Deconstructing Complex Search Tasks: A Bayesian Nonparametric Approach for Extracting Sub-tasks

Rishabh Mehrotra  
Dept. of Computer Science  
University College London  
London, UK  
R.Mehrotra@cs.ucl.ac.uk

Prasanta Bhattacharya  
Dept. of Information Systems  
National University of Singapore  
Singapore  
prasanta@comp.nus.edu.sg

Emine Yilmaz  
Dept. of Computer Science  
University College London  
London, UK  
emine.yilmaz@ucl.ac.uk

Abstract

Search tasks, comprising a series of search queries serving a common informational need, have steadily emerged as accurate units for developing the next generation of task-aware web search systems. Most prior research in this area has focused on segmenting chronologically ordered search queries into higher level tasks. A more naturalistic viewpoint would involve treating query logs as convoluted structures of tasks-subtasks, with complex search tasks being decomposed into more focused sub-tasks. In this work, we focus on extracting sub-tasks from a given collection of on-task search queries. We jointly leverage insights from Bayesian nonparametrics and word embeddings to identify and extract sub-tasks from a given collection of on-task queries. Our proposed model can inform the design of the next generation of task-based search systems that leverage user’s task behavior for better support and personalization.

1 Introduction

Search behavior, and information behavior more generally, is often motivated by tasks that prompt search processes that are often lengthy, iterative, and intermittent, and are characterized by distinct stages, shifting goals and multitasking (Kelly et al., 2013; Mehrotra et al., 2016). Current search systems do not provide adequate support for users tackling complex tasks, due to which the cognitive burden of keeping track of such tasks is placed on the searcher. Ideally, a search engine should be able to understand the reason that caused the user to submit a query (i.e., the actual task that caused the query to be issued) and be able to guide users to achieve their tasks by incorporating this information about the actual informational need. Clearly, identifying and analyzing search tasks is an extremely important activity not only for search engine providers but also other web based frameworks like spoken dialogue (Sun et al., 2015) and general recommendation systems (Mehrotra et al., 2014) in their effort to improve user experience on their platforms.

Previous work in the area have proposed a number of methods for identifying and extracting task knowledge from search query sessions (Mehrotra and Yilmaz, 2015b; Wang et al., 2013; Lucchese et al., 2011; Verma and Yilmaz, 2014; Mehrotra and Yilmaz, 2015a). However, while some tasks are fairly trivial and single-shot (e.g. "latest Taylor Swift album"), others are more complex and often involve multiple steps or sub-tasks (e.g. "planning a wedding").

Deciphering sub-tasks from search query logs becomes an important problem since users might exhibit different search preferences as well as expend different amounts of search effort while executing the sub-tasks. For example, while planning a wedding, users might choose to spend more time and effort on searching for a suitable venue, while spending considerably less on the choice of a wedding cake. However, even before we can analyze the variance in search effort across sub-tasks, it becomes imperative to successfully identify and extract sub-tasks for a specific task from search query logs. This turns out to be a complex problem for two reasons. First, the number of sub-tasks in a given task is not a
parameter than can be explicitly defined, and is generally task dependent. Second, while similar sounding queries like "wedding planning checklist" and "wedding dress" belong to the same task, they inherently represent different sub-tasks. This necessitates the use of advanced distancing techniques, beyond the usual bag-of-words or TF-IDF approaches.

In our current study, we propose a method for extracting search sub-tasks from a given collection of queries constituting a complex search task, using a non-parametric Bayes approach. Our generative model is not restricted by a fixed number of sub-task clusters, and assumes an infinite number of latent groups, with each group described by a certain set of parameters. We specify our non-parametric model by defining a Distance-dependent Chinese Restaurant Process (dd-CRP) prior and a Dirichlet multinomial likelihood (Blei and Frazier, 2011). Further, we draw on recent advancements that emphasize the superiority of embedding based distancing approaches over others, especially when comparing documents with less or no common words (Mikolov et al., 2013). We enrich our non-parametric model by working in the vector embedding space and propose a word-embedding based distance measure (Kusner et al., 2015) to encode query distances for efficient sub-task extraction.

