International Journal of Information Retrieval Research
Volume 13 • Issue 1

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

DOI: 10.4018/IJIRR.325072 *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.

Contribution and Outline

The key contribution is a novel retrieval framework for the discovery-oriented information-seeking, on which the role of *document-term* (DT) and *query-term* (QT) relevance measures are synthesis. Additional contributions set proposed information retrieval framework apart from the previous researches:

(1). We investigated diverse relevance aspects of a data retrieval strategy, e.g., user search efforts, relevance types, relevance attributes, exploitation vs. exploration balance, potential assistance for learning, etc., particularly to a '*vague*' scholarly search over a vast scientific corpus. The aim is to assist the user to find relevant documents and browse through influential ones alongside, via exploiting the QTP within matched documents.
(2). We observe that, contextual factors of traditional relevance measures (such as *tf-idf*) are common with *term-term* proximity and correlation among contextual parameters and generalization may play a key role. Hence, both (similarity and proximity) measures are amalgamate in retrieval framework.
(3). The experimental analysis validate usability of the *intuitive* information retrieval framework based on *implicit* and *explicit* relevance factors, and its feasibility for enhanced tradeoffs among *relevance*, *novelty*, and *diversity* (*coverage* and *topicality*), and overall retrieval (*precision, recall, f-measure*). The superiority of proposed framework over the baseline system is evident to balance the *exploitation-exploration* tradeoffs during information search task. Finally, the analysis confirms the implication modelling each *search-interaction* and *evolving intent estimation* for relevance manifestation.

The proposed work asserts that research effort is first in this direction and significant contributions in the research directions on the exploratory information-seeking and inclusion of additional relevance attributes/factors related to user efforts.

The organization of paper as follows: section 2 discussed the related research efforts and prospect or proposed work. Section 3 elaborates, the components of proposed work, including conceptual modelling, definition of proposed proximity measures, and algorithms. The experimental analysis and assessment is presented in section 4. Section 5 briefly summaries the finding of the proposed proximity-based relevance measures with retrieval framework for an exploratory information-search task. At last, a concluded.

Material and Methods

We emphasize that various proximity factors affects user searches manifold within matched results? The paper investigates the following questions and co-related issues of relevance measures in the context of exploratory search:

(i). When *Query term proximity* (QTP) relevance measure is useful?
(a). When the closeness among query terms is required.
(b). When the matched search results are not ordered on search preferences.
(c). When there was too many or few search results are retrieved.
(ii). In what ways does *term-term* proximity affect the overall relevance?
(a). It promotes the matched document with higher proximity in user query;
(b). It guides to a list of cohesive results over a matched document for futuristic search.

In the proposed work, *query-terms* (QTs) statistics are placed on the top of documents terms (DTs) statistics, and a novel information retrieval framework is derived. A dataset of over 50 thousands scientific publication records extracted from *DBLP, ACM*, and other sources. The dataset contains

meta-information, e.g. *title, abstract, keywords, author names, and publication year* of each paper. The aim of the work is to exploit the document-based similarity and term-based proximity to model search intent and extraction. The relevance-based measure, i.e. *algorithm relevance*, *topical relevance* and *affective relevance*, is complemented by *QTP* measures, and validated in various feasibility assessments over potential user search trails (extracted from *TREC QA 2004 Database*).

RELATED WORK

A prior research effort that relates to what we describe in this paper falls into areas: (i) *Theoretical Frameworks of Exploratory Search and Search trails*, and (ii) *Proximity based relevance manifestation*.

Exploratory Search Frameworks and Search Trails

User's information search tasks are abstract in nature, and can be made more intelligible by referring to physical activity. A framework is a conceptual structure intended to serve as a support or guide for the building of something that expands the structure into something useful. So far, the depiction of exploratory search is entirely unclear, though various real-world search behaviour exists and analogous to an exploratory search. In the recent years, various collective or collaborative search behaviours of humans or swarms are adapted into potential algorithms, e.g. *food search behaviour* of *Bees* and *Ants* are driven by foraging in its core.

Marcia Bates proposed the *berry-picking* framework (Bates, 1989), to presents a theoretical foundation for the traditional information retrieval systems. *Berry-picking* is a metaphoric construct of individual's behaviours in the selection of the berries from the bushes and analogy with information search. Though, the berries are spread on the bushes and do not come in bunches, hence a searcher chooses a pieces of information at a time, assess its relevance with search needs and follows. The principal aim of the framework is to direct the seeker towards relevant information part. The *search-thought* evolution and *focused* search are the key drivers.

Another framework is *information-foraging*, proposed (Pirolli, 1997) to illustrate the organism's physical behaviour for the sustenance. The central belief of *foraging* theory is rooted in forging behaviours of organism, exploitation of food intake over a given amount of time. The analogy with information-seeking is evident, as seekers are hunters and gaze at potential prey (information) that can be identified and accessed (Chi, Pirolli, Chen, & Pitkow, 2001; Chi, Pirolli, & Pitkow, 2000; Pirolli, Card, & Wege, 2000). The analogy leads to the development of potential search strategies via adapting collective and collaborative foraging behaviour, e.g. *Bees* and *Ant* swarms etc. The main aim is to support *information-collection* and *supervisory* during the hunt.

Users are expected to have a sense on the search goals and futuristic information trails (Zhai & Lafferty, 2017). In contrast, *exploratory browsing* portrays user-interactions in foraging of relevant sources to complete search task, e.g. implicit series of glimpses to steers closer to the results. In *information foraging* search for food sources comes first and food extraction follows. For the accomplishments of the user search, a tradeoffs in *focused search* followed *exploratory browsing* essential. As higher browsing leads to a lesser learned individual and higher *focused learning* end up with a few relevant objects.

Proximity-Based Relevance Manifestation

The primary focus of information retrieval (IR) systems has been to optimize for relevance, as existing retrieval approaches used to rank documents or evaluate IR systems do not account for "*user effort*". Most of the existing IR strategies rely on the term statistics estimated from the document to query terms, e.g. *document length* (*length*), *term frequency* (*tf*), *inverse document frequency (idf)*, etc, (Cosijn, & Ingwersen, 2000). These *document- term* (DTs) measures essentially are used to position/rank a document in the relevance order for user search (Saracevic, 2006; Barry, 1994). Though, DT measures rarely undertook the coherence aspects of the *user queries*, for example the *proximity of query terms*

(QTs) within a matched result set or document. Intuitively, the proximity of query-terms (QTs) can be exploitation addition to traditional DT measures for the manifestation of overall relevance score. The overall relevance scheme will produces semantically improved results, as QT measures promotes the scores of documents (or results set), in which the matched query terms are proximate to each other (Büttcher, Clarke, & Lushman, 2006; Schenkel, Broschart, Hwang, Theobald, & Weikum, 2007). Though, proximity heuristics has been largely under-explored in traditional models; primarily due to uncertainty on *how we can model proximity* and *how much is significant* into an existing retrieval model (Rasolofo, & Savoy, 2003; Qiao, Du, & Wan, 2017; Ye,He, Wang, & Luo, 2013).

Interestingly, the proximity measure is conceptually appealing in the user information-seeking behaviors. Indeed, several existing studies have covers the proximity aspects into seeking behavior (Keen, 1991; Keen, 1992; Beigbeder, & Mercier, 2005; Hawking, & Thistlewaite, 1995; Rasolofo & Savoy, 2003) and the outcomes are not conclusive. The studies assert the significance of proximity in query terms within matched result, but unable to establishes the clear directions for modeling a proximity information retrieval framework. The conceptualization of proximity aspects and its consequence on the overall *relevance* scheme are the two main concerns raised in the studies. The *proximity heuristic* has also been *indirectly* captured in some retrieval models through using larger indexing units than words that are derived based on term proximity (e.g., (Svore, Kanani, & Khan, 2010), but these models can only exploit proximity to a limited extent since they do not measure the proximity of terms (Huang, Kusner, Sun, Sha, & Weinberger, 2016).

With the above limitations, the proposed work explores the possibility to accommodate both relevance measures (*document-terms* and *query-terms*) into information retrieval framework. The proposed framework steers both the key drivers, i.e. *focused search* based on *document-term* similarity measures and *exploratory browsing* based on *query-term* proximity measure within matched results, eventually to personalize the support in information-seeking tasks.

