informationresearch

VOL. 21 NO. 4, DECEMBER, 2016

Moving beyond text highlights: inferring users' interests to improve the relevance of retrieval

Vimala Balakrishnan, Yasir Mehmood and Yoganathan Nagappan

Abstract

Introduction. *Studies have indicated that users' text highlighting behaviour can be further manipulated to improve the relevance of retrieved results. This article reports on a study that examined users' text highlight frequency, length and users' copy-paste actions.*
Method. *A binary voting mechanism was employed to determine the weights for the feedback, which were then used to re-rank the original search results. A search engine prototype was built using the Communications of the ACM test collection, with the well-known BM25 acting as the baseline model.*
Analysis. *The proposed enhanced model's performance was evaluated using the mean average precisions and F-score metrics, and results were compared at the top 5, 10 and 15. Additionally, comparisons were also made based on the number of terms used in a query, that is single, double and triple terms.*
Results. *The findings show that the enhanced model significantly outperformed BM25, and the rest of the models at all document levels. To be specific, the enhanced model showed significant improvements over the frequency model. Additionally, retrieval relevance was found to be the best when the query length is two.*
Conclusions. *Users' post-click behaviour may serve as a significant indicator of their interests, and thus can be used to improve the relevance of the retrieved results. Future studies could look into further extending this model by including other post-click behaviour such as printing or saving.*

change font

Introduction

Online searching has become part of many people's work and daily lives, including activities such as research, shopping and entertainment (Clay and Esparza, 2012). For example, it is common for people to: seek information from Wikipedia, search

through Google, buy products from eBay or Amazon.com, etc. However, searching for relevant items or services can be a daunting task due to the amount of information, and this is further exacerbated by a lack of searching skills. Web users tend to not know (or care) about the heterogeneity of Web content, the syntax of query languages and the art of phrasing queries, often resulting in them spending a lot of time looking for relevant items on the Internet (Manning, Raghavan and Schutze, 2009; Varathan, Tengku Sembok, Abdul Kadir and Omar, 2014).

To solve the problem of users' lack of query skills, relevance feedback is commonly used. Relevance feedback is a process involving users in the development of information retrieval systems, and aims to improve search results and increase user satisfaction. According to Baeza-Yates and Ribeiro-Neto (1999), '*in a relevance feedback cycle, the user is presented with a list of the retrieved documents and, after examining them, marks those which are relevant*'. In fact, relevance feedback has been shown to be an indicator of users' interests, which can then be used to improve their satisfaction (Claypool, Le, Wased and Brown, 2001; Fox, Karnawat, Mydland, Dumais and White, 2005; Liu, Gwizdka and Liu, 2010). Two components of relevance feedback have evolved, namely, query expansion (i.e., automatic relevance feedback) and term reweighting. Query expansion involves the addition of new terms to the initial query automatically, using techniques such as pseudo-relevance feedback (i.e., users get improved retrieval performance without further interaction), thesaurus-based or other types of expansions, with studies demonstrating the accuracy of the interpretation of original queries to improve by this technique (Belkin *et al.*, 2004; Crabtree, Andreae and Gao, 2007; Walker, Robertson, Boughanem, Jones and Sparck, 1998). In contrast, term reweighting refers to the modification of term weights according to the relevance judgement by users (Baeza-Yates and Ribeiro-Neto, 1999). In other words, this technique increases the term weights in relevant documents whilst decreasing those in irrelevant documents.

Relevance feedback is usually gathered through explicit and/or implicit feedback. Explicit feedback requires the users to provide feedback for products or services rendered, with methods including ranking, rating, commenting and answering questions (Balakrishnan, Ahmadi and Ravana, 2016; Claypool *et al.*, 2001; Núñez-Valdéz *et al.*, 2012). Explicit feedback is well understood, easy to implement and fairly precise. The approach, however, requires the users to engage in additional activities beyond their normal searching behaviour, hence resulting in higher user costs in time and effort. Additionally, not all users like to be involved in providing explicit feedback as the repetitive and frequent way of obtaining the relevance judgement causes a cognitive overload

for the users (Claypool *et al.*, 2001; Zemerli, 2012). Users are generally busy reading or looking for the right document or item during a search session, and thus they do not often provide explicit feedback. In fact, during an experiment it was found that only a small group of users agreed to give relevance judgments, and sometimes they had to be paid to provide this information (Spink, Jansen and Ozmultu, 2000). Similarly, previous work on GroupLens found that users rated many fewer documents than they read (Sarwar *et al.*, 1998). Thus, even though explicit ratings are fairly precise in recognizing user interests, their efficacy is limited.

