalig@microsoft.com

Tak W. Yan
Microsoft Corporation
1065 La Avenida, Mountain View, CA 94043
takyan@microsoft.com

ABSTRACT

We describe a fast query evaluation method for ad document retrieval in online advertising, based upon the classic WAND algorithm. The key idea is to localize per-topic term upper bounds into homogeneous ad groups. Our approach is not only theoretically motivated by a topical mixture model; but empirically justified by the characteristics of the ad domain, that is, short and semantically focused documents with natural hierarchy. We report experimental results using artificial and real-world query-ad retrieval data, and show that the tighter-bound WAND outperforms the traditional approach by 35.4% reduction in number of full evaluations.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models, Selection process

General Terms

Algorithms, Experimentation, Performance

Keywords

Query evaluation, WAND, ad selection, ad relevance

1. INTRODUCTION

WAND is a fast algorithm for query evaluation in information retrieval (IR) tasks [1]. We consider an additive scoring function of a query-document pair,

$$s(q,d) = \sum_{t \in q \cap d} w(t,q) \times w(t,d). \quad (1)$$

For a vector space model in the $tf \cdot idf$ space, $w(t,q)$ is a normalized $tf \cdot idf$ weight of term t in query q: $w(t,q) = tf_{t,q} \times idf_t / ||q||$, and $w(t,d)$ is a normalized $tf \cdot idf$ weight of term t in document d: $w(t,d) = tf_{t,d} \times idf_t / ||d||$.

One key idea of WAND is to augment each term t in index with an upper bound on $w(t,d)$, i.e., the term's multiplicative contribution to any document,

$$u_t = \max_d (w(t,d)), \quad (2)$$

and then efficiently compute an upper bound on $s(q,d)$ by summing the above document-independent term upper bounds, weighted by $w(t,q)$,

$$u(q,d) = \sum_{t \in q \cap d} w(t,q) u_t \geq s(q,d). \quad (3)$$

The first-level retrieval then only passes documents with $u(q,d) \geq \theta$, a threshold, to the second-level full evaluation.

One problem with this approach is that a term upper bound might not be tight enough, resulting in a high false-positive error in the first-level retrieval. For the web search task, this problem may not be so severe, since documents are relatively long and exhibit many topics. Indeed, the tightness of a term upper bound reflects the trade-off between the first-level computational efficiency and the false-positive error. For the ad selection task, however, a global term upper bound may be far from optimal, given the following observations: (1) an ad document is typically short, and (2) contains very focused and few topics. One example would be a popular query like "iphone deal" may never appear in an often-seen "credit score" ad.

2. METHODOLOGY

We propose to associate with each term a small set of upper bounds, each derived from a homogeneous group of ads, e.g., from a same ad category, denoted as k,

$$u_{t,k} = \max_{d \in k} (w(t,d)), \quad (4)$$

and an upper bound on $s(q,d)$ for ad group k is then,

$$u_k(q,d) = \sum_{t \in q \cap d} w(t,q) u_{t,k} \geq s(q,d), \forall d \in k. \quad (5)$$

The rationale behind our approach is that $w(t,d)$ is generated from a mixture model, where each mixture component corresponds to a homogeneous group or topic k.

$$p(w(t,d)) = \sum_k p(k;\beta) p(w|k;\gamma_{t,k}). \quad (6)$$

The topical mixture weight $p(k;\beta)$ is typically modeled as a multinomial parameterized by β. Given a topic k, $w(t,d)$ follows a continuous distribution $p(w|k;\gamma_{t,k})$, parameterized by $\gamma_{t,k}$ dependent upon the term t and the topic k. Two sensible choices of the weight distribution $p(w|k;\gamma_{t,k})$ are uniform $\mathcal{U}(a,b)$ and Gaussian $\mathcal{N}(\mu,\sigma^2)$. The more homogeneous the ads within a same topic and the more heterogeneous the ads across different topics, the tighter (with a smaller variance) and the less overlapping (with widely spread means) the weight distributions of both choices would be. Therefore, given a sound model assumption, the derived local per-topic

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

term upper bounds would be significantly tighter than the global upper bounds.

On the other hand, learning latent topics from ads requires relatively expensive offline computation, yet an appealing approximation is to leverage the domain knowledge of hierarchical structure on ads, e.g., the ad category. Suppose that the topic k of an ad d is encoded into the ad id, deriving local upper bounds has only $\mathcal{O}(N)$ time complexity, same as the traditional WAND.

The traditional WAND maintains two critical invariants by using global term upper bounds: (1) any document with id $\leq$ the current id has already been considered and thus can be skipped, and (2) within the posting list of a term t, any document with id $<$ the current posting has already been considered. In our approach local upper bounds are no longer document-independent, hence the above two invariants will not be admitted. One solution is to localize the inverted index per topic k, and run first-level WAND retrieval in parallel on local indices, using local per-topic term upper bounds. The only communication required among parallel runs is synchronizing the threshold θ and the top K results. Localizing inverted index can be implemented physically by splitting the index, or virtually by encoding the topic k into higher bits of ad id. In the latter case, special care needs to be taken for cross-topic skips. Specifically, the skip() operation of the WAND iterator will degrade to the simple next() operation when crossing topical zones in a term posting list.