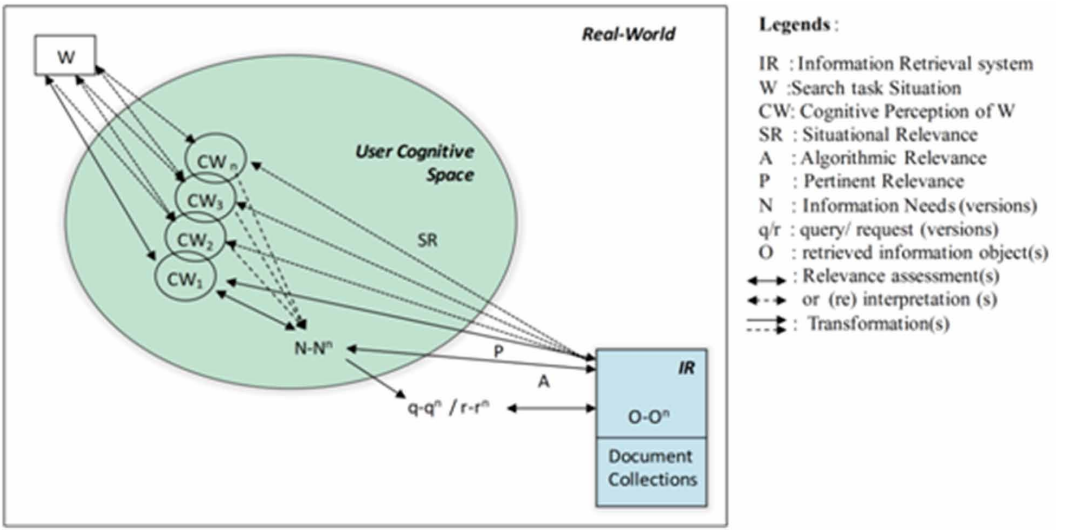
PROMOTING DOCUMENT RELEVANCE

The retrieval of relevant information requires either formal awareness understanding of complex schema and content to formulate a data retrieval request or assistance from information system (IS) in this task (Zhao, & Yun, 2009; Van Rijsbergen, 1977; Rasolofo & Savoy, 2003). A relevance measure is primarily responsible to outline relevant data objects and eventually navigate user search towards a *region-of interest*. In this paper, a novel information retrieval framework is proposed, to balance the *exploitation-exploration* in an information-seeking.

For a search task (W), the user has cognitive perceptions (CW) and equivalent abstraction as information's needs (N). Though, the cognitive version of a search task may evolve with progression of search over the time. Information needs (N) are metamorphosed into data request 'Query' (q) and often transformations ($q\text{-}q^n$), fetches respective results ($r\text{-}r^n$) from the information objects, as shown in Figure 1. The retrieval of $r\text{-}r^n$ is purely based on the system/algorithmic relevance between information objects (feature-based) and query terms.Additionally, topical relevance (T) and Pertinent relevance (P) is derived from user relevance feedback, to define a co-relation among information objects ($O\text{-}O^n$) and needs ($N\text{-}N^n$), and cognitive information need (CW) respectively.

The term-term proximity of user query within a matched document plays a pivotal role to promote the document among entire document set (Salton & Buckley, 1988; He, Huang, & Zhou, 2011; Miao, Huang, & Ye, 2012). The *QTP* capsulate the topical relevance (T), algorithmic relevance (A) and pertinence relevance (P) of user's search context. The manifestation of relevance measures is adapted to achieve a user-centric data exploration, and eventually improve focused searching (exploitation). The proximity measures with different perceptive are proposed in addition to traditional similarity based measure enhances the retrieval of potential data objects, thus to achieve a balanced exploratory browsing (exploration) of news information objects.

Figure 1. Illustrations of the overall search context, relevance types, and development of information needs (N)



Proposed Framework

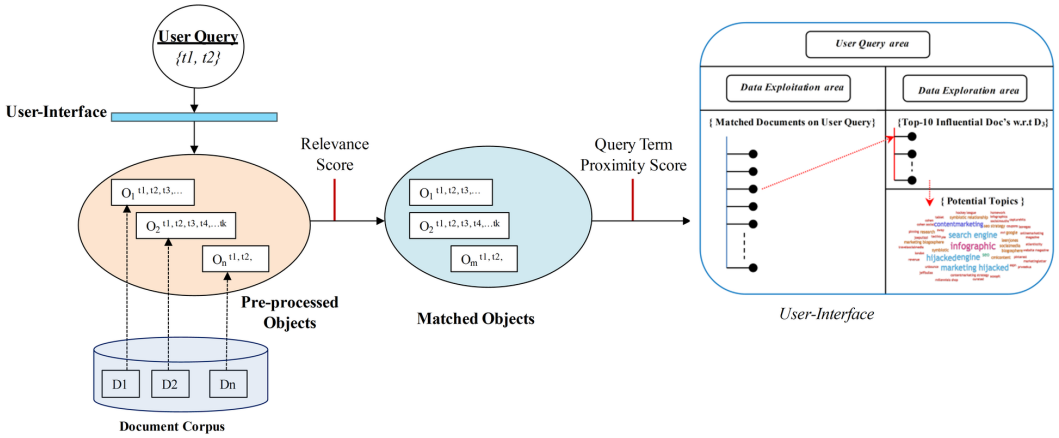
The traditional relevance measure of information retrieval are solely based on the occurrence of the query terms within a document, and implies the term-frequency (*tf*) weight to a term in a document to emphasis the relevance of the document for the user query. Similarly, inverse document frequency (*idf*) enhances the search context bubbles and extracts larger set of query results containing query terms, contrary to *tf* that, implies smaller set of query search results and from a '*local context*'. Eventually, *tf* offers results for *focused search* and *idf* for the *exploratory browsing* during the search. Both, traditional unable to capture users contextual preference in search query, i.e. *query term proximity*, *query term semantics*, etc in retrieval framework. Query term proximity (QTP) is one potential measures that incorporates query terms contextual relevance into information retrieval framework. The QTP measures are proposed to relate proximity either implicitly or explicitly, and simply evaluated using positional identifier.

The query term proximity is characterized either *implicitly* or *explicitly*, a *explicit* proximity deal with the distance between the positions of a pair of query terms within a matched document, whereas *implicit* proximity measured based on the length of a text segment covering all the query terms. A schematic view of proposed framework is illustrated in Figure 2. In the proposed retrieval framework, *implicit* proximity measures, i.e. *Query term Coverage* and *Span,* and *explicit* measures, i.e. *Minimum_Pair_Distance*, *Avg_Distance*, *Match_Distance*, and *Different_Position* are synthesized with traditional DT measures, to enhance the relevance of a matched document. The adapted definitions of each proximity measures are described in next section, with a working example.

Relevance Manifestation and Query Term Proximity (QTP) Estimation

The traditional information retrieval system exploits the term frequency and inverse document (*tf-idf*) to the extract the relevant documents (Sadakane & Imai, 1999). The presence of query term within matched document has little role to play during the document retrieval. The closeness among query terms is important in discovery-oriented information seeking as user query act as exemplar source (Borlund, 2003; Mottin, Lissandrini, Velegrakis, & Palpanas, 2014; Song, Taylor, Wen, Hon, & Yu, 2008). The affinity between query terms has important role in retrieval of relevant document and eventually specifying the position of the document among matched documents.

Figure 2. Schematic view of proposed information retrieval framework



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 proximity among two query terms characterizes the importance between two matched documents. Next, we elaborate the definitions of various query term proximity factors with formalized notions. We will use the following short document *d* as an example to explain our definitions.

$$Document\ (D) = \{t_1^1, t_2^2, t_1^3, t_3^4, t_5^5, t_4^6, t_2^7, t_3^8, t_4^9\}$$

For simplicity, the paragraph and sentence boundaries are ignored. Therefore, the positions of each term reflect the actual ordering in which the terms occur in the document. Let t_1^1 denotes position of a term t_1 in the position vector within *document D*, $Pos_{t1}^D = \{1, 3\}$, and $TF_{t1}^D = \{2\}$ denote term frequency of query term t_1. The aim is to develop measures which capture proximity information from all of the query terms.

