LUCAS CHAVES LIMA

UTILITY MODELS FOR COMPLEX SEARCH TASKS

Dissertation proposal presented to the Graduate Program in Ciência da Computação of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Master in Ciência da Computação.

Advisor: Rodrygo Luis Teodoro Santos

Belo Horizonte
November 2016

Contents

Chapter 1

Introduction

Search engines are the primary means by which people locate online information, serving billions of queries over hundreds of billions of documents every day. The range of online tasks has grown a lot in the past decade [25], and with that, the users have become more demanding, requiring and needing help from search engines to complete more complex tasks through the internet.

While modern web search engines already have a great perform at serving high-quality results or even direct answers for individual user queries, they frequently underperform at supporting the user to achieve successfully the task underlying that query, forcing the user to perform a series of non-trivial steps. For instance, a user planning a wedding may be obliged to submit multiple queries such as wedding dresses, wedding bouquet, churches in Belo Horizonte, and spend some effort to combine all these retrieved documents to complete the task. This project addresses the problem of retrieving documents that help in the completion of a task behind a user's query known as Task Completion Problem.

The primary goal of this dissertation is to produce a ranking for these complex tasks, in particular, complex queries that involve task planning. This project has drawn the attention to three complementary problems: The first problem is to identify and understand the task underlying a user's query. Second, we want to identify documents that are useful(not just relevant) trying in a certain way to mix usefulness and relevance of documents, seeking to address the multiple sub-tasks associated with the user's search task. One promising direction here is to consider each sub-task s as conveying a transactional search intent [6]. Third, produce a ranking combining the documents retrieved in problem 2 to support the completion of as many sub-tasks as possible with minimum redundancy on the supported sub-tasks preventing the user from the complexities of the underlying process. In our work, we will assume that the first problem is already a problem solved and focuses then just on problems 2 and 3.

Within this proposal, we expect to produce three major outcomes: a) Define the difference between utility and relevance towards the completion of a task b) Create strategies to retrieve documents that are useful and not just relevant and c) Propose a utility model for ranking search complex tasks.

The remainder of this project is organized as follows. The second chapter defines complex tasks, overviews previous approaches related to our problem of task completion, review some intent detection strategies and prior works associated with the diversification problem. Chapter 3 displays the methodology by first giving a overview of the project detailing our proposed solution, followed by evaluations strategies which will be used in this project. Lastly, Chapter 4 describe the schedule plan for this project.

Chapter 2

Related Work

In this chapter, we present some key concepts for formalizing the Task Completion Problem. We start by discussing the prior works approaches existing in the literature and how they differ from our proposal. Then, we discuss the intent detection problem, and how we are going to use it in our project. Further, we present the diversification problem and introduce two orthogonal directions a) Subtopic Mining and b) Document Ranking.

2.1 Task completion approaches

A task is an activity to be performed to accomplish a goal [15]. Resolving complex tasks with current search technology requires us to use multiple search sessions and various search strategies, and then laboriously combine and integrate information over sessions [4]. Considering that the subtasks were already identified in Task Understanding (subtopic mining), the goal of these approaches is to produce a ranking U of documents that support the completion of as many sub-tasks in T as possible and with minimum redundancy on the supported sub-tasks [34]. Also in [4], they propose a plan for task-completion involving the production of an engine that supports humans in the process of solving complex tasks using UI's combining all the subtasks needed to complete the complex task involved. In [16] the authors' approach uses the related queries from the task understanding subtask as the task set. For each related query individually, they retrieve documents using ChatNoir 2 and combine the top-N retrieved documents of the different related queries in various ways.

Different from the other previous works, this proposal has as primary objective the retrieval of useful (and not just relevant) resources for the completion of the task. Is believed that documents with high transactional intent(documents that perform any web-mediated activity) tend to be more helpful in the context of utility for completeness of the task. After identified the useful documents for each sub-task, we use diversification strategies to produce a final ranking with high coverage(sub-tasks) with low redundancy between the documents.

