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 subtasks 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

Proceedings of NAACL-HLT 2016, pages 599–605,
San Diego, California, June 12-17, 2016.

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.

2 Related Work

Web search logs have been extensively studied to generate insights and provide explicit cues about the information seeking behavior of users, that would improve their search experiences. There have been attempts to extract in-session tasks (Jones and Klinkner, 2008; Lucchese et al., 2011; Spink et al., 2005), and cross-session tasks (Wang et al., 2013; Kotov et al., 2011; Li et al., 2014) from query sequences based on classification and clustering methods. Hagen *et al.* (Hagen et al., 2013) have recently presented a cascading method for logical session detection that can also be applied to search mission detection. Kotov et al (Kotov et al., 2011) and Agichtein et al (Agichtein et al., 2012) have studied the problem of cross-session task extraction via binary same-task classification. Unfortunately, pairwise predictions alone cannot generate the partition of tasks, and post-processing is needed to obtain the final task partitions (Liao et al., 2012).

Some previous attempts have been made to support people engaged in complex tasks by allowing them to take notes and record results that they already examined (Donato et al., 2010), or to provide task continuation assistance (Morris et al., 2008). Jones et al. (Jones and Klinkner, 2008) was the first work to consider the notion that there may be multiple sub-tasks associated with a user's informational needs. However, they fall short of proposing a method to identify a task from queries. Since most of these task extraction methods are based on relating a user's current query to one of her previous tasks, these methods cannot be directly used in finding and extracting sub-tasks. As a result, while task extraction methods abound, very little has been done to explicitly identify the sub-tasks from within complex search tasks.

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. planing 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) i

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 α. 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 ith query assignment, the index of the query with whom the ith 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 α 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 α, and an exchangeable Dirichlet distribution with parameter λ, N M-word queries are drawn as follows,

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

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 subtask 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:

$$V_q = \frac{1}{n_{terms}} \sum_i \frac{n_{q_{t_i}}}{\Sigma_q n_q} V_{t_i} \quad (1)$$

where t_i is the i-th term in the query q, $n_{q_{t_i}}$ 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,

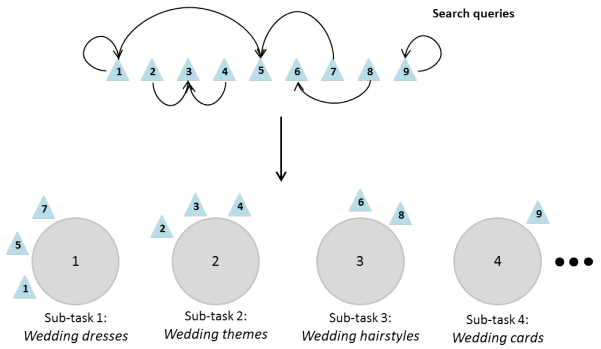

Figure 1: Visual formulation of the proposed approach. The tables represent the different sub-tasks while each triangle represents the search queries. Query assignment leads to sub-task assignments.

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

$$p(z_i^{new}|z_{-i}, x) \propto p(z_i^{new}|D, \alpha) \\ p(x|t(z_{-i} \cup z_i^{new}), G_0) \quad (2)$$

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

Proposed Approach			LDA		
sub-task 1	**sub-task 2**	**sub-task 3**	**sub-task 1**	**sub-task 2**	**sub-task 3**
wedding hairstyles	used wedding dresses	wedding card holders	wedding insurance	christian wedding vows	make wedding invitations
wedding hair dos	colorful bridal gowns	indian wedding program	destination wedding	brides	wedding cakes pictures
curly wedding hairstyles	preowned wedding dresses	wedding program	wedding planning book	cheap wedding dresses	planners
pictures of wedding hair	wedding attire	regency wedding cards	party supply stores	tea length wedding dresses	wedding colors
CRP			**QC-HTC**		
sub-task 1	**sub-task 2**	**sub-task 3**	**sub-task 1**	**sub-task 2**	**sub-task 3**
wedding planning kit	wedding theme	wedding insurance	wedding insurance	christian wedding vows	cheap dresses
destination wedding	wedding guide	weddings in vegas	destination wedding	plus size bridesmaid	wedding cakes pictures
wedding table decorations	save the date ideas	wedding cakes pictures	financing wedding rings	wedding colors	pricing weddings
1930s wedding pictures	wedding vacation	planning a wedding	party supply stores	tea length wedding dresses	wedding dresses discounts

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

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.