Implicit Query Term Proximity Measures

The *implicit* proximity measured based on the length of a text segment covering all the query terms (Zhao, Huang, & Ye, 2014). In this query terms *span* and *minimum coverage* of query term are two key aspects. The adapted definitions of both measures are as presented below:

Definition 1: Query terms *Span* is estimated as *the length of the shortest segment in result/document that covers all user query terms occurrences, including repeated occurrences.* For example, in the short document d, the *Span* value is 7 for the query $\{t_1, t_2\}$.

Definition 2: Query terms *Minimum coverage* is estimated as *the shortest segment of the result/document which covers all the user query terms which are present in that document,* and formalized as below:

$$MinCoverage\ Score = \frac{\text{Query\_len}}{\text{Min\_len}} * \frac{1}{\left(\text{Diff} + 1\right)} \tag{1}$$

where, *QL* is the *length of user input query*, *Min_len* is the minimum coverage of query terms, and *Diff* is *number of query terms* which are not present in documents. For a user query $\{t_1, t_2\}$ the *MinCover* would be 2, but if the query is $\{t_1, t_2, t_4\}$, its *MinCover* would be 5.

Explicit Query Term Proximity Measures

An *explicit* proximity deals with the distance between the positions of a query term pair within matched document, hence defined between pair-wise distances between query term occurrences, and an aggregate of distances for overall proximity values (Kusner, Sun, Kolkin, & Weinberger, 2015). For a user query $\{t_1, t_2, t_3\}$ and a document with matched all terms, three different pair $\{t_1,t_2\}$, $\{t_1, t_3\}$, and $\{t_2, t_3\}$, the closest distance for each pair is 1.

Definition 3: Query terms *Minimum pair distance* is estimated as *the smallest pair distance value of all query terms pairs*, and formalized as below:

$$Min\_pair\_dist = Min(q_1,q_2 \in D, q_1 = q_2 \left\{Distance\left(q_1,q_2;D\right)\right\} \tag{2}$$

For example, the minimum pair distance of the example document *d* for query $Q = \{t_1, t_2, t_3\}$ is 1.

Definition 4: Query terms *Average distance* is estimated *as the average distance between every path of query terms for all position combinations within document.* Average distance measures sums each possible combination of distances values, to promote query terms that are consistently co-occurring, and formalized as below:

$$AverageDistance = \sum\nolimits_{l \le i \le j \le n} \frac{Freqi * Frej)}{\sum \left\| positionk - positionl \right\|} \tag{3}$$

Here, *positionk* and *positionl* indicates the k^{th} *position* of i^{th} and j^{th} query terms respectively. For a document *d* and a user query $\{t_1, t_4, t_5\}$ the average distance is $(1 + 2 + 3)/3 = 2$.

Definition 5: Query terms *Match distance* is estimated *the smallest distance achievable when each co-occurrence of a query terms is uniquely matched to another occurrence of a term,* and formalized as below:

$$MatchDist = Max\ (q_1,q_2 \in Q \cap D, q_1 = q_2 \left\{Distance\left(q_1,q_2;D\right)\right\} \tag{4}$$

Definition 6: Query terms *Different average position* is estimated as *the difference between the average positions of term$_1$ and term$_2$ within a matched document (D).* A query terms position vector is employed to measure the average position of each query terms, and then calculate the difference in terms positions.

Each proximity measure implies different perspective of user relevance via term-term of query in the retrieval of documents; therefore all the 06 proximity measures are aggregated and amalgamated into retrieval framework for the extraction of personalized documents, with enhance result diversification during information-seeking. Next, overall information retrieval framework and term score scheme is described.

Proximity Retrieval Framework

The user information seeking begins with basic information search of relevant data objects; *algorithm 1*(in Table 2) presents the adapted strategy. Table 1 lists all the mathematical notations mentioned in various algorithms, e.g. *informational search* algorithm, and overall *Relevance manifestation algorithm* and *equations*. The algorithm initiates an informational search with pre-processing of user initial query (Q_i), e.g. single term (Q_i^w), text query ($Q_i^{text,}$), phrase query (Q_i^{phrase}), via traditional measures. The text-processing measures such as, stemming, lemmatization are applied on the input query text to remove noisy words, and separate keywords of user needs. Next step is to match the relevant data objects from the corpus, based on a relevance measures. The proposed strategy employ traditional tf-idf based approach to extract initial list of *m matched documents* (*MatchedDocList* ($D_1,D_2,\ldots D_m$)) for Q^t_i *with t terms*, and further re-ranked based on *intent estimate*.

The *algorithm 2* (in Table 3) describes the overall *relevance manifestation* for proposed framework and details of relevance components, e.g. *relevance feedback, pseudo-relevance, query term proximity*, etc, to model the search intent evolutions and intermediate interactions into retrieval strategy. The extracted document will be displayed on the *user interface* (UI), and placed for the further data play. Initially, proposed strategy extract results based on *implicit* relevance measures and subsequently adapts *explicit* relevance measures for re-ranking the same list, the re-ranking eventually applied for enhanced tradeoffs among *relevance*, *diversity* and *novelty* within extracted results. The combination 'α_1*total_tf-idf+ α_2*total_proximity_score' is imposed as *implicit* relevance measures to generate the initial result to a user and then, the proposed strategy employ traditional *document-terms* (such as *tf-idf)* measure to extract initial list of m *matched documents* (*MatchedDocList* ($D_1,D_2,\ldots D_m$)) for Q^t_i *with t terms*, and further re-ranked based on intent estimate, initially results are based on the basic similarity measure and proximity measures are coupled on the top of these measures in proposed result retrieval framework. The query term proximity measurers are intended enhance relevance within matched result and balance *exploitation-exploration* tradeoffs during user information–seeking.

The proposed retrieval framework consists of two main factors: *document-term* (DT) scores and *query-term* (QT) scores. The details meta-information related to manifestation of query term proximity based relevance is given in *algorithm* 2, and for document-term relevance is as per cosine similarity (below mentioned *equation)*. For, a *document vector d* and *query vector q* with *t query terms*, the similarity is expressed as:

Table 1. Mathematical notations for algorithms and equations

Symbol	*Meaning*
Q_i	*Initial User Query*
QTP	*Query Term Proximity*
Document-term	*Terms within a document*
MatchedDocList	*List of total document matched for* Q_i
TerV	*Term vector*
posV	*Position vector of document and query terms*
w_1	*Weight coefficient of Document-term(tf-idf score)*
w_2	*Weight coefficient of Query-term (QTP_score)*
c_i	*Meta coefficient of proximity measure of* i^{th}*document*

Table 2. Proposed algorithm for user information search

Algorithm1: User Information Search	
Input: *User Initial Query* Q_i (Q_i^{t1} or $Q_i^{t1,t2,t,...,tk}$ or Q_i^{phrase}) .	
Output*: Matched Document List* (D_1, D_2,...., D_m).	
Initialize	
Submit Initial User Query Q_i,	*// UI provide alternates to apply Qi //*
one_word_query Q_i^w: $\{Q_i^{t1}\}$	*// Qi with Single term //*
text_search_query Q_i^{text}: $\{Q_i^{t1,t2,t3...tn}\}$	*//Qi with multiple terms //*
Phrase_Query Q_i^{phrase}: $\{Q_i^{phrase}\}$	*//Qi with structured text query //*
Begin	
if *Qi is a one_word_query, then continue*	
else if *Qi is phrase query*	
Preprocess Input Query, *if* $\{Q_i^{t1,t2,t3...tn}\}$ or Q_i^{phrase}	
remove (stop words, stemming, lemmatization)	
MatchedDocList= Extract (D_1,D_2,…D_n); for each Q_i,	
For each documents in *MatchedDocList* (D_1,D_2,…D_m)	// if m document matched//
$Final\_score_{doci} = (w_1 * DT\_scores) + (w_2 * QTP\_score);$	
end	
Visualize Top-K ranked Documents (D_1,D_2,…D_m) .	

$$Cosine\left(D_i, Q\right) = \frac{\sum_{j=1}^{t}\left(d_{ij}.q_i\right)}{\sqrt{\sum_{j=1}^{t} d_{ij}^{\ 2}.q_j^{\ 2}}} \quad A = \pi r^2$$

While, estimation of a query-term (QT) scores is algorithmic, as may involves several inherent query term proximity elements, e.g. *QT_Span*, *QT_coverage*, *QT_MinDist*, *QT_AvgDist*, etc, within matched documents. Each proximity measure represents different semantic rationale of query term proximity for information retrieval. The computation of each proximity value is according to definitions discussed in *section 3.2.1* and *3.2.2*, though each proximity elements is aggregated to derive a single QTP score of proposed retrieval framework.

