Information Processing and Management 54 (2018) 726–739

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A Wikipedia powered state-based approach to automatic search query enhancement

Kyle Goslin*, Markus Hofmann

Institute of Technology Blanchardstown, Department of Informatics and Engineering, Blanchardstown Road North, Dublin 15, Ireland

ARTICLE INFO

Keywords:
Automatic Search Query Enhancement
Query drift
Information retrieval
Wikipedia

ABSTRACT

This paper describes the development and testing of a novel Automatic Search Query Enhancement (ASQE) algorithm, the Wikipedia N Sub-state Algorithm (WNSSA), which utilises Wikipedia as the sole data source for prior knowledge. This algorithm is built upon the concept of iterative states and sub-states, harnessing the power of Wikipedia's data set and link information to identify and utilise reoccurring terms to aid term selection and weighting during enhancement. This algorithm is designed to prevent query drift by making callbacks to the user's original search intent by persisting the original query between internal states with additional selected enhancement terms. The developed algorithm has shown to improve both short and long queries by providing a better understanding of the query and available data. The proposed algorithm was compared against five existing ASQE algorithms that utilise Wikipedia as the sole data source, showing an average Mean Average Precision (MAP) improvement of 0.273 over the tested existing ASQE algorithms.

1. Introduction

The process of searching for content on the web is typically done by entering search terms into a text field on the front-end of a search engine. This process however, can be seen as a one-size-fits-all approach, generalising the user's requirements, needs, background knowledge with search engines and overall search ability. Automatic Search Query Enhancement (ASQE) algorithms are designed to modify a user search query e.g., the terms entered by a user into a search engine text field, through the addition, removal or correction (Vilares, Alonso, Doval, & Vilares, 2016) of search terms to improve precision / recall of a search query.

ASQE algorithms often require additional information e.g., a data source containing terms, that can be used as a source of candidate expansion terms e.g., terms that possibly relevant to the user's query, as a method to gauge the importance of available terms. Recent ASQE algorithms (ALMasri, Berrut, & Chevallet, 2013; Boston, Fang, Carberry, Wu, & Liu, 2014; Bruce, Gao, Andreae, & Jabeen, 2012, Zhao, Liu, Qin, & Li, 2014; Zingla, Chiraz, & Slimani, 2016) have shown the use of dynamic data sources, such as Wikipedia,[1] which offers high quality and ever changing articles with common fields and structure, can be beneficial to the ASQE process. Unlike utilising static document collections and thesauri[2] which require expert knowledge to maintain, Wikipedia has shown to be beneficial as a source of prior knowledge for domain specific query enhancement such as in the area of patent retrieval (Al-Shboul & Myaeng, 2014; Sharma, Tripathi, & Tripathi, 2015). Many ASQE approaches however, become so focused on the term identification process, they do not consider if the terms are related to the user's query as a whole. It can often be the case that too

* Corresponding author.
E-mail address: kylegoslin@gmail.com (K. Goslin).

[1] https://www.wikipedia.org.

[2] https://wordnet.princeton.edu/.

http://dx.doi.org/10.1016/j.ipm.2017.10.001
Received 23 August 2016; Received in revised form 4 October 2017; Accepted 10 October 2017
Available online 04 November 2017
0306-4573/ © 2017 Elsevier Ltd. All rights reserved.

many enhancement terms are added or that the terms that have been selected are of questionable relevance or conceptually distant causing *query drift* (Shtok, Kurland, Carmel, Raiber, & Markovits, 2012).

ASQE is built upon a number of different IR techniques that can be further enhanced by utilising Wikipedia. These technique include term weighting (Karisani, Rahgozar, & Oroumchian, 2016), linguistic understanding (Selvaretnam & Belkhatir, 2016), relevance calculation (Zhao & Callan, 2012), term disambiguation (Habibi, Mahdabi, & Popescu-Belis, 2016; Yadav & Kumar, 2016) and similarity assessment based up Wikipedia articles (Jiang, Zhang, Tang, & Nie, 2015). The additional data available in Wikipedia can be beneficial to users, as during search, users with little knowledge about the area of search have shown to perform worse due to their lack of prior knowledge when compared to domain experts (Monchaux, Amadieu, Chevalier, & Marin, 2015). He and Ounis (2009) identified two possible reasons for the failure of ASQE, low query quality and topic drift. As search queries can be overly simple or complex, recent research has moved towards understanding the structural and syntactic complexity of search queries (Roy, Agarwal, Ganguly, & Choudhury, 2016), which can further improve ASQE techniques. The context of a user during search plays an important role as it often does not exist for the user at the beginning of a search session (Fourney & Dumais, 2016).

