A Survey of Query Auto Completion in Information Retrieval

Fei Cai
Science and Technology on Information Systems Engineering Laboratory,
National University of Defense Technology,
Changsha, China
University of Amsterdam, The Netherlands
f.cai@uva.nl

Maarten de Rijke
University of Amsterdam, The Netherlands
derijke@uva.nl

Full text available at: http://dx.doi.org/10.1561/500000055
Editorial Scope

Topics

Foundations and Trends® in Signal Processing publishes survey and tutorial articles in the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multirate signal processing
- Multi-resolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing

Information for Librarians

Foundations and Trends® in Signal Processing, 2016, Volume 10, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.
A Survey of Query Auto Completion in Information Retrieval

Fei Cai
Science and Technology on Information Systems Engineering Laboratory,
National University of Defense Technology,
Changsha, China
University of Amsterdam, The Netherlands
f.cai@uva.nl

Maarten de Rijke
University of Amsterdam, The Netherlands
derijke@uva.nl
# Contents

1 Introduction 2
   1.1 Motivation 8
   1.2 Historical notes 4
   1.3 Aims of this survey 6
   1.4 Structure of this survey 6

2 Query Auto Completion 8
   2.1 Problem formulation 8
   2.2 Aims of query auto completion 10
   2.3 A taxonomy of solutions to query auto completion 11

3 Heuristic Approaches to Query Auto Completion 14
   3.1 Time-sensitive query auto completion models 14
   3.2 User-centered query auto completion models 21
   3.3 Summary 25

4 Learning-based Approaches to Query Auto Completion 26
   4.1 Learning from time-related characteristics 27
   4.2 Learning from user interactions 30
   4.3 Summary 34

5 Evaluation 35
Abstract

In information retrieval, query auto completion (QAC), also known as type-ahead [Xiao et al., 2013, Cai et al., 2014b] and auto-complete suggestion [Jain and Mishne, 2010], refers to the following functionality: given a prefix consisting of a number of characters entered into a search box, the user interface proposes alternative ways of extending the prefix to a full query. Ranking query completions is a challenging task due to the limited length of prefixes entered by users, the large volume of possible query completions matching a prefix, and the broad range of possible search intents. In recent years, a large number of query auto completion approaches have been proposed that produce ranked lists of alternative query completions by mining query logs.

In this survey, we review work on query auto completion that has been published before 2016. We focus mainly on web search and provide a formal definition of the query auto completion problem. We describe two dominant families of approaches to the query auto completion problem, one based on heuristic models and the other based on learning to rank. We also identify dominant trends in published work on query auto completion, viz. the use of time-sensitive signals and the use of user-specific signals. We describe the datasets and metrics that are used to evaluate algorithms for query auto completion. We also devote a chapter to efficiency and a chapter to presentation and interaction aspects of query auto completion. We end by discussing related tasks as well as potential research directions to further the area.
1 Introduction

1.1 Motivation

Word completion is a user interface feature that offers the user a list of words after one or more letters have been typed. It has been around as a feature of word editors and command shells for nearly half a century. In information retrieval, word completion is best known in the form of query auto completion, which provides users with suggested queries as they begin to enter their query in the search box. The user’s incomplete input is often called a query prefix and the suggested queries are often called query completions. As shown in Figure 1.1, these query completions often start with the characters that the user has entered into the search box (see Figure 1.1a) and are adjusted as the user continues typing (see Figure 1.1b). Query auto completion (QAC) helps users to formulate their query when they have an intent in mind but not a clear way of expressing in a query. It helps to avoid possible spelling mistakes, especially on devices with small screens. It has been reported that, when submitting English queries, using QAC by selecting suggested query completions has saved more than 50% keystrokes of global Yahoo! searchers in 2014 [Zhang et al., 2015]. In addition, cutting down the search duration of users implies a lower load on the search engine, which results in savings in machine resources and maintenance [Bar-Yossef and Kraus, 2011].
1.1. Motivation

(a) A list of four query completions for the prefix “information.”

(b) An updated list of four query completions for the prefix “information re.”

Figure 1.1: Examples of query auto completion.

Importantly, the use of QAC is not limited to web search engines such as Baidu, Bing, Google, or Yandex. QAC has also become part of other common services for search tasks. Facebook\(^1\) provides QAC for finding friends, typically providing four suggestions; Twitter\(^2\) uses QAC for searching tweets, also offering up to four suggestions; QAC is employed for product search in online shopping stores such as Amazon,\(^3\) which often produce up to ten suggestions; and it assists users to formulate requests for specific videos in online video-sharing websites such as Youtube,\(^4\) where users also get up to ten suggestions. As with web search, users of QAC functionality in other settings are helped to find an ideal query to address a particular search task. Research indicates that using the suggested queries can greatly improve user satisfaction, especially for informational queries [Song et al., 2011b].

\(^1\)https://www.facebook.com
\(^2\)https://twitter.com
\(^3\)http://www.amazon.com
\(^4\)https://www.youtube.com
1.2 Historical notes

Query auto completion is a particular form of word prediction: when a writer writes the first letter or letters of a word, a word predictor lists possible words as choices and if the intended word is listed amongst the choices, it can be selected. In a different form of word prediction, the most likely next words (instead of completions) are listed. The core idea goes back at least half a century: Longuet-Higgins and Ortony [1968] describe a method to help decrease the number of keystrokes needed in order to complete a word, in their case commands and identifiers entered by developers. Another important motivation was to help individuals with physical disabilities increase their typing speed. Early work on word prediction focused both on assessing the cognitive benefits of word prediction and on algorithmic developments. For instance, early studies found that reduction of keystrokes through word prediction is often offset by having to scan the list of predicted words and selecting the desired word [Vanderheiden and Kelso, 1987]. Displaying five suggested words in a vertical manner was found to provide a reasonable balance between keystrokes saving and cognitive load [Swiffin et al., 1987].