Algorithm 2 demonstrates how proximity measure is evaluated for each matched result document for query keywords (terms). The evaluation begins with an input of total *MatchedDocList* and query term vectors. The *MatchedDocList* is generated previously on DT measures for a input query, and now each document in this list goes for relevance manifestation of query proximity. Each proximity measures definition is imposed to evaluate the equivalent value combine into *QTP_score* with a suitable weight factors. In the experimental studies confirms the feasibility synthesis of both factors into retrieval framework.

EXPERIMENTAL SETUP AND PERFORMNACE ASSESSMENT

A dataset of over 50 thousands scientific publication records extracted from *DBLP, ACM*, and other sources related to *computer science* domain. The dataset contains meta-information, e.g. *title, abstract,*

Table 3. Proposed algorithm for query term proximity estimation

Algorithm2: QTP_scores evaluation	
Input: *Query Terms* $Q_i^{(t_1,t_2,\ldots,t_n)}$ and *Matched document List* $(D_1,D_2,\ldots,D_m)$	
Output**:** *QTP_score*$(D_{1\,(Score)}, D_{2\,(Score)},\ldots, D_{m(Score)})$	
Begin	
for each document D_i in *MatchedDocList*	*// total matched documents based on document-term score//*
$QTP\_score_i = 0$,	*// j=06, as 06 different proximity measures adapted //*
for each *QTP measure j*	
$document\_score = QTP\_score\_evalution_j\,(document_i)$;	
$QTP\_score_i \mathrel{+}= c_i * document\_score;$	*//computation of each score proximity score is per definitions//*
return QTP_score	
end	

keywords, author names, and publication year of each paper. The aim of the work is to exploit the document-based similarity and query term-based proximity to model search intent and extraction. The relevance-based measure, i.e. *algorithm relevance*, *topical relevance* and *affective relevance*, is complemented by proximity measures. The performance analysis in next section aims to focus on evaluating the following goals in line with research questions specified in the introduction,

- Feasibility analysis of proposed relevance measures, e.g. *query-term statistics* (QTP) into novel *information retrieval framework*.
- Assert the impact of proposed framework for the document *promotion*, on information *novelty*, *search result diversification* (*coverage* and *topicality*) and overall *retrieval* indicators (*precision*, *recall* and *f-measure*).

In an exploratory information-seeking, a user performs searches in trails. A *search trail* (ST) describes the user's search behaviours in a context, and various search intent *meta-information*. A search trail indicates the information seeking coverage and overall knowledge acquisition, thus to investigate the overall performance 05 most frequent *user searches trails* (STs), extracted from *TREC 2004 QA* database, listed in Table 4. Each search trail consists of a chain of user queries with intermediate relevance feedbacks to accomplish the search goal. A search trail goes through multiple user-interactions, e.g., for relevance feedback, query revisions, etc. For simplicity, these search trails are presented as a sequence of keyword queries. All the experiment evaluation is performed keeping search trails in focus.

Table 4. List of search trails (STs)

Sno.	*Search Trails (STs) description (as chain of query terms:* $Q_i^{term,}\ Q_{i+1}^{term}\ldots\ldots\ Q_{i+k}^{term})$
ST_1	*{ machine learning, image processing, supervised learning }*
ST_2	*{computer vision, operating system, centralized network }*
ST_3	*{interactive modeling, user intention, interactive interface }*
ST_4	*{radio networks, cognitive radio, mobile networks, measurement of radio networks}*
ST_5	*{database management system, structured database, mysql, transaction, DB normalization}*

Overall Relevance Manifestation

The first observation is aimed to establish the significance of each *Query term proximity* scores for the retrieval of relevant and no-relevant documents. For this evaluation, the 2000 records are extracted based on *tf-idf* based weight scheme and labeled as relevant and non-relevant. Next, both *implicit QTP measures* (*Span* and *MinCoverage*) are employed in 2000 matched documents to assess the significance of each score (average of *normalized relevance score*), listed in Table 5 and Table 6. The Query terms *Span* measure shows less effective growth for *relevance*, as result under *relevant* column are lesser on most of the search trails, though *MinCover* measure is now indeed slightly smaller on relevant documents than on non-relevant documents in most cases, suggesting the existence of weak signals. Similar, exercise is done for explicit measures, listed in Table 6. The results are clearly indicative of the fact that *MaxDist* results are still non preferable, both *AvgDist* and *MinDist* are consistent; particularly *MinDist* delivers better than among *explicit* query term proximity measures.

A consolidated information retrieval performance on the user search trails on the *average precision* with aggregated *query proximity scores,* shown in figure 3. The evaluation of *precision* is as '*the fraction of relevant instances among the retrieved instances*' and performed under the document set extracted for the search trails. Figure 3 depicts the *average precision* delivers by a relevance measures, among the top-2000 document/ results.

Information Retrieval Performance

Information Novelty: To evaluate the overall *information novelty*, three aspects of search results are considers, for search trails (STs) listed in Table 4. Traditionally, the novelty described by three main factors among extracted results: *number of unique results, number of re-retrieved results, and number of useful results*. For the simplicity of experimental evaluation, numbers of result objects are clustered into three sets from complete list of first 2000 relevant result (initially related results) for each search tasks (STs), shown in Figure 4. A noteworthy point is any unique pattern is not identified from result

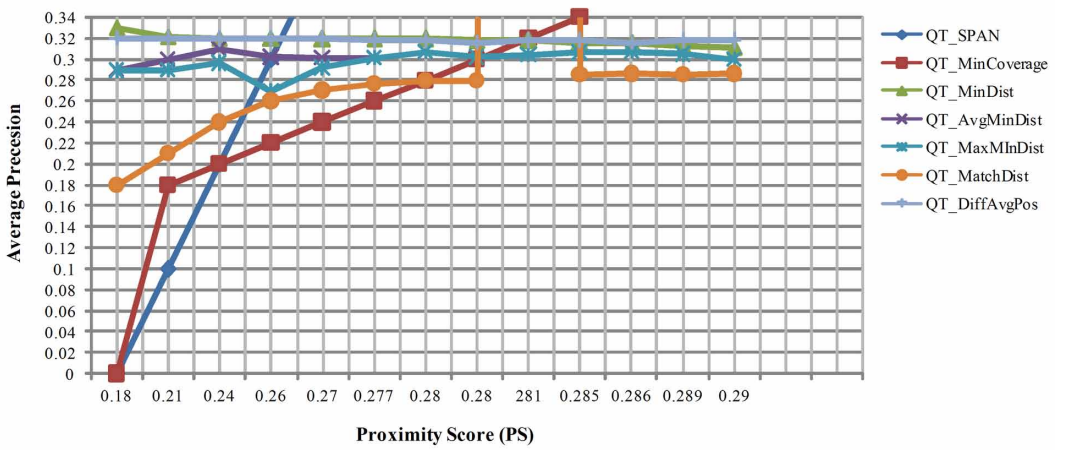
Table 5. Performance of implicit query term proximity measures

Search Tail (STs)	*Span*		*MinCoverage*	
	Relevant	**Non-relevant**	**Relevant**	**Non-relevant**
ST_1	46.43	50.78	27.63	30.9
ST_2	150.9	104.13	127.93	127.93
ST_3	57.67	56.25	20.13	20.13
ST_4	153.48	103.38	56.88	56.88
ST_5	156.398	108.38	9.857	5.857