2.2 Intent detection

There are several studies that use the taxonomy proposed by Broder, that classify the queries intent into three categories: Informational, navigational and transactional [6]. Considering this taxonomy, search engines must be able to identify the intent of a query and provide the right answer to this specific intent. The intent of a query is defined by the goal underlying the user's query. There are different methods used to identify the different intents; our approach is only interested in one: Transactional. The transactional query is intent to documents that mediate some web-activity. To classify this intent, Ho Kang [20] collected anchor text related to each hyperlink redirecting to known transactional pages to use it as training data for a query classification module. In [24] Li, proposed an approach that uses document filtering, term filtering, and synonym expansion to identify pages with transactional intent. We believe that the utility of a document d can be estimated based on its predicted suitability for this type of intent. Thus, we propose to adapt some of these techniques to identify useful documents for the completion of a task.

2.3 Diversification

The Diversification problem is an instance of the classical NP-Hard problem known as maximum coverage problem [18]. Prove of this np-hardness can be found in [2]. The author proves the np-hardness by reducing the maximum coverage problem to the search diversification problem. Since the diversification problem is NP-hard, it has no known efficient exact solution, but fortunately, admits efficiently approximations. Some greedy approximations can be used to solve this optimization problem, which already has interesting results of approximations. This algorithm guarantees that the greedy solution is the best known polynomial solution. Over the years, a probabilistic rankings approach has been used for ranking in search results. However, these techniques assume that the relevance of the documents retrieved in the ranking is independent, which doesn't hold in realistic scenarios[10; 11].

The diversification problem is tackled from two different perspectives, ambiguity in the user's query and redundancy in the search results. From this conception, a way to handle the query ambiguity is retrieving documents that ensure the high coverage of potential information needs. In turn, redundancy can be tackled by ensuring that the retrieved documents provide a great novelty on their covered needs [30]. These two perspectives give the intuition that an ambiguous query represents not just one but multiple possible information needs [Spärck-jones et al.]. Coverage and novelty conflict in their objectives. This conflict happens because a ranking with maximum coverage may choose documents of particular needs (topics) over other needs. Similarly, a ranking with maximum novelty may cover each need as early as possible in the ranking, but not all possible needs may be covered.

There are two orthogonal directions in diversity: The subtopic mining and document ranking. The subtopic mining consists in the generation of possible topics t behind a user's query. There are different approaches to this meaning. Some exploit search logs for query expansion based on path-constrained random walks [14], others propose to exploit query reformulations provided by major web search engines as a mean to uncover different query aspects [27]. While subtopic mining focuses on topics underlying a query, our project goes to another direction, mining subtasks. The purpose of task understanding (mining subtasks) is to understand the possible tasks users might be trying to achieve underlying a query [34].

The document ranking already has several approaches in the literature for the search result diversification problem. Some of the Coverage-based approaches accomplish this goal by directly estimating how well each document covers the multiple aspects of the query, regardless of how well these aspects are covered by the other retrieved documents [30]. The coverage of a ranking can be estimated by topics [9] or relevance [27]. The authors in [2] investigated the diversification problem by adopting a taxonomy for queries and documents. In their work, two documents retrieved as a result for a query are considered similar if they are classified in one or more common categories covered by the query. In another hand, novelty-based approaches directly correlate the retrieved documents one to another, despite how well they cover diverse query aspects, to promote novel information. For example, documents can be compared in terms of content dissimilarity [8], the divergence of language models[37], or relevance score correlation [26]. Indeed, while a coverage-based strategy primarily focuses on resolving query ambiguity, a novelty-based strategy focuses on avoiding redundancy in the search results.

As defined in the previous chapters, the aim of the task completion problem is to test the usefulness of a retrieval system in supporting a user achieve a task. The structure of the Diversification problem is related to our problem of Task Completion. However, while most of the diversity strategies propose coverage related to the relevance of a document over a topic, our problem refers to a different dimension of the coverage problem, the coverage of high utility documents for each sub-tasks towards the completion of a task underlying the user's query.

Chapter 3

Methodology

In this chapter, we present the Methodology adopted to propose utility models for supporting complex search tasks.

3.1 Proposed Solution

This project addresses the problem of retrieving documents that supports users in the completion of a task behind his query known as Task Completion Problem. Formally, given a query q describing a complex search task, which potentially encompasses multiple sub-tasks T, unknown to a retrieval system, the goal of the system is to produce a ranking U of documents that support the completion of as many sub-tasks in T as possible and with minimum redundancy with respect to the supported sub-tasks.