In this research, we pose the question: does a state based approach to ASQE which persists the user's original query during all IR processes improve the precision of search results for a given user query?. The objective of this research is to utilise the available data from the Wikipedia data set and term relevance metrics for each data source to automatically enhance search queries to improve the precision of search results. The theory behind this research is that a state based approach provides a number of chances for re-occurring terms be identified, and when an alternative method for selection of available candidate terms is used, it will allow optimal candidate terms to surface. In addition to this, during the process of gathering of related data for a given query, if a reference is made back to the original query to provide constant relevance persistence of the user's original intent with selected candidate terms, the collection of enhancement terms will be more relevant to the user.

To do this, a novel state and sub-state based approach to ASQE was developed with a stem query e.g., the user's original unmodified query with additional selected terms, and a term window based selection process of candidate terms. The proposed algorithm, the Wikipedia N Sub-state Algorithm (WNSSA) is described and tested using 50 of the TREC-9 & 50 TREC 2014 Web Topics[3] on the ClueWeb12 full data set[4]. In addition to this, five existing ASQE algorithms that utilise different aspects of the Wikipedia data set were implemented and analysed. Relevance calculations for each algorithm are performed using the Average Precision@10 results from each of the enhanced sample topics and the overall Mean Average Precision@10 for each tested algorithm.

The main contributions defined in this paper include: 1) A novel state based approach to the process of gathering and identifying candidate enhancement terms for ASQE; 2) Stem query generation and utilisation between states during the Information Retrieval (IR) process for ASQE, consisting of the user's original query and identified relevant enhancement terms from the current internal states to prevent query drift; 3) Term window based selection of enhancement terms for ASQE; and 4) A comprehensive cross-analysis of five existing ASQE algorithms that utilise Wikipedia as the sole data source for candidate enhancement terms against the proposed algorithm. This paper begins in Section 2 reviewing the area of ASQE and query drift, with a focus on algorithm that utilise Wikipedia as the data source for expansion terms. Section 3 provides an overview of the methodology followed. The core focus of this paper is the developed ASQE algorithm, the WNSSA, which is described in Section 4. Section 5 outlines the ASQE algorithms tested and the data sets utilised during the testing and analysis process.

To provide an understanding of the results, Section 6 discusses the results of the testing process of the five existing Wikipedia powered ASQE algorithms and the proposed algorithm. Section 7 concludes this research outlining the key findings. As work on the WNSSA is ongoing, Section 8 outlines the future work for this study.

2. Related work

ALMasri et al. (2013) proposed a Wikipedia based semantic query enrichment algorithm, whereby semantically related terms are extracted from Wikipedia and then used as Pseudo Relevance Feedback (PRF). This process is achieved through the following steps: Collect all articles $S(q)$ which are entitled by the user's query q. Each article $a \in S(q)$ has the probability $P(a|q)$ of being used in the enrichment process. The probability is defined as $P(a|t) = \frac{|O(a)|}{\sum_{a_i \in S(t)} \left| O(a_i) \right|}$, where $O(a)$ is the set of articles that a points to. The expansion set ES of selected n number of articles for user query q are defined as $ES(q, n) = \bigcup_{a \in S(q)} f(a, \lceil n \times P(a|q) \rceil)$. The collection of terms for query q are built from a union of article titles in the enrichment set. A weight is attached to each between 0 and 1, whereby 1 is most important and 0 is least important. The weight for each of the terms is defined as $weight(t, q_e) = \alpha \times SIM(a_q, a_t)$, whereby α is a tuning parameter between 0 and 1.

Boston et al. (2014) proposed a tool titled Wikimantic which exploits Wikipedia articles and their inter-article reference relations which has shown to be effective for short queries. They define an AtomicConcept as a simple form of a concept. Each article is considered a series of terms which was generated by an AtomicConcept. The prior probability of $P(A)$ generating terms, where A is an individual article is defined as $P(A) = \frac{number\ of\ incoming\ links}{number\ of\ links\ in\ Wikipedia}$. As most of the articles in Wikipedia are linking to other articles, the authors define the probability of article A generating term t is defined as $P(t|A) = \frac{count(t, A)}{number\ of\ words\ in\ A}$.

Due to the limitation that not all articles will have a variety of different terms to check their probability with, the Microsoft n-

[3] http://trec.nist.gov/data/web_topics.html.

[4] http://lemurproject.org/clueweb12/.