Table 6. Performance of explicit query term proximity measures

STs	*MinDist*		*MaxDist*		*AvgDist*		*MatchDist*		*Diff_avg_pos*	
	Rel	**Non-rel**	**Rel**	**Non-rel**	**Rel**	**Non-rel**	**Rel**	**Non-rel**	**Rel**	**Non-rel**
ST_1	16.18	30.64	89.41	82.78	43.3	52.25	43.3	52.25	43.3	52.25
ST_2	39.35	39.83	415.02	133.62	148.82	72.07	148.82	72.07	148.82	72.07
ST_3	19.15	31.77	49.92	48.52	32.25	39.33	32.25	39.33	32.25	39.33
ST_4	61.15	67.91	146.92	100.42	96.65	82.73	96.65	82.73	96.65	82.73
ST_5	7.66	11.68	13.97	15.31	10.57	13.38	10.57	13.38	10.57	13.38

Figure 3. Average precision of proximity measure

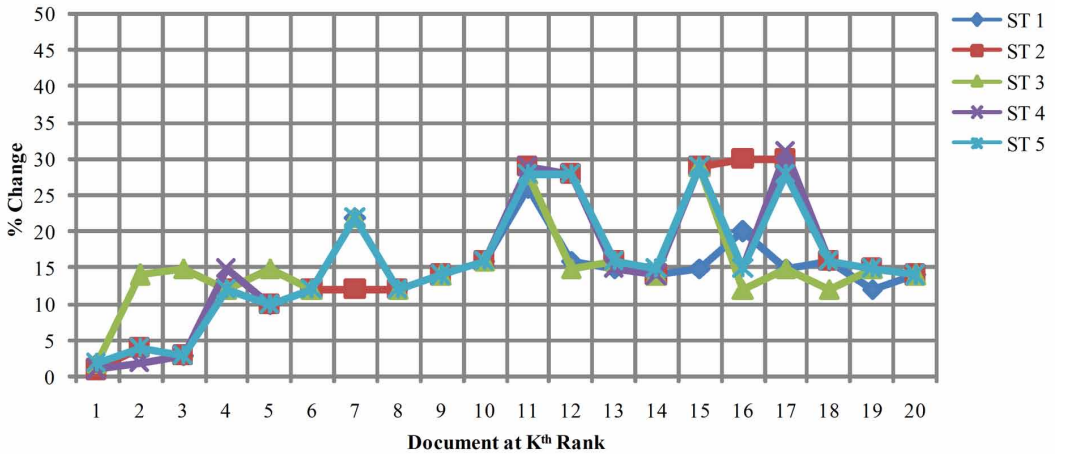


although, high degree of results common and task ST_2 lines are higher than ST_1 task, which shows that the is performing better even for un-cleaned explorations.

The *information novelty* hints at the total results for each search trails during exploratory information-seeking, for a search task with more opportunity for feedback offers higher number of results (in ST_5) and lesser intermediate relevance feedback leads to higher number of re-retrieved results (in ST_3). Traditionally, information novelty in a search task is related to, but not identical to, the homonymous information retrieval concept: a document is novel it belongs to a semantic area of interest to a person for which no documents have been seen in the recent past. A noteworthy point is any unique pattern is not identified form result although, high degree of results are common and task T_2 lines are higher than T_1 task, which shows that the is performing better even for un-cleaned explorations. Figure 4 illustrates the achieved information novelty during the search task the proposed system *information novelty.*

Simil*arly*, the performance of the proposed retrieval framework is listed, on the indicators (*Precision, Recall, and f-measure*), in Table 7. The conational definitions of each indicators is adapted

Figure 4. The novelty introduced among top-20 ranks/positions



for the simplicity of evaluation, such as *precision* is a ratio of '*relevant objects to total relevant retrieved*', *recall(sensitivity)is* ratio of *'relevant objects with total retrieved',* and *f-measure* is a *harmonic mean* of *precision* and *recall* values. The *precision* characterizes '*how useful the search results are*' and *recall 'how complete results are'* of an information search. These definitions are applied through subsequent evaluations.

Balancing Exploitation-Exploration Tradeoffs

For the proposed framework *exploitation-exploration* balance is validated though a detailed analysis on concerned parameters: *MAP,* $P_{10,}$ P_{20}, P_{100} and B_{pref} for user effort in *exploitation* and *MAR,* $R_{10,}$ R_{20}, R_{100} *and* E_{pref} for *Exploration-efforts* over search tails listed in Table 8.

The evaluation indicators *MAP,* $P_{10,}$ P_{20}, P_{100} and B_{pref} are accepted for the assessment of *focused search* efforts in an IR (Qiao, Du, & Wan, 2017). *MAP* characterizes the mean of the precision scores obtained after each relevant document is retrieval; B_{pref} is a preference-based IR measure that considers whether relevant documents are ranked above irrelevant ones. There are generally 10 search results in one page in most of IR systems, and P_{10} indicate the precision of 1st page (*as all users prefer to view page 1*); similarly P_{20} is *precision* in page 1 and page 2 (most users will click next page at least once). P_{100} means the precision in pages 1–10 (*most users will not see the pages after page 11*). The precision scores clearly indicate the significant enhanced exploitation and resultant into reduced user efforts.

Similarly, *MAR,* $R_{10,}$ R_{20}, R_{100} *and* E_{pref} adapted to evaluate exploration effects. The *Recall* measure emphasis towards retrieval of potentially relevant results additions to precisely matched result for user's current search, *MAR* is the mean recall scores after retrievals and E_{pref} is a preference-based indicator of whether how relevant documents are predicted. R_{10} means the recall score on in page 1 results; R_{20} means the recall in page 1 and page 2. Moreover, there are 20 results in one page in some IR systems; R_{100} means the precision in pages 1–10.

Estimating Optimal Weights of Relevance Components: The adapted framework amalgamates query term proximity scores and feedback components addition with document-term statistics, as in

Table 7. Overall Information retrieval performance (Precision, Recall and F1-score) of framework

User Search Tails (STs)	*Precision*	*Recall*	*F-Score*
ST_1	0.280	0.549	0.371
ST_2	0.120	0.594	0.200
ST_3	0.150	.415	0.147
ST_4	0.173	0.569	0.265
ST_5	0.047	0.251	0.079

Table 8. Performance on the exploration and exploitation aspects

STs	*Exploitation efforts*					*Exploration efforts*				
	MAP	P_{10}	P_{20}	P_{100}	B_{pref}	*MAR*	R_{10}	R_{20}	R_{100}	E_{pref}
ST_1	0.871	0.956	0.756	0.613	0.513	0.896	0.809	0.889	0.687	0.587
ST_2	0.891	0.885	0.701	0.673	0.673	0.928	0.789	0.870	0.784	0.684
ST_3	0.658	0.833	0.722	0.341	0.344	0.732	0.933	0.780	0.483	0.430
ST_4	0.779	0.660	0.862	0.475	0.389	0.897	0.760	0.964	0.729	0.799
ST_5	0.800	0.556	0.97 0	0.432	0.532	0.858	0.892	0.970	0.606	0.605

equation 5. Though balancing weights is crucial for overall tradeoffs, particularly for the % aggregate change at particular position/rank, as it signifies the effects on the overall exploration-exploitation. For simplicity of evaluation % change covers the change in precision and the recall due to updated of associated

$$FinalScore\left(Doc_i\right) = w_i{*}\mathrm{DT}_{\mathrm{score}} + w_i{*}\mathrm{QT}_{\mathrm{score}} \tag{5}$$

weights (*α1, α2, α3*) of relevance factors. To estimate the % change in the results sets, a correlation is formalize as follows:

$$\%change\left(to\ \ attain\ \ Accuracy\right) = \left(\left(Precision * Recall\right) / Precission\right) \tag{6}$$