On the other hand, implicit feedback such as mouse clicks (i.e., click-through data) can be mined unobtrusively and used to determine users' preferences (Agichtein, Brill, Dumais and Ragno, 2006; Agrawal, Halverson, Kenthapadi, Mishra and Tsaparas, 2009; Carterette and Jones, 2008; Chuklin, Markov and Rijke, 2015; Claypool *et al.*, 2001; Dupret and Liao, 2010; Feimin *et al.*, 2010; Joachims, Granka, Pan, Hembrooke and Gay, 2005; Yu, Lu, Sun, and Zhang, 2012). For instance, mouse clicks have been used for online advertising to estimate the relevance of an advertisement to a search result page or a document (Chatterjee, Hoffman and Novak, 2003; Graepel, Candela, Borchert and Herbrich, 2010).

One of the main difficulties in estimating relevance from click data is due to position bias, that is, a document appearing in a higher position is more likely to attract user clicks even though it is irrelevant. Subsequent studies on implicit feedback hence progressed to include users' post-click behaviour, which refers to users' actions on the selected documents or results. These include dwell time (i.e., reading or display time) (Balakrishnan and Zhang, 2014; Fox *et al.*, 2005; Oard and Kim, 1998), printing (Oard and Kim, 1998), scrolling (Claypool *et al.*, 2001; Guo and Agichtein, 2012; Oard and Kim, 1998), and mouse or cursor movements (Buscher, White, Dumais and Huang, 2012; Guo and Agichtein, 2010; Huang, White and, Buscher, 2012), with results indicating post-click behaviour to provide valuable implicit feedback that could indicate the relevance of selected documents. Huge volumes of implicit feedback data can be gathered easily and unobtrusively. Furthermore, no mental efforts are required from the users (Claypool *et al.*, 2001; Manning *et al.*, 2009).

One of the recent implicit feedback techniques that has been explored is text highlight or text selection, which involves selecting a block of text to indicate its relevancy to the user. Generally, people make some form of mark, such as highlights, annotations, comments, circles, etc., on documents to indicate interests or relevance (Shipman, Price, Marshall, Golovchinsky

and Schilit, 2003). Similar assumptions have been made in information retrieval studies, whereby users' annotations and text highlighting behaviour was used to improve document relevance. However, research focusing on such behaviour is scarce.

Studies of users' text highlighting behaviour thus far have examined the frequency of text highlighting, that is, it is assumed that the more text highlights a document contains, the more relevant the document is to the user (Balakrishnan and Zhang, 2014; White and Buscher, 2012). Determining a document's relevance based only on the frequency of text highlighting may be inadequate because factors such as the length of the highlighted text and users' post-selection actions may also indicate users' interests (White and Buscher, 2012). Furthermore, according to Buscher *et al.* (2012), copy-paste and reading aid were the two main reasons leading users to highlight text. In fact, both text highlighting and copying are considered to be very strong indicators of users' interests (Hauger, Paramythis, and Weibelzahl, 2011; Hijikata, 2004). Therefore, these indicators can potentially be used together to improve retrieval relevance.

The current study aims to extend the above-mentioned works by further exploring and manipulating users' text highlighting behaviour. To be precise, the study intends to improve document retrieval relevance by analysing three parameters: (i) frequency of text highlight, (ii) length of text highlight, and (iii) user's copy-paste action. The traditional ranking algorithm, Okapi BM25 was used as the baseline (i.e., without users' feedback) and Communications of the ACM (CACM) was used as the test collection. To evaluate the retrieval effectiveness of the proposed model, an experiment was conducted using a self-developed prototype search engine. The retrieved results were analysed at the top 5, 10 and 15 document levels, and also compared by query lengths. As will be shown, the findings show that the proposed model consistently yields significant improvements over BM25, and the rest of the feedback models.

