Promoting Document Relevance Using Query Term Proximity for Exploratory Search

Vikram Singh, National Institute of Technology, Kurukshetra, India*

https://orcid.org/0000-0001-6315-0872

ABSTRACT

In the information retrieval system, relevance manifestation is pivotal and regularly based on document-term statistics, i.e., term frequency (tf), inverse document frequency (idf), etc. Query term proximity (QTP) within matched documents is mostly under-explored. In this article, a novel information retrieval framework is proposed to promote the documents among all relevant retrieved ones. The relevance estimation is a weighted combination of document statistics and query term statistics, and term-term proximity is simply aggregates of diverse user preferences aspects in query formation, thus adapted into the framework with conventional relevance measures. Intuitively, QTP is exploited to promote the documents for balanced exploitation-exploration, and eventually navigate a search towards goals. The evaluation asserts the usability of QTP measures to balance several seeking tradeoffs, e.g., relevance, novelty, result diversification (coverage, topicality), and overall retrieval. The assessment of user search trails indicates significant growth in a learning outcome (due to novelty).

KEYWORDS

Exploratory Search, Information Retrieval, Query Term Proximity, Relevance, Retrieval Strategy

INTRODUCTION

Information-seeking is a fundament endeavor of human being and several information search systems has been deigned to assist a user to pose queries and retrieves informative data to accomplish search goals. The traditional systems strongly trust user's capability of phrasing precise request and perform better if requests are short and navigational. A potential obstacle to such systems is an astonishing rate of information overload that makes difficult to a user for identifying useful information. Therefore nowadays, search focus is shifting from finding to understanding information (White & Roth, 2009), especially in *discovery-oriented* search. When a user wants information for learning purpose, decision making or other cognitive activity, the conventional search methodologies are not capable to assist, though data exploration is helpful. A data exploration synthesis *focused* search

*Corresponding Author

and *exploratory* browsing, to discover the interesting data objects. Though, exploration become a recall-oriented navigation over complex and huge datasets using short typed ill-phrased data request (Idreos, Papaemmanouil, & Chaudhuri, 2015; White, 2016; Marchionini, 2006), and thus requires strong support for adaptive relevance measures in retrieval framework (Nandi, & Jagadish, 2011).

In the data deluge, retrieval of relevant data requires either formal awareness of complex schema and content for the formulation of a data retrieval request or assistance from information system (Kersten, Idreos, Manegold, & Liarou, 2011; Huston, Culpepper, & Croft, 2014). For both situations, the system employs *implicit* measures to outline matched objects and *explicit* measures to eventually steer search towards a *region-of-interest*. Most existing retrieval models score a document predominantly on documents-terms statistics, i.e. *document lengths*, *query-term frequencies*, *inverse document frequencies*, etc (Van, 1977; Daoud & Huang, 2013). Intuitively, the *query terms proximities* (QTPs) within pre-fetched result set/documents could be exploited for re-position/re-raking of the documents/results in which the matched query terms are close to each other. For example, an information search considering the query '*exploratory search*' on two documents, both matching the two query terms once:

Doc_1: {…*exploratory search*………}.
Doc_2: {….*exploratory*….*search*….}.

Intuitively, *document_1* should be ranked higher, as occurrences of both query terms are closest to each other. In compare to the *document_2*, where both query terms are far apart and their combination does not necessarily imply the meaning of '*exploratory search*'.

The *term-term* affinity within matched document has role to play during the retrieval and eventually to position the document in appropriate relevance (Salton & Buckley, 1988; Borlund, 2003; Verma, 2016). For an information search, a user specify data request in more than one terms with an anticipated inherent closeness. The closeness in query terms characterizes structural constraints of a user query and the importance between two matched documents in an information-seeking. The *query term proximity* is one measure, however, has been principally under-explored in traditional retrieval framework and models; mainly due to intrinsic design concerns (*how we can model proximity*) and its overall usability (*what it serve*) into a retrieval model.

This paper systematically explores the query term proximity heuristic, to guide the user's information-seeking by deliberating both *document-terms* (DTs) and *query-terms* (QTs) relevance means. The focus is on three *research questions* (RQs):

RQ1: *What constitutes relevance in exploitation and exploration? What relevance type is most significant?*

RQ2: *How can query term proximity (QTP) be adapted with document-terms relevance to optimize information exploitation and eventually exploration efforts?*

RQ3: *Finally, how to design an information retrieval framework that, account user' search task while rewarding or penalizing both relevance measures.*

For a given *document corpus* and *query terms* (in user's query), relevance manifestation is done across three factors: coincidence of QTs with DTs (*intra-document* and *inter-document* relevance) and *Span* of QTs (*intra-document*) and distance of QTs (*intra-document*). The significance of measures is derived via a study on user defined *search trails* (STs), and eventually the impact of overall retrieval framework on the exploration efforts.

Related Content

A Presentation-Preserved Compositional Approach for Integrating Heterogeneous Systems: Using E-Learning as an Example
Fang-Chuan Ou Yang (2013). *Modern Library Technologies for Data Storage, Retrieval, and Use (pp. 210-229).*
www.irma-international.org/chapter/presentation-preserved-compositional-approach-integrating/73778

Annotation of Medical Images
Chia-Hung Weiand Sherry Y. Chen (2012). *Intelligent Multimedia Databases and Information Retrieval: Advancing Applications and Technologies (pp. 74-90).*
www.irma-international.org/chapter/annotation-medical-images/59953

An Online Measure of Discernment
Hazel C. V. Traüffer, Corné L. Bekker, Mihai C. Bocarneaand Bruce E. Winston (2013). *Online Instruments, Data Collection, and Electronic Measurements: Organizational Advancements (pp. 254-270).*
www.irma-international.org/chapter/online-measure-discernment/69745

Short-Term Load Forecasting for a Captive Power Plant Using Artificial Neural Network
Vidhi Tiwariand Kirti Pal (2022). *International Journal of Information Retrieval Research (pp. 1-11).*
www.irma-international.org/article/short-term-load-forecasting-captive/289613

A Hybrid User-Centric Approach for Efficient Web Service Selection
Neerja Negiand Satish Chandra (2020). *International Journal of Information Retrieval Research (pp. 1-20).*
www.irma-international.org/article/a-hybrid-user-centric-approach-for-efficient-web-service-selection/249698