3 Extracting Sub-tasks

Consider a collection of queries \( Q \) issued by search engine users, trying to accomplish certain search tasks. Quite often, these search tasks (e.g. planning a trip) are complex and conceptually decompose into a set of sub-tasks (e.g. booking flights, finding places of interest etc), each of which warrants the user to further issue multiple queries to solve. It is important to note that while the queries are observed, the inherent sub-tasks and their numbers are latent. Given a collection of on-task queries, extracted using a standard task extraction algorithm, our goal is to extract these sub-tasks from the on-task query collection.

The distance dependent Chinese restaurant process (dd-CRP) (Blei and Frazier, 2011) was recently introduced to model random partitions of non-exchangeable data. To extract sub-tasks, we consider the dd-CRP model in an embedding-space setting and place a dd-CRP prior over the search tasks.

3.1 Nonparametric Priors for Sub-tasks

The Chinese restaurant process (CRP) is a distribution on all possible partitions of a set of objects (in our case, queries). The generative process can be described via a restaurant with an infinite number of tables (in our case, sub-tasks). Customers (queries)
enter the restaurant in sequence and select a table $z_i$ to join. They pick an occupied table with a probability proportional to the number of customers already sitting there, or a new table with probability proportional to a scaling parameter $\alpha$. The dd-CRP alters the CRP by modeling customer links not to tables, but to other customers.

In our sub-task extraction problem, each task is associated with a dd-CRP and its tables are embellished with IID draws from a base distribution over mixture component parameters. Let $z_i$ denote the $i$th query assignment, the index of the query with whom the $i$th query is linked. Let $d_{ij}$ denote the distance measurement between queries $i$ and $j$, let $D$ denote the set of all distance measurements between queries, and let $f$ be a decay function. The distance dependent CRP independently draws the query assignments to sub-tasks conditioned on the distance measurements,

$$p(z_i = j | D, \alpha) \propto \begin{cases} f(d_{ij}) & \text{if } j \neq i \\ \alpha & \text{if } j = i \end{cases}$$

Here, $d_{ij}$ is an externally specified distance between queries $i$ and $j$, and $\alpha$ determines the probability that a customer links to themselves rather than another customer. The monotonically decreasing decay function $f(d)$ mediates how the distance between two queries affects their probability of connecting to each other. The overall link structure specifies a partition: two queries are clustered together in the same sub-task if and only if one can reach the other by traversing the link edges. $R(q_1:N)$ maps query assignments to sub-task assignments.

Given a decay function $f$, distances between queries $D$, scaling parameter $\alpha$, and an exchangeable Dirichlet distribution with parameter $\lambda$, N M-word queries are drawn as follows,

1. For $i \in [1,N]$, draw $z_i \sim \text{dist- } \text{CRP}(\alpha, f, D)$.
2. For $i \in [1,N]$,
   (a) If $z_i \notin R^*_q$, set the parameter for the $i$th query to $\theta_i = \theta_q$. Otherwise draw the parameter from the base distribution, $\theta_i \sim \text{Dirichlet}(\lambda)$.
   (b) Draw the $i$th query terms, $w_i \sim \text{Mult}(M, \theta_i)$.

We experimented with 3 different values of alpha and reported the best performing results. We next define the distance and decay functions which help us find task-specific query distances.

### 3.2 Quantifying Task Based Query Distances

Word embeddings capture lexico-semantic regularities in language, such that words with similar syntactic and semantic properties are found to be close to each other in the embedding space. We leverage this insight and propose a novel query-query distance metric based on such embeddings. We train a skip-gram word embeddings model where a query term is used as an input to a log-linear classifier with continuous projection layer and words within a certain window before and after the words are predicted. We next describe how we use these query term embedding vectors to define query distances.