Related work

Implicit user feedback can be generally divided into two categories: the query actions and the physical user reactions. The query actions refer to ways in which the user interacts with the search engine (e.g., clicks, key strokes) whereas the physical reactions are users' unconscious behaviour (e.g., eye movements, heart rate). Unlike the latter category, which requires special devices to collect data, users' query actions can be easily captured during a search session. The current study intends to exploit these query actions, specifically users' text highlighting behaviour.

Inferences drawn from implicit feedback are considered to be less reliable compared to explicit feedback, but on the other hand, large quantities of data can be gathered unobtrusively (Jung, Herlocker, and Webster, 2007). Studies focusing on implicit feedback have investigated various user behaviour, such as mouse clicks (Agichtein *et al.*, 2006; Agrawal *et al.*, 2009; Balakrishnan and Zhang, 2014; Claypool *et al.*, 2001; Dupret and Liao, 2010; Feimin *et al.*, 2010; Yu *et al.*, 2012), dwell time (Balakrishnan and Zhang, 2014; Fox *et al.*, 2005; Hassan, Jones and Klinkner, 2010; Huang, White and Dumais, 2011), eye tracking (Joachims *et al.*, 2007), and mouse movements (Buscher *et al.*, 2012; Guo and Agichtein, 2010; Huang *et al.*, 2012), among others. Techniques such as mouse clicks are based on the assumption that the clicked documents are relevant to the search queries, however this may not be accurate. Joachims *et al.* (2005) reported two main issues: trust bias in which users trusted the ranking quality of the search engine and only clicked the first few results, and quality bias which refers to users' varying behaviour for the same query in different search engines.

Generally, studies examine the search logs to understand users' behaviour and interests because they can automatically capture user interaction details. In addition, these data can be analysed to optimize retrieval performances (Jordan, Simone, Thomas, and Alexander, 2010), help query suggestions (Huanhuan *et al.*, 2008) and enhance the ranked results (Agichtein *et al.*, 2006; Balakrishnan and Zhang, 2014). More recent studies have looked into ways to improve retrieval by investigating other clicking behaviour, such as Xu, Chen, Xu, Li, and Elbio (2010) who used click rate and last click to predict the relevant labels or Uniform Resource Locators (URL). Although an overall improvement was observed, using last click as an interest indicator may not be accurate as well, because there are different reasons behind the last click. For example, users who left the last documents may have either succeeded in finding useful documents (good abandon) or failed to find relevant documents (bad abandon), and hence began a new search (Huang *et al.*, 2011).

Studies have also progressed into examining users' post-click behaviour (i.e., actions performed after clicking on a link). A simple technique would be the dwell time, whereby it is assumed that if a document is relevant, the user may spend longer time on it than other documents (Buscher, Elst, and Dengel, 2009). In a more recent study, dwell time was further experimented from three different angles, that is, display time (i.e., interval time between open and close of the document), dwell time (i.e., reading time) and decision time (i.e., decision-making time to select document), with results indicating dwell time topped the list in predicting document relevance compared to the other two

indicators (Liu and Belkin, 2010). Unfortunately, similar to mouse click problems, dwell time does not work well in all cases, because it is strongly dependent on the length of the document. In fact, it has been argued that spending more time on a page does not necessarily translate into higher user interests or relevance. In other words, the correlation between time and length cannot be directly inferred as a correlation between time and degree of interests (Zemirli, 2012). As implicit feedback techniques do not necessarily predict users' interests on their own, many studies have combined multiple implicit feedback. For instance, Claypool *et al.* (2001) found dwell time and scrolling predicted relevance in Internet browsing, Oard and Kim (1998) found dwell time and printing significantly indicated users' interests, whereas Guo and Agichtein (2012) improved document relevance using dwell time, scrolling and cursor movements.