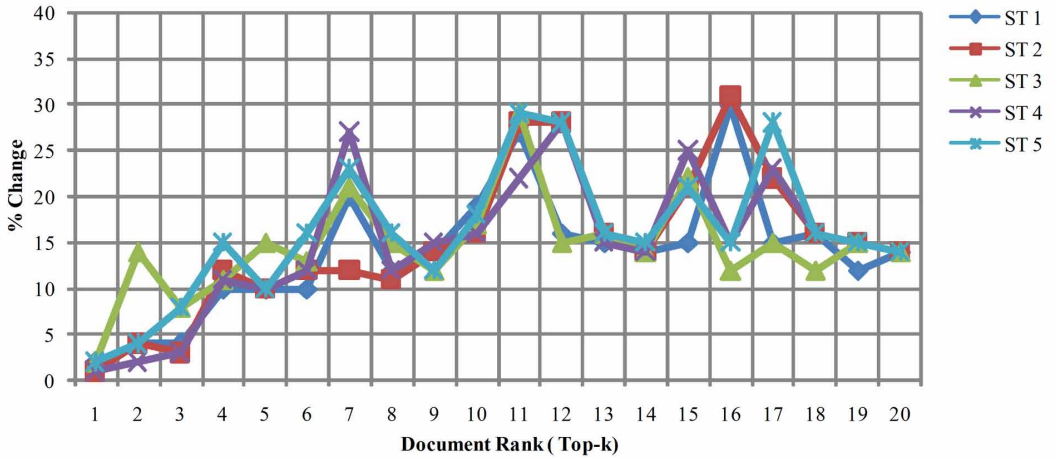
Equation 6 asserts that optimal weights for relevance factors could also affects the optimal efforts for exploration and exploitation. Figure 5, Illustrates the effects of overall change occurred in top-k positions due to varying level of weights (from 0.1 to 0.9) for each document-term (*tf-idf*) and query-term (*QTP*) based measures. Institutively, for various searches trails the change rate are uniform in rank 1-5, and rank 16-20, while at rank positions between 8-12 and 14-18 higher updating occurred for the varying multiplying factors. Intuitively, the weight combination such as 0.4 & 0.6 delivers the optimal balance between information exploitation and exploration.

*Balancing Information Exploitation-Exploration tradeoffs:*The performance indicator, such as *Precision, Recall, F-measure, accuracy, error*, etc provide quantitative estimate of performance, and could be arranged into relations for the for the simplicity of evaluation in terms of *MAP* and *MAR*. Though *MAP* and *MAR* values are intuitively complements, and correlated in a reciprocal relation, as follows:

$$trdaeoff\ \ ratio = \left(MAP * qts + MAR * qts\right) / Avg\left(MAP + MAR\right) \tag{7}$$

where, *MAP* and *MAR* is mean average *precision* and *recall* respectively, $number\ of\ feed_{ints}$ is feedback interactions on each search sessions. *Table 4* indicates that both *MAP and MAR* evaluated after each user

Figure 5. Optimal relevance weights (for favourable exploitation-exploration level)

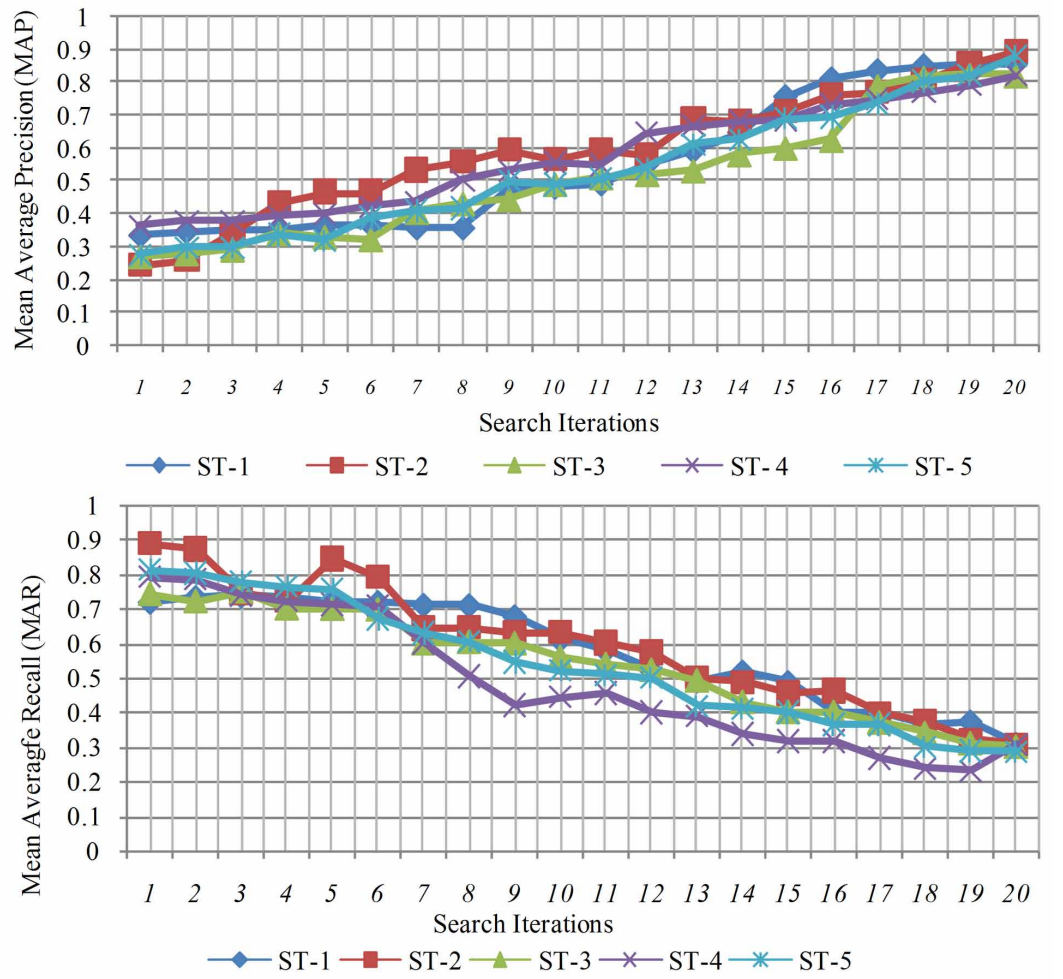


search iterations, to indicate the *growth* in user *knowledge state* in the search progression. An overall estimates of the *expanding precision* and *shrinking* area for *exploitation,* and *exploration* respectively for a search trail (ST). In Figure 6 (a), clearly depicts that whenever the user reformulate the query or impose new query the *MAP* values increased (as user *certainty* improves) and *MAR* decreases.

Search Result Diversity (% Information Coverage within Top-k results): Most of the *state-of-art* IR systems offer result re-ranking in order to achieve higher result diversifications (Selvarangam, & Kumar, 2014). For this often adapts additional relevance measures, in this paper *query term proximity* (QTP) and *users relevance feedback* measure are amalgamated into relevance framework, to promote the results relevance and eventually intent-aware results diversification. The experimental analysis also confirms that both measures steers the retrieval process and promote a result from the matched results set. In the experimental analysis, result diversification is undertaken under two different aspects: *Information Coverage* and *Information Topicality*.

Information coverage mainly affected by the % of changes (*re-ranking*) in top-k results set that implies the re-positioning of results into top-k and personalized results. A user often reviews results

Figure 6. Balancing exploration-exploitation tradeoffs:(a) via MAP (b) via MAR



places on *top-10* or *top-20* positions, due to which diversity among top query results is significant. The results diversification (on information coverage among top-k searched result is formalized as:

$$\%\ of\ \ change\left(among\ \ Top-k\ \ result\right)=\frac{1}{i}\sum_{1\le j\le i}\left|\left(\left(pos_{tf}-\frac{1}{ws}\right)-\frac{j-1}{ws}\right)\right|+1/\left(x^2\ /\ ws\right) \quad (8)$$

where, *ws* represent *the system capacity to show number of result per page*, pos_{tf} indicate the *result rank via tf-idf* based measure and x is number of relevant results. In *equation 8*, the % change (in *coverage*) characterizes the re-positioning of results (*within top-k*) during the information-seeking, mainly due to marking of user feedback relevance and query term proximity within matched document. Figure 7 illustrates the % change occurred due re-ranked positions of results among top-k results. The result established the feasibility of both *QTP-score* into retrieval framework, as steers to significant diversification (improved coverage within top-k results) .The *query term proximity* characterizes the user preference in matched results and directly imposed into relevance mechanism, eventually for a personalized extraction of information. Similarly, *adapted QTP* measures significantly improve the coverage of relevant results, revised search intents are applied in information retrieval. The peaks indicate the point of a relevance-interactions and degree of *relevance feedback*.