For a search task like “planning a wedding”, frequent queries include wedding checklist, wedding planning and bridal dresses. Ideally, checklist and planning related queries constitute a different sub-task than bridal dresses. Given the overall context of weddings, words like checklist and dresses are more informative than generic words like weddings. To this end, we classify each word as background word or subtask-specific word using a simple frequency based approach on the given collection of on-task query terms and use a weighted combination of their embedding vectors to encode a query’s vector:

$$V_q = \frac{1}{n_{\text{terms}}} \sum_i n_{q_i} V_{t_i}$$

where $t_i$ is the i-th term in the query $q$, $n_{q_i}$ is the number of queries in the current task which contain the term $t_i$. We encode each query by its corresponding embedding vector representation $V_q$ and take the cosine distance of these vectors while defining $d_{ij}$. We consider a simple window decay $f(d) = 1[d < a]$ to only considers queries that are separated from the current query for a given sub-task, by a distance of, at most, $a$.

### 3.3 Posterior Inference

The posterior of the proposed dd-CRP model is intractable to compute because the dd-CRP places a prior over a combinatorial number of possible customer configurations. We employ a Gibbs sampler,
wherein we iteratively draw from the conditional distribution of each latent variable, given the other latent variables and observations. The Gibbs sampler iteratively draws from

$$p(z_{i}^{new}|z_{-i}, x) \propto p(z_{i}^{new}|D, \alpha) p(x|t(z_{-i} \cup z_{i}^{new}), G_0)$$

(2)

The first term is the dd-CRP prior and the second is the likelihood of observations ($x$) under the partition, and $t(z)$ is the sub-task formed from the assignments $z$. We employ a Dirichlet-Multinomial conjugate distribution to model the likelihood of query terms.

Queries are assigned to sub-tasks by considering sets of queries that are reachable from each other through the query assignments. Notice that many configurations of query assignments might lead to the same sub-task assignment. Finally, query assignments can produce a cycle, e.g., query 1 linking with 2 and query 2 linking with 1. This still determines a valid sub-task assignment: all queries linked in a cycle are assigned to the same sub-task. Figure 1 provides a pictorial representation of the sub-task assignment process.

### 4 Experimental Evaluation

In this section, we evaluate the robustness of the proposed sub-task extraction framework. In addition to qualitative analysis of the extracted sub-tasks, we perform a user judgment study to evaluate the quality of the extracted sub-tasks.

#### 4.1 Dataset & Baselines

We make use of the AOL log dataset which consists of 20M web queries collected over three months (Pass et al., 2006). The dataset comprises of five fields viz. the search query string, the query time stamp, the rank of the selected item (if any), the domain of the selected items URL (if any), and a unique user identifier. We augment on-task queries extracted from the AOL logs with the related searches output from different search engines by making use of their APIs.

To compare the performance of the proposed sub-task extraction algorithm, we baseline against a number of methods including state-of-the-art task extraction systems, in addition to parametric and non-parametric clustering approaches:

1. **QC-HTC** (Lucchese et al., 2011): a frequently used search task identification method.

2. **LDA** (Blei et al., 2003): a topic model based baseline which aggregates queries (similar to tweet aggregation as proposed in (Mehrotra et al., 2013)) in a session to form a document and learns an LDA model on top of it.

3. **vanilla-CRP**: a vanilla non-parametric CRP model (Wang and Blei, 2009).

4. **Proposed Approach**: the proposed embedding based dd-CRP model.

#### 4.2 Qualitative Evaluation

Table 1 shows some exemplar sub-tasks identified by the proposed model and the baseline methods using a CRP, QC-HTC and a LDA process. Each task is visualized using four search queries that were most frequently executed in relation to that sub-task, but not in any specific order among themselves. The task selected for this illustration was that of planning a wedding, and the three sub-tasks identified using our proposed method, for this particular task were wedding hairstyles, wedding dresses, and wedding cards. In comparison, however, the baseline methods failed to identify diagnostic clusters. For instance, LDA grouped wedding insurance, wedding planning books and wedding cards as a single sub-task, while CRP grouped wedding planning kits, wedding dresses and wedding decorations into
a single sub-task. Our proposed method, however, demonstrated remarkably good discriminant validity, as is clear from Table 1.