Early algorithmic work on word prediction, that is, work from the 1970s and 1980s, can be grouped into three classes: character predictors, word completers, and combinations of the two [Darragh, 1988]. Character predictors accelerate input by making likely letters faster and easier to select. Word completors support text input by offering words based on an initial prefix of one or more letters entered by the user. Combined systems do both. The Reactive Keyboard introduced by Darragh et al. [1990] is a combined approach that also incorporates many facilities of today’s query auto completion methods, including the use of popular words and of task appropriate words, both as identified from past query behavior.

Witten and Darragh [1992] provide an extensive treatment of work on word completion as an assistive technology up to the early 1990s. The Reactive Keyboard mentioned above is a prototypical example for the period as it incorporates the functionality of word completion and accelerates typewritten communication with a computer. Its word completion is mostly based on a standard dictionary; later approaches extended it to use adaptive modeling, which produces completions based on previously entered text. Similarly, a user’s previously entered text, i.e., the search context, is taken into account.
1.2. Historical notes

for personalized query auto completion [Bar-Yossef and Kraus, 2011]. Furthermore, a user’s long-term search history has been explored for query auto completion [Cai et al., 2016a], but it has not been incorporated into a word completion system. Basically, word completion relies on a special tree structure to match the input and its completions [Witten and Darragh, 1992, Darragh et al., 1990], which is similar to the data structure used in query auto completion. In addition, both word completion and query auto completion often build on simple lexical models based on n-grams, which can be directly extracted from a text corpus or query log [Cai and de Rijke, 2016b].

Sentence completion is a slightly different task: given an initial fragment of a sentence, identify the remaining part of the sentence that the user intends to write. This task setting is motivated by processes such as answering questions sent by email in a call center, or writing letters in an administrative environment [Grabski and Scheffer, 2004]. Grabski and Scheffer [2004], Bickel et al. [2005], Nandi and Jagadish [2007] study this task based on lexical statistics of text collections.

The use of word prediction for search, goes back to the Emacs text editor [Ciccarelli, 1978], which offered a single line of feedback, not a list of suggestions. In the early 2000s, Raskin [2000] advocated the use of incremental search in which a user gets instant feedback as they enter a query based on what may exist in the content search over what he called delimited search, which basically is a traditional search interface allowing search in three steps: a user submits a query, the system computes a result page, the user receives the result page. Since then, query auto completion has been used by internet browsers, development environments, web sites, desktop search, operating systems, databases, email clients, and search engines, which are the focus of this survey. Around the middle of the 2000s, QAC was being offered by a growing number of search engines. For instance, Google Suggest, which launched in 2004, was an early showcase of query auto completion in the search engine setting. And around the same time Kayak, a travel metasearch engine, completed location related queries with travel related suggestions such as airport codes.

5 http://looksgoodworkswell.blogspot.nl/2005/12/distracting-or-narrowing-looking.html
6 https://www.kayak.com
1.3 Aims of this survey

Because of the clear benefits of QAC, a considerable number of algorithmic approaches to QAC have been proposed in the past few years. Query logs have proved to be a key asset underlying most of the recent research. In this paper we survey this research. We focus on summarizing the literature on QAC and provide a general understanding of the wealth of QAC approaches that is currently available. Our contributions in this survey can be summarized as follows:

1. We provide researchers who are working on query auto completion or related problems in the field of information retrieval with a good overview and analysis of state-of-the-art QAC approaches. In particular, for new researchers to the field, the survey can serve as an introduction to the state-of-the-art.

2. We offer a comprehensive perspective on QAC approaches by presenting a taxonomy of existing solutions. In addition, we present solutions for QAC under different conditions such as available high-resolution query logs, in-depth user interactions with QAC using eye-tracking, and elaborate user engagements in a QAC process. We also discuss practical issues related to QAC.

3. We present a detailed discussion of core challenges and promising open directions in QAC.

Our focus in this survey is on effectiveness of QAC. We do, however, include chapters on efficiency (Chapter 6) and on presentation and interaction aspects of query auto completion (Chapter 7).

1.4 Structure of this survey

The rest of this survey is organized as follows. Chapter 2 describes the problem of QAC as well as some closely related tasks; Chapter 3 and Chapter 4 detail published QAC approaches based on heuristic models and on learning algorithms, respectively; Chapter 5 presents publicly available benchmarks as well as metrics for QAC evaluation. In Chapter 6 we discuss efficiency and robustness aspects of query auto completion. In Chapter 7 we include a survey of studies into interface and interaction aspects of query auto completion. We include a short overview of three areas of research that are closely
1.4. Structure of this survey

related to QAC, i.e., query suggestion, query expansion and query correction in Chapter 8. We conclude the survey in Chapter 9 and point out follow-up research directions.
References


Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. Context-aware query suggestion by mining click-through and session data. In Proceedings of the 14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD ’08, pages 875–883, New York, NY, USA, 2008. ACM.


References


References


Katja Hofmann, Anne Schuth, Shimon Whiteson, and Maarten de Rijke. Reusing historical interaction data for faster online learning to rank for IR. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM ’13, pages 183–192, New York, NY, USA, 2013a. ACM.


References


References


Zhen Liao, Daxin Jiang, Enhong Chen, Jian Pei, Huanhuan Cao, and Hang Li. Mining concept sequences from large-scale search logs for context-aware query suggestion. ACM Transactions on Intelligent Systems and Technology, 3(1):Article 17, 2011.


References


References