Information topicality, other aspects of importance of result diversification of information search. *Topicality* of retrieved result with information needs is pivotal, mainly to achieve information exploitation and intent-aware exploration. *Information topicality* often characterizes the pertinence of user intents or the material's degree of information provided; and utility, or the item's usefulness in fulfilling the information need. For the evaluation of the proposed framework, *topicality* is defined as *influence* of extracted information that implies the presence of search terms within top-k results, formalized as:

$$\%\ of\ \ influence\left(among\ \ Top-k\ \ result\right)=\frac{1}{i}\sum_{1\le j\le i}\left|\left(\left(newRelS-\frac{1}{ws}\right)-\frac{j-1}{ws}\right)\right|+1/\left(x^2\ /\ ws\right) \quad (9)$$

Where, *WS* is *number of results per page*, *NewRelS* is the relevance score via document-terms and *x* is number of relevant results., Hence, *equation 9* characterizes the relevance feedback to improve the *topicality* in information search, e.g. in ST_1 the *influence* improves 35% and 12% among *top-5* and *top-10* results respectively. This implies an assertion that with higher the search-interactions opportunity the overall informational *topicality* betters among searched data, e.g. in ST_3 relevance opportunities are high leads to better *topicalities,* also depicted in Figure 8. Other assertion related

Figure 7. Diversification within top-k results (% change among top-k rank/positions)

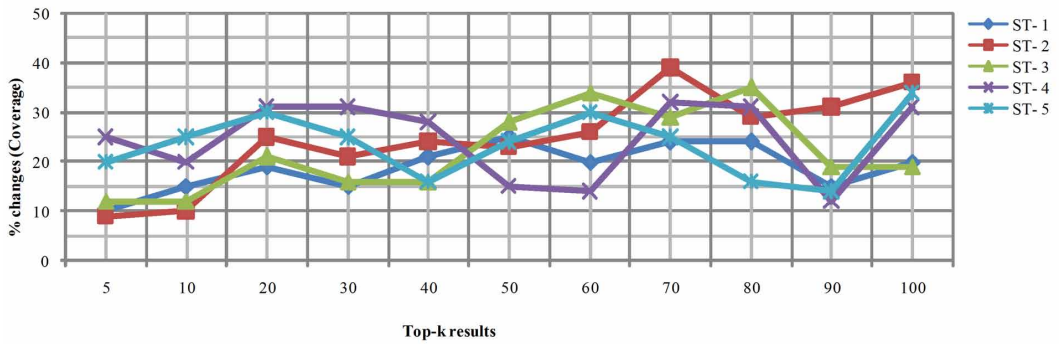
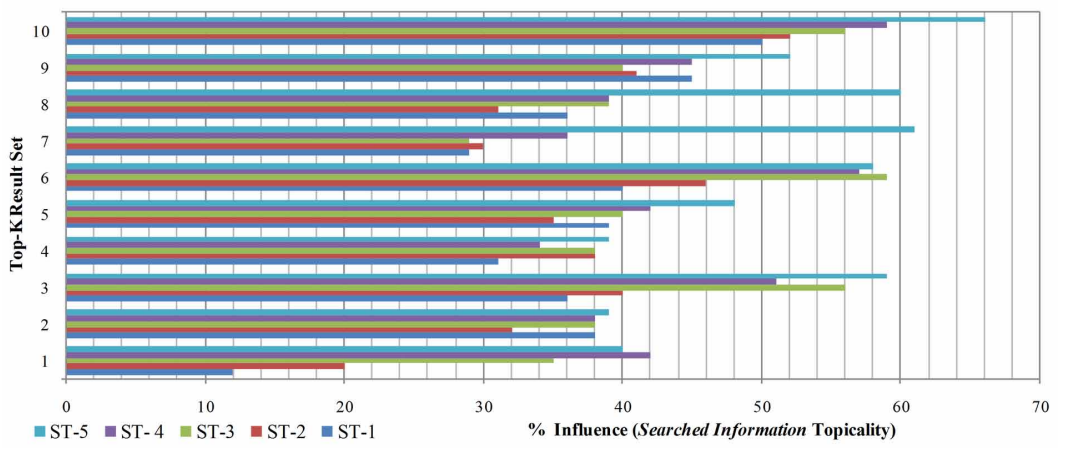


Figure 8. Information topicality (% of influence among top-k results)



to *number of terms* in user query, such as in ST_5 contains queries with comparatively higher number of *query terms* and delivers significantly higher overall *topicality* (*influence for the ST*). It is clearly inferred that both facets influence overall *result topicality* for a search and thus required to be modelled into *retrieval framework* and *intent estimate.*

DISCUSSION AND ANALYSIS

The key objective of work discussed is to incorporate a proximity measures with existing term-weighting scheme in retrieval framework. Eventually, viewed as a problem of search result *re-ranking* within top-k matched documents/objects. This also ensures that these *N documents* have an ample supply of *query-terms*. Consequently, performance analysis among the top 2000 documents from retrieval run and examines the correlation between the measures outlined in the previous section and the relevant and non-relevant documents in this set of documents. The proposed strategy provides a balance on both search behaviors, and assigns higher weights initially (initial iterations of retrieval of data) on exploration and emphasis on exploitation on later iterations. The focal point of retrieval shift towards *exploitation* of related data objects, in later stages, eventually to extract highly related objects. The assertion is based on the observation confirms the solution to *RQ 1*, as user's *uncertainty* at initial phase improves with a search progression with expanded precision among retrieved results and enhanced *exploitation circle*. Though, achieving the change in the focus is complex task and requires adaptive decision-making, on when to change the weights on both factors.

Further, additional relevance measures are adapted to enhance relevance of results in line with evolution of search intents, experiment result assert the feasibility of relevance factors. Thus validate the *RQ 2*, with improved overall retrieval performance and result diversification on both aspects: information coverage and topicality among top-k result set. The adaption of additional relevance factors also plays pivotal role, for the promotion of results objects among matched documents.

Similarly, for *RQ 3* capturing the user search interactions and intent evolution is a fruitful direction in exploratory information-seeking. The opportunity of search related interactions, e.g. *relevance feedback, query reformulations, new query insertion*, significantly affects entire search evolution and personalization of results. Further, adaption of additional relevance measures navigate the entire exploration process to real region-of-interest, as *query term proximity* (QTP) promotes the document within matched list to characterize the importance of document. The experimental analysis confirms the usability for the proximity to enhance the overall exploration and controlling the focal shift.

CONCLUSION

In this paper, document-terms (DTs) and query-terms (QTs) statistics based relevance measures are combined to steers the balanced informational search. The discussed strategy, begin with an emphasis on *exploration* (*relevant* data objects) based retrieval and less focused on *exploitation*-based retrieval (on *exact- match* data objects), though gradually reversed on progression of search. This setting is based on cognitive theory that, in beginning of search task a user is uncertain of *information needs* and unclear of precise query terms, which gradually enhances and eventually results into more specific goals. To drive the proposed strategy, particularly for the *relevance manifestation*, six different contextual proximity measures are formalized. Each measure captures different aspects of query terms proximity during *query* formulation and subsequently for score evaluation of relevant document. The QT based scores are adapted to promote a document (within all matched document), and incorporate user-preferences context in retrieval framework. The experimental assessment validates the significant growth on several tradeoffs of information search, e.g. *Information novelty,* overall *relevance*, *search result diversification* (on both aspects, *Coverage* and *Topicality*), and overall *information retrieval* (on indicators *precision*, *recall*, and *f-measure*).

The *future scope* of current work may include an *adaptive query completion* approach based on term proximities (DTs and QTs) within matched documents. The terms prediction will emulate a term/word level intents for improved exploratory information-seeking. Another, to adapt the user relevance feedback mechanism via an improved *user interface* (UI) and persuasive visualization of proximity based terms clouds and overlap among documents, to support the multi-session searches and collaborative intent modelling.

DECLARATION OF CONFLICTING INTERESTS

All authors declare that they have no conflicts of interest.

FUNDING STATEMENTS

The author(s) received no financial support for the research, authorship, and/or publication of this article.