Given a representation of the possible sub-tasks associated with a complex search task, our challenge is to produce a ranking that is useful towards the completion of this task. To make this possible to a massive amount of documents, we propose a approach in which a initial ranking R is produced from any ranking approach with pottentially relevant documents from the entire corpus, and a posterior utility-oriented ranking U is obtained by re-ranking the initial ranking R. It is noteworthy, while the initial ranking R can be produced by any relevance-oriented ranking approach, the final ranking U must explicitly address the utility requirement inherent to the task completion problem. In particular, we propose to estimate the utility of each retrieved document $d \in R$ with respect to each identified sub-task $s \in S$.

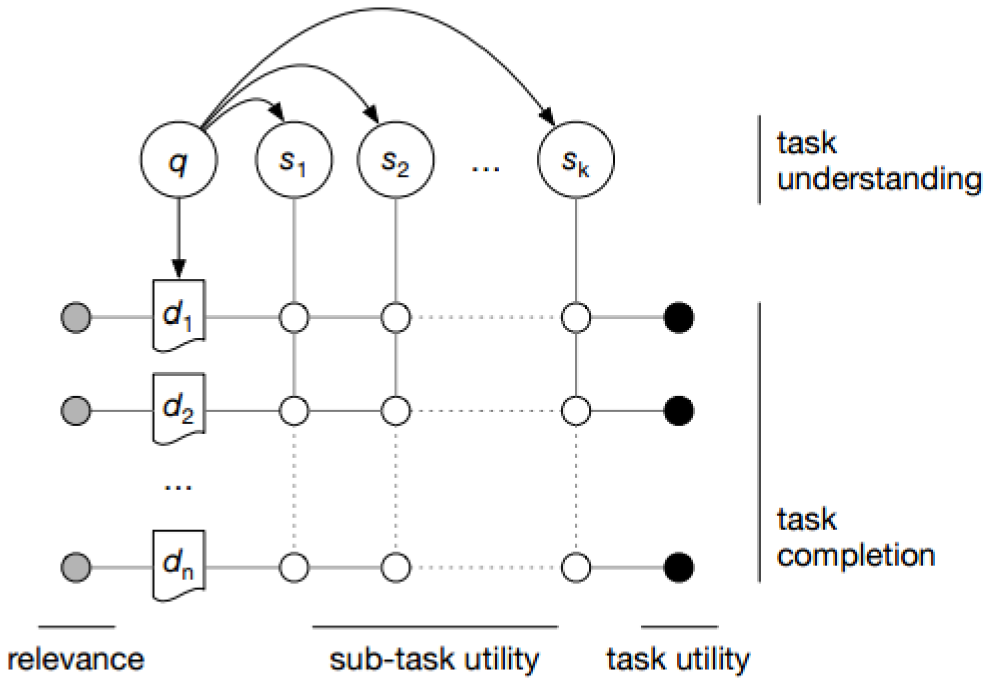

Figure 3.1. Overview of the proposed approach for the problem

The figure 1 shows the steps that might be taken during the process. The first step here is to find the possible sub-tasks underlying the user's query. Any subtopic mining strategy could be used in this step, however, we won't be focusing on this and will assume that the sub-tasks are known. The second step is to estimate the relevance of the (grey dots) of each document underlying a query, at first we will use any rank strategy well known in the literature assuming that this especific problem is not our focuses. Third, given the known sub-tasks for the problem we must ensure that the documents are not just relevant, but also useful. One promising direction to estimates the utility of a document is to consider each sub-task s as conveying a transactional search intent [6], in which case the utility of a document d can be estimated based on its predicted suitability for this type of intent [24].

However, estimating the relevance (gray dots) of each document and its utility with respect to each identified sub-task (white dots) is not enough to help the user on the completion of the task. In order to produce task-level utility estimates for each document (black dots), and hence the final utility-oriented ranking U, the interplay between the sub-tasks addressed by the various documents must be considered. Precisely, we must ensure that the final ranking U provides the maximum coverage of the actual sub-tasks T (proxied by S) underlying the user's query q. At the same time, U should also incur minimum redundancy with respect to these sub-tasks, so that the user can find useful documents for each individual sub-task as early as possible. This is a classical NP-hard problem, which fortunately admits efficient approximations [18], which have recently been exploited for the related problem of search result diversification [30]. While diversification problem, proposes the maximum diversity related to different topics, our problem has maximum diversity with respect to the sub-tasks.