### 4.3 User Study

Evaluation of tasks and sub-tasks is an open research question. Owing to the absence of ground truth data on sub-task classification, we resort to user judgments in order to validate the quality of sub-tasks extracted. We select a sub-task at random and then choose a randomly selected pair of queries from that sub-task. Next, we ask the judges, recruited via AMT\(^1\), to affirm or deny if the two queries should be assigned to the same sub-task category. We repeat this process for a total of 100 iterations and compare the results with the ones predicted by our proposed approach, as well as with the ones predicted by the baselines.

We report the proportion of correct matches (i.e. proportion of times our predicted sub-task classifications matched the expert judgments) in Fig. 2. The label agreement among the judges was 85.4% and the performance differences were statistically significant. It is clear that our proposed method outperforms both, task extraction & topic model based baselines in making correct sub-task classifications.

### 5 Results & Discussion

Web search tasks are often complex and comprise several constituent sub-tasks. In this paper we offer a non-parametric Bayesian approach to identifying sub-tasks by grouping search queries using an embedding based dd-CRP approach. The proposed model combines insights from Bayesian non-parametrics and distributional semantics to extract sub-tasks which are not only meaningful but are also coherent. We evaluate our proposed method on the popular AOL search log dataset augmented with related search queries and demonstrate superiority over comparable approaches such as LDA and CRP. Further, we contend that our proposed approach is significantly more useful in online environments where the number of sub-tasks is never known apriori and impossible to ascertain or approximate.

In future work, we intend to consider hierarchical extensions for extracting hierarchies of tasks-subtasks. Further, using an embedding based distancing scheme, we offer an improvement in empirical performance over prior clustering approaches that have used either a bag-of-words or TF-IDF based solution. Our method offers search engine providers with a novel method to identify and analyze user task-behavior, and better support task decisions on their platforms.

### Acknowledgments

This work was supported in part by a Google Faculty Research Award.

---

\(^1\)https://www.mturk.com/mturk/welcome

---

<table>
<thead>
<tr>
<th>Proposed Approach</th>
<th>LDA</th>
<th>CRP</th>
<th>QC-HTC</th>
</tr>
</thead>
<tbody>
<tr>
<td>sub-task 1</td>
<td>sub-task 2</td>
<td>sub-task 3</td>
<td>sub-task 1</td>
</tr>
<tr>
<td>wedding hairstyles</td>
<td>used wedding dresses</td>
<td>wedding card holders</td>
<td>wedding insurance</td>
</tr>
<tr>
<td>wedding hair dos</td>
<td>colorful bridal gowns</td>
<td>wedding program</td>
<td>destination wedding</td>
</tr>
<tr>
<td>curly wedding hairstyles</td>
<td>preowned wedding dresses</td>
<td>wedding program</td>
<td>wedding planning book</td>
</tr>
<tr>
<td>pictures of wedding hair</td>
<td>wedding attire</td>
<td>regency wedding cards</td>
<td>party supply stores</td>
</tr>
<tr>
<td>wedding planning kit</td>
<td>wedding guide</td>
<td>weddings in vegas</td>
<td>destination wedding</td>
</tr>
<tr>
<td>wedding table decorations</td>
<td>save the date ideas</td>
<td>wedding cakes pictures</td>
<td>financing wedding rings</td>
</tr>
<tr>
<td>1930s wedding pictures</td>
<td>wedding vacation</td>
<td>planning a wedding</td>
<td>party supply stores</td>
</tr>
</tbody>
</table>

Table 1: Qualitative Analysis of Sub-Tasks extracted by different approaches.
References