REFERENCES

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, *45*(3), 149–159. doi:10.1002/(SICI)1097-4571(199404)45:3<149::AID-ASI5>3.0.CO;2-J

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online review, 13*(5), 407-424.

Beigbeder, M., & Mercier, A. (2005, March). An information retrieval model using the fuzzy proximity degree of term occurences. In *Proceedings of the 2005 ACM symposium on Applied computing* (pp. 1018-1022). ACM. doi:10.1145/1066677.1066912

Beigbeder, M., & Mercier, A. (2005, March). An information retrieval model using the fuzzy proximity degree of term occurences. In *Proceedings of the 2005 ACM symposium on Applied computing* (pp. 1018-1022). ACM. doi:10.1145/1066677.1066912

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, *54*(10), 913–925. doi:10.1002/asi.10286

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, *8*(3).

Büttcher, S., Clarke, C. L., & Lushman, B. (2006, August). Term proximity scoring for ad-hoc retrieval on very large text collections. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 621-622). ACM. doi:10.1145/1148170.1148285

Chi, E. H., Pirolli, P., Chen, K., & Pitkow, J. (2001, March). Using information scent to model user information needs and actions and the Web. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 490-497). ACM. doi:10.1145/365024.365325

Chi, E. H., Pirolli, P., & Pitkow, J. (2000, April). The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a web site. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*(pp. 161-168). ACM. doi:10.1145/332040.332423

Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management*, *36*(4), 533–550. doi:10.1016/S0306-4573(99)00072-2

Daoud, M., & Huang, J. X. (2013). Modeling geographic, temporal, and proximity contexts for improving geotemporal search. *Journal of the American Society for Information Science and Technology*, *64*(1), 190–212. doi:10.1002/asi.22648

Hawking, D., & Thistlewaite, P. (1995, December). Proximity operators-so near and yet so far. In *Proceedings of the 4th text retrieval conference* (pp. 131-143). ACM.

He, B., Huang, J. X., & Zhou, X. (2011). Modeling term proximity for probabilistic information retrieval models. *Information Sciences*, *181*(14), 3017–3031. doi:10.1016/j.ins.2011.03.007

Huang, G., Guo, C., Kusner, M. J., Sun, Y., Sha, F., & Weinberger, K. Q. (2016). Supervised word mover's distance. In Advances in Neural Information Processing Systems (pp. 4862-4870).

Huston, S., Culpepper, J. S., & Croft, W. B. (2014). Indexing word sequences for ranked retrieval. [TOIS]. *ACM Transactions on Information Systems*, *32*(1), 3. doi:10.1145/2559168

Idreos, S., Papaemmanouil, O., & Chaudhuri, S. (2015, May). Overview of data exploration techniques. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (pp. 277-281). ACM. doi:10.1145/2723372.2731084

Keen, E. M. (1992). Some aspects of proximity searching in text retrieval systems. *Journal of Information Science*, *18*(2), 89–98. doi:10.1177/016555159201800202

Kersten, M. L., Idreos, S., Manegold, S., & Liarou, E. (2011). The researcher's guide to the data deluge: Querying a scientific database in just a few seconds. *PVLDB Challenges and Visions, 3*(3).

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, June). From word embeddings to document distances. In *International Conference on Machine Learning* (pp. 957-966). ACM.

Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM*, *49*(4), 41–46. doi:10.1145/1121949.1121979

Miao, J., Huang, J. X., & Ye, Z. (2012, August). Proximity-based rocchio's model for pseudo relevance. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (pp. 535-544). ACM.

Michael Keen, E. (1991). The use of term position devices in ranked output experiments. *The Journal of Documentation*, *47*(1), 1–22. doi:10.1108/eb026869

Mottin, D., Lissandrini, M., Velegrakis, Y., & Palpanas, T. (2014). Exemplar queries: Give me an example of what you need. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, *7*(5), 365–376. doi:10.14778/2732269.2732273

Nandi, A., & Jagadish, H. V. (2011). Guided interaction: Rethinking the query-result paradigm. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, *4*(12), 1466–1469. doi:10.14778/3402755.3402797

Pirolli, P. (1997, March). Computational models of information scent-following in a very large browsable text collection. In CHI (Vol. 97, pp. 3-10). doi:10.1145/258549.258558

Pirolli, P., Card, S. K., & Van Der Wege, M. M. (2000, May). The effect of information scent on searching information: visualizations of large tree structures. In *Proceedings of the working conference on Advanced visual interfaces* (pp. 161-172). ACM. doi:10.1145/345513.345304

Qiao, Y. N., Du, Q., & Wan, D. F. (2017). A study on query terms proximity embedding for information retrieval. *International Journal of Distributed Sensor Networks*, *13*(2), 1550147717694891. doi:10.1177/1550147717694891

Rasolofo, Y., & Savoy, J. (2003, April). Term proximity scoring for keyword-based retrieval systems. In *European Conference on Information Retrieval* (pp. 207-218). Springer, Berlin, Heidelberg. doi:10.1007/3-540-36618-0_15

Sadakane, K., & Imai, H. (1999). Text Retrieval by using k-word Proximity Search. In *Proceedings 1999 International Symposium on Database Applications in Non-Traditional Environments (DANTE'99)(Cat. No. PR00496)* (pp. 183-188). IEEE.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. doi:10.1016/0306-4573(88)90021-0

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. doi:10.1016/0306-4573(88)90021-0

Saracevic, T. (2016). The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really? *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *8*(3), i-109. doi:10.1007/978-3-031-02302-6

Schenkel, R., Broschart, A., Hwang, S., Theobald, M., & Weikum, G. (2007, October). Efficient text proximity search. In *International Symposium on String Processing and Information Retrieval* (pp. 287-299). Springer, Berlin, Heidelberg. doi:10.1007/978-3-540-75530-2_26

Selvarangam, K., & Kumar, K. R. (2014, November). Interestingness of measures: A statistical prospective. In *2014 International Conference on Contemporary Computing and Informatics (IC3I)* (pp. 209-213). IEEE. doi:10.1109/IC3I.2014.7019800

Song, R., Taylor, M. J., Wen, J. R., Hon, H. W., & Yu, Y. (2008, March). Viewing term proximity from a different perspective. In *European Conference on Information Retrieval* (pp. 346-357). Springer, Berlin, Heidelberg. doi:10.1007/978-3-540-78646-7_32

Svore, K. M., Kanani, P. H., & Khan, N. (2010, July). How good is a span of terms?: exploiting proximity to improve web retrieval. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 154-161). ACM. doi:10.1145/1835449.1835477

Van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *The Journal of Documentation*, *33*(2), 106–119. doi:10.1108/eb026637

Van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *The Journal of Documentation*, *33*(2), 106–119. doi:10.1108/eb026637

Verma, M. (2016). Going Beyond Relevance: Incorporating Effort in Information Retrieval. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 1173-1173). ACM.

White, R. W. (2016). *Interactions with search systems*. Cambridge University Press. doi:10.1017/CBO9781139525305

White, R. W., & Roth, R. A. (2009). Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *1*(1), 1–98. doi:10.1007/978-3-031-02260-9

Ye, Z., He, B., Wang, L., & Luo, T. (2013). Utilizing term proximity for blog post retrieval. *Journal of the American Society for Information Science and Technology*, *64*(11), 2278–2298. doi:10.1002/asi.22916

Zhai, C., & Lafferty, J. (2017, August). A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Forum, 51*(2), 268-276. ACM. doi:10.1145/3130348.3130377

Zhao, J., Huang, J. X., & Ye, Z. (2014). Modeling term associations for probabilistic information retrieval. [TOIS]. *ACM Transactions on Information Systems*, *32*(2), 7. doi:10.1145/2590988

Zhao, J., & Yun, Y. (2009, July). A proximity language model for information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 291-298). ACM. doi:10.1145/1571941.1571993

Vikram Singh received his MTech in Computer Engineering from the Jawaharlal Nehru University, New Delhi, India in the year 2009. In 2009 he was hired as the Data Analytics at Maruti Suzuki India Limited, and Assistant Professor at National Institute of Technology, Kurukshetra, India in 2013 and currently pursuing Ph.D. He writes and presents widely on research issues of Exploratory Analysis, Information sciences, Human-Computer Interaction, and Data Science.