3.2 Evaluation

As a document corpus, we will use the publicly available ClueWeb12 dataset, a crawl of the Web comprising 733,019,372 English documents, collected between February 10, 2012, and May 10, 2012.2. The assessment of the effectiveness of the proposed models will follow the standard experimentation framework provided by the Text Retrieval Conference (TREC). For our experimental analysis of the Task completion, we will evaluate a ranked list of documents that could be relevant to any task a user may be trying to achieve given a query. Each document will be analysed based on it usefulness and relevance. We will divide the documents in 3 different categories of utility: key(essential), useful(more documents might be needed), and not useful. For the relevance, we also classify the documents onto 3 categories: Hrel(high relevant), Rel(relevant), Non(does not provide information about task). Given these judgments, the quality of each ranked list will then be evaluated using diversity metrics such as ERR-IA and alpha-NDCG [33]. Our approaches will be compared to competing approaches participating in TREC 2016, as well as other suitable alternatives from the literature.

Chapter 4

Schedule Plan

The schedule containing the planning of the tasks for the next year of the master's degree will be described in 7 tasks described below.

- **Data Analysis:** Perform data processing to extract information from the dataset.
- **Related Work:** Update of knowledge about the subject and advanced topics.
- **Define experimental setup:** Define the experiments that will be used to prove our hypothesis
- **Implementation of baselines and literature approaches for ranking:** Implementation of prior work and baselines.
- **Implementation of the Hypothesis:** Implementation of our hypotheses for the problems 2 and 3 to evaluate our results.
- **Result Analysis:** Evaluation of hypothesis results and analysis development.
- **Writing:** Documentation of the project and writing the review.

Table 4.1. Master's project Schedule

	2016		***2017***											
Task	***Nov***	***Dez***	***Jan***	***Feb***	***Mar***	***Apr***	***May***	***Jun***	***Jul***	***Aug***	***Sept***	***Oct***	***Nov***	***Dez***
A		x	x			x	x	x						
B	x	x	x	x	x	x	x	x	x	x	x	x	x	x
C			x	x	x			x	x					
D				x	x	x								
E			x	x	x			x	x	x	x			
F					x	x	x				x	x	x	x
G	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Bibliography

[1] Agichtein, E. and Zheng, Z. (2006). Identifying "best bet" web search results by mining past user behavior. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 902--908, New York, NY, USA. ACM Press.

[2] Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. (2009). Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5--14, New York, NY, USA. ACM.

[3] Allan, J., Croft, B., Moffat, A., and Sanderson, M. (2012). Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne. *SIGIR Forum*, 46(1):2--32. ISSN 0163-5840.

[4] Balog, K. (2015). Task-completion engines: A vision with a plan. In *Proceedings of the First International Workshop on Supporting Complex Search Tasks*, SCST '15.

[5] Bonchi, F., Maarek, Y., Chi, T., and Donato, D. (2010). Do you want to take notes? identifying research missions in yahoo search pad. In *Proceedings of WWW 2010: Proceedings of the 19th international conference on World Wide Web, Raleigh, North Carolina, USA*.

[6] Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, 36(2):3--10. ISSN 0163-5840.

[7] Cao, H., Hu, D. H., Shen, D., Jiang, D., Sun, J.-T., Chen, E., and Yang, Q. (2009). Context-aware query classification. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 3--10, New York, NY, USA. ACM.

[8] Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st*

Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98, pages 335--336, New York, NY, USA. ACM.

[9] Carterette, B. and Chandar, P. (2009). Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1287--1296, New York, NY, USA. ACM.

[10] Chaudhuri, S., Das, G., Hristidis, V., and Weikum, G. (2006). Probabilistic information retrieval approach for ranking of database query results. *ACM Trans. Database Syst.*, 31(3):1134--1168. ISSN 0362-5915.

[11] Cooper, W. S. (1971). The inadequacy of probability of usefulness as a ranking criterion for retrieval system output. Technical report, University of California, Berkeley.

[12] Dan Morris, Meredith Ringel Morris, G. V. (2008). Searchbar: a search-centric web history for task resumption and information re-finding. *ACM Conference on Human Factors in Computing Systems*, pages 1207–1216.

[13] Gade, M., Hall, M. M., Huurdeman, H., Kamps, J., Koolen, M., Skove, M., Toms, E., and Walsh, D. (2015). Report on the first workshop on supporting complex search tasks. *SIGIR Forum*, 49(1):50--56. ISSN 0163-5840.