3. EXPERIMENTS

We first conducted experiments using an artificial data set generated from a mixture model as described in Section 2. Formally, given a query $q = \{t\}_{t=1}^{|q|}$ with length $|q|$ and a set of N ad documents indexed by $D = \{d\}_{d=1}^{N}$ from C topics indexed by $Z = \{k\}_{k=1}^{C}$, we wish to retrieve the top K most relevant ads. The generative process is as follows,

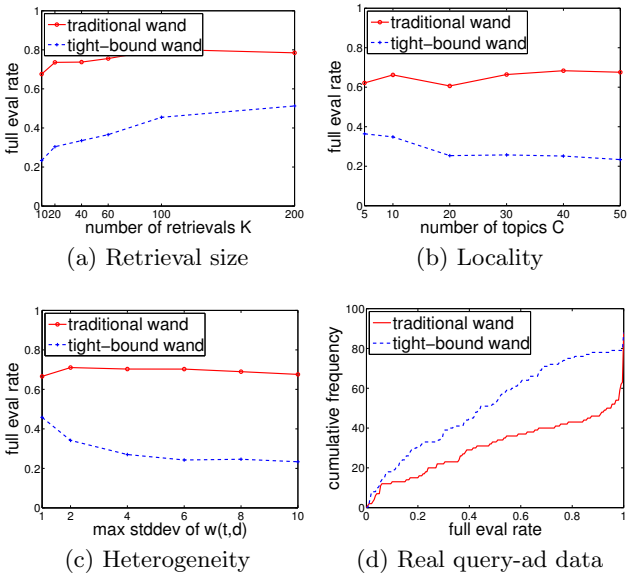
1. $w(t,q) \sim \mathcal{U}(0,1), \forall t \in q.$
2. $k(d) \sim \text{Multinomial}(\beta), \forall d.$
3. $\mu_{t,k} = \mu(w(t,d)|k) \sim \mathcal{U}(0,1), \forall t \in q, k.$
4. $\sigma_{t,k} = \sigma(w(t,d)|k) \sim \mathcal{U}(1,h), \forall t \in q, k.$
5. $w(t,d) \sim \max\left(0, \mathcal{N}\left(\mu_{t,k(d)}, \sigma^2_{t,k(d)}\right)\right), \forall t \in q, d.$

This generative model fits well with typical IR models such as the vector space model in $tf \cdot idf$ space and Markov random field (MRF) for term dependencies. With the generated data set, the global inverted index and the local per-topic indices were then readily derived.

The objective of our experiments is to compare the computational efficiencies between the traditional and the proposed tighter-bound WAND. The scoring function takes the general form as in Eq. (1). The evaluation metric is primarily the full evaluation rate (FER $= N_{\text{full}}/N$, where N_{full} is the number of fully evaluated ads), while preserving the perfect precision and recall. We also wish to gain empirical insights into how the two algorithms behave under different characteristics of the data, by varying retrieval size K, locality C, and heterogeneity $h = \max(\sigma(w(t,d)|k))$. We chose the following parameter values as our baseline configuration as summarized in Table 1, while performing univariate movement in each experiment. The experimental results are plotted in Figures 1(a), 1(b), and 1(c), where each data point is an average full evaluation rate over 10 independent random runs or queries. As the results show, the proposed tighter-bound WAND outperforms the traditional approach significantly over all parameter configurations, with 2-4 folds reduction in number of full evaluations. We observe that as the number of topics and the maximum std dev of $w(t,d)$ increase, the gain in computational efficiency is widened. As retrieval set size increases, the difference in computational reduction becomes smaller (more like web search).

Table 1: Baseline Parameters

Parameter	Description	Value
N	number of ads	$10,000$
K	number of retrievals	10
$\|q\|$	query lengh	3
C	number of topics	50
h	max stddev of $w(t,d)$	10



(a) Retrieval size (b) Locality

(c) Heterogeneity (d) Real query-ad data

Figure 1: Full evaluation rate comparison between the traditional and tighter-bound WAND.

Finally, we report experimental results with a real-world query-ad retrieval data set that contains a random sample of 100 queries and 160K ads from 1.4K categories. The empirical CDF of full evaluation rates are shown in Figure 1(d). The average full evaluation rate of the traditional WAND is 65.3%, while the tighter-bound WAND yields 42.2%, a 35.4% relative reduction in computation.

4. REFERENCES

[1] A. Z. Broder, D. Carmel, M. Herscovici, A. Soffer, and J. Zien. Efficient query evaluation using a two-level retrieval process. *CIKM 2003.*