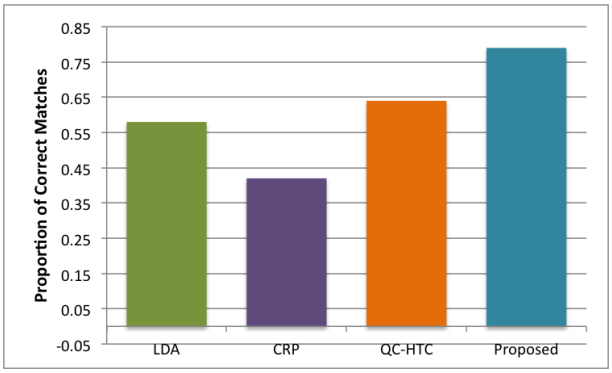

Figure 2: Judgments results for sub-task validity across compared approaches.

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 nonparametrics 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

References

[Agichtein et al.2012] Eugene Agichtein, Ryen W White, Susan T Dumais, and Paul N Bennet. 2012. Search, interrupted: understanding and predicting search task continuation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 315–324. ACM.

[Blei and Frazier2011] David M Blei and Peter I Frazier. 2011. Distance dependent chinese restaurant processes. *The Journal of Machine Learning Research*, 12:2461–2488.

[Blei et al.2003] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

[Donato et al.2010] Debora Donato, Francesco Bonchi, Tom Chi, and Yoelle Maarek. 2010. Do you want to take notes?: identifying research missions in yahoo! search pad. In *Proceedings of the 19th international conference on World wide web*, pages 321–330. ACM.

[Hagen et al.2013] Matthias Hagen, Jakob Gomoll, Anna Beyer, and Benno Stein. 2013. From search session detection to search mission detection. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 85–92.

[Jones and Klinkner2008] Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 699–708. ACM.

[Kelly et al.2013] Diane Kelly, Jaime Arguello, and Robert Capra. 2013. Nsf workshop on task-based information search systems. In *ACM SIGIR Forum*, volume 47, pages 116–127. ACM.

[Kotov et al.2011] Alexander Kotov, Paul N Bennett, Ryen W White, Susan T Dumais, and Jaime Teevan. 2011. Modeling and analysis of cross-session search tasks. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 5–14. ACM.

[Kusner et al.2015] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. 2015. From word embeddings to document distances. In *ICML*.

[Li et al.2014] Liangda Li, Hongbo Deng, Anlei Dong, Yi Chang, and Hongyuan Zha. 2014. Identifying and labeling search tasks via query-based hawkes processes. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 731–740. ACM.

[Liao et al.2012] Zhen Liao, Yang Song, Li-wei He, and Yalou Huang. 2012. Evaluating the effectiveness of search task trails. In *Proceedings of the 21st international conference on World Wide Web*, pages 489–498. ACM.

[Lucchese et al.2011] Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. 2011. Identifying task-based sessions in search engine query logs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 277–286. ACM.

[Mehrotra and Yilmaz2015a] Rishabh Mehrotra and Emine Yilmaz. 2015a. Terms, topics & tasks: Enhanced user modelling for better personalization. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 131–140. ACM.

[Mehrotra and Yilmaz2015b] Rishabh Mehrotra and Emine Yilmaz. 2015b. Towards hierarchies of search tasks & subtasks. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 73–74. International World Wide Web Conferences Steering Committee.

[Mehrotra et al.2013] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892. ACM.

[Mehrotra et al.2014] Rishabh Mehrotra, Emine Yilmaz, and Manisha Verma. 2014. Task-based user modelling for personalization via probabilistic matrix factorization. In *RecSys Posters*.

[Mehrotra et al.2016] Rishabh Mehrotra, Prasanta Bhattacharya, and Emine Yilmaz. 2016. Characterizing users' multi-tasking behavior in web search. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 297–300, New York, NY, USA. ACM.

[Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Morris et al.2008] Dan Morris, Meredith Ringel Morris, and Gina Venolia. 2008. Searchbar: a search-centric web history for task resumption and information refinding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1207–1216. ACM.

[Pass et al.2006] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *InfoScale*, volume 152, page 1.

[Spink et al.2005] Amanda Spink, Sherry Koshman, Minsoo Park, Chris Field, and Bernard J Jansen. 2005.

Multitasking web search on vivisimo. com. In *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on*, volume 2, pages 486–490. IEEE.

[Sun et al.2015] Ming Sun, Yun-Nung Chen, and Alexander I Rudnicky. 2015. Understanding users cross-domain intentions in spoken dialog systems. In *NIPS Workshop on Machine Learning for SLU and Interaction*.

[Verma and Yilmaz2014] Manisha Verma and Emine Yilmaz. 2014. Entity oriented task extraction from query logs. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1975–1978. ACM.

[Wang and Blei2009] Chong Wang and David M Blei. 2009. Variational inference for the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, pages 1990–1998.

[Wang et al.2013] Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ryen W White, and Wei Chu. 2013. Learning to extract cross-session search tasks. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1353–1364. International World Wide Web Conferences Steering Committee.