[14] Guo, J., Xu, G., Li, H., and Cheng, X. (2008). A unified and discriminative model for query refinement. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 379--386, New York, NY, USA. ACM.

[15] Hackos, J. T. and Redish, J. C. (1998). *User and Task Analysis for Interface Design.* John Wiley & Sons, Inc., New York, NY, USA. ISBN 0-471-17831-4.

[16] Hagen, M., Göring, S., Keil, M., Anifowose, O., Othman, A., and Stein, B. (2015). Webis at TREC 2015: Tasks and total recall tracks. In [34].

[17] Hassan, A. and White, R. W. (2012). Task tours: Helping users tackle complex search tasks. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1885--1889, New York, NY, USA. ACM.

[18] Hochbaum, D. S., editor (1997). *Approximation Algorithms for NP-hard Problems.* PWS Publishing Co., Boston, MA, USA. ISBN 0-534-94968-1.

[19] Jeh, G. and Widom, J. (2003). Scaling personalized web search. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 271--279, New York, NY, USA. ACM.

[20] Kang, I.-H. (2005). *Transactional Query Identification in Web Search*, pages 221--232. Springer Berlin Heidelberg, Berlin, Heidelberg.

[21] Kotov, A., Bennett, P. N., White, R. W., Dumais, S. T., and Teevan, J. (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*, SIGIR '11, pages 5--14, New York, NY, USA. ACM.

[22] Krieger, M. and Ahn, D. (2010). Tweetmotif: exploratory search and topic summarization for twitter. In *In Proc. of AAAI Conference on Weblogs and Social*.

[23] Li, L., Zhou, K., Xue, G.-R., Zha, H., and Yu, Y. (2009). Enhancing diversity, coverage and balance for summarization through structure learning. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 71--80, New York, NY, USA. ACM.

[24] Li, Y., Krishnamurthy, R., Vaithyanathan, S., and Jagadish, H. V. (2006). Getting work done on the web: Supporting transactional queries. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 557--564, New York, NY, USA. ACM.

[25] Marchionini, G. (2006). Exploratory search: From finding to understanding. *Commun. ACM*, 49(4):41--46. ISSN 0001-0782.

[26] Rafiei, D., Bharat, K., and Shukla, A. (2010). Diversifying web search results. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 781--790, New York, NY, USA. ACM.

[27] Santos, R. L., Macdonald, C., and Ounis, I. (2010). Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 881--890, New York, NY, USA. ACM.

[28] Santos, R. L., Macdonald, C., and Ounis, I. (2011). Intent-aware search result diversification. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 595--604, New York, NY, USA. ACM.

[29] Santos, R. L., Macdonald, C., and Ounis, I. (2013). Learning to rank query suggestions for adhoc and diversity search. *Inf. Retr.*, 16(4):429--451. ISSN 1386-4564.

[30] Santos, R. L. T., Macdonald, C., and Ounis, I. (2015). Search result diversification. *Found. Trends Inf. Retr.*, 9(1):1--90. ISSN 1554-0669.

[Spärck-jones et al.] Spärck-jones, K., Robertson, S. E., and S, M. Paper ambiguous requests: implications for retrieval tests, systems and theories.

[32] Toms, E. G., Villa, R., and McCay-Peet, L. (2013). How is a search system used in work task completion? *J. Information Science*, 39(1):15–25.

[33] Voorhees, E. M. and Buckland, L. P., editors (2012). *Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012*, volume Special Publication 500-298. National Institute of Standards and Technology (NIST).

[34] Voorhees, E. M. and Ellis, A., editors (2015). *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*, volume Special Publication 500-319. National Institute of Standards and Technology (NIST).

[35] White, R. W. (2016). *Interactions with Search Systems*, volume 4 of *10*. New York, NY : Cambridge University Press, 2016, The address, 3 edition. ISBN 9781107034228 1107034221. An optional note.

[36] Xu, Y. and Mease, D. (2009). Evaluating web search using task completion time. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 676--677, New York, NY, USA. ACM.

[37] Zhai, C. X., Cohen, W. W., and Lafferty, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 10--17, New York, NY, USA. ACM.

[38] Zhu, Y., Lan, Y., Guo, J., Cheng, X., and Niu, S. (2014). Learning for search result diversification. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 293--302, New York, NY, USA. ACM.