Another post-click behaviour explored to infer users' interests is mouse or cursor movement, with the assumption that *'the more a mouse moves, the more a user is interested in the Web page or document'* (Guo and Agichtein, 2010, 2012; Huang *et al.*, 2011; Koumpouri and Simaki, 2012; Zemerli, 2012). In a study by Guo and Agichtein (2010), the authors compared cursor movements with dwell time, with findings indicating strong associations for some of the cursor features. For instance, it was observed that the further down the searcher moves the cursor, the more likely it is that s/he finds the page to be relevant. Additionally, the authors also observed lower speed of cursor movements to be indicative of *reading*, which is more likely to happen when the page is relevant. Similarly, Huang *et al.* (2011) evaluated several features of mouse movements (i.e., cursor trail length, movement time and cursor speed) on the result pages of the Microsoft Bing search engine. Their results showed that cursor movements not only improve the search results ranking, but also help query classifications.

Re-finding, i.e., users returning to the pages that have been visited previously, hence indicating interests in that page (Tyler and Teevan, 2010; Tyler, Wang, and Zhang, 2010) is another post-click activity. A page review or revisit, on the other hand, refers a user returning to the same document or item during a search session, usually by the back button. This is in contrast to re-finding, in the sense that the user does not search for the same URL using a query, browser history or bookmarked items; instead the user clicks on the same item by returning to the same search results page. Obendorf, Weinreich, Herder, and Mayer (2007) conducted a small experiment with 25 users to investigate page review behaviour. They found 31% of reviews were accomplished by the back button, and 72.6% of the reviews frequently took place in less than one hour, suggesting that page

reviews take place frequently in a Web search, especially on a short-term basis. Other user actions such as printing, copying and bookmarking can also be interpreted as implicit indicators of relevance. That is, the users performed a certain action because they are interested in the corresponding document (Guo and Agichtein, 2012; Koumpouri and Simaki, 2012; Oard and Kim, 1998). In fact, it has been revealed that scrolling time and copying text from titles and/or snippets to be the best predictors of user satisfaction (Koumpouri and Simaki, 2012). Alternatively, Bullock, Jäschke and Hotho (2011) used users' URL tags to infer their Web interests, and found that tagging data helps improve the results relevance scores.

Interestingly, very few researchers have looked into text highlights. For instance, White and Buscher (2012) conducted an experiment using Microsoft Bing search engine and a plug-in tool to record text highlighting behaviour on result pages, studying 389 queries containing text highlights. Their results showed 6% improvement in the precision compared to their baseline model for the top ten results. Similarly, Balakrishnan and Zhang (2014) compared four implicit feedback approaches, namely dwell time, click-through data, page review and text highlight. Their results revealed text highlight to have the best precisions for top ten, fifteen and twenty-five document levels compared to the other techniques, albeit with insignificant differences ($p > 0.05$). Both these studies examined text highlighting by looking at the frequency of its occurrences. It is believed that with further assessment of this particular behaviour, document relevance can be improved.

Drawing inspirations from these post-click behaviour studies (Balakrishnan and Zhang, 2014; Koumpouri and Simaki, 2012; White and Buscher, 2012), we propose to further exploit users' text highlighting behaviour by focusing not only on the frequency of highlights, but also on the length of highlights and users' copy-paste actions.

Research methods

This section provides details on the methods employed in the current study. The proposed enhanced model basically works in the following manner:

i. A user sends a query
ii. The traditional BM25 ranking algorithm returns a set of initial ranked results
iii. Users' behaviour is tracked and gathered (i.e., text highlighting and copy-paste actions). The user's interaction with the original search engine result page is monitored and recorded in a search log. The recorded user's implicit feedback (i.e., length and frequency of highlighted text, and copy-paste action) will be then weighted by a feedback

weighting scheme called the binary voting mechanism.

iv. Scores from the binary voting mechanism are fed into the re-ranking algorithm. More details are given in the following subsections.

v. A new set of improved results is displayed for the same query in the next search.

The specific details of the model are elaborated in the following sub-sections.

User behaviour

Similar to previous studies, an assumption has to be made to tune the setting of the re-ranking algorithm applicable in this study. Generally, studies on implicit feedback assume certain actions are performed because the users are interested in the particular result or document, and hence it may be of relevance. For example, Zemerli (2012), who inferred users' interests based on reading time, saving or bookmarking and printing, assumed that '*if the given document is saved or printed, this means that it has attracted the interest of the user, so it must be considered as relevant*'. White and Buscher (2012) and Balakrishnan and Zhang (2014) assumed that the more highlights a document contains (i.e., frequency), the more relevant it is to the user.

Similarly, the current study assumes the following:

Assumption 1: If a given document's content is highlighted, this means it has attracted the interest of the user, so it must be considered relevant.

A document containing more highlights is deemed to be more relevant (Balakrishnan and Zhang, 2014; White and Buscher, 2012).

- Frequency of text highlight – when a word or a portion of text is highlighted, the frequency of text highlight for the document is increased accordingly. As an example, assume a search for abstract is performed, and the user reviewed three documents as shown in Figure 1. Document three is deemed to be more relevant than documents two and one as it contains the highest frequency of highlights (i.e., eight).

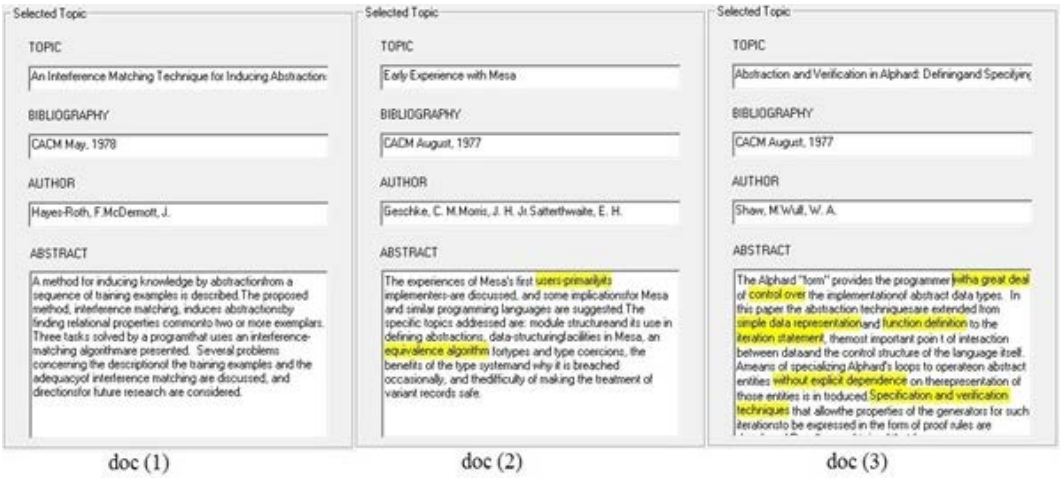

Figure 1: Frequency of text highlight

Assumption 2: A document containing longer highlighted texts is deemed to be more relevant.

- Length of text highlight – refers to the number of words that has been highlighted, with a similar assumption that the longer the highlighted text, the more relevant the document is to the user.

Assumption 3: If a copy-paste action is performed after the highlights, then it must be considered relevant

- Copy-paste – this refers to the post-selection action, in which when a user performs a copy-paste action, then the count for this variable is increased.

Baseline ranking model

Okapi BM25 is one of the most established probabilistic term weighting models used by search engines to rank matching documents to their relevance to a given search query. The model and its variants have been extensively described and evaluated in the field of information retrieval, and hence serve as a strong, reproducible baseline algorithm (Robertson and Zaragoza, 2009). It has been used in various studies as the baseline model (Bidoki, Ghodsnia, Yazdani, and Oroumchian, 2010; Dang, Bendersky, and Croft, 2013; Zhou, Liu and Zhang, 2013). Therefore, the ranking model also acts as the baseline in this study.

Re-ranking algorithm

We propose the following re-ranking algorithm, which is based on BM25 and binary voting (Equation 1).

$$Ranked\ score = BM25 + \left(1 - \left(\frac{1}{FRE+LEN+CP}\right)\right) \quad (1)$$

where

- Ranked score refers to the final score (i.e., weight) upon incorporating users' feedback
- FRE is the number of occurrences of text highlights. The value is set to a zero by default, and it is incremented by one when a user highlights a word or a block of text (i.e., when the action is performed)
- LEN indicates the length of the highlighted text. The initial value is also set to a zero. The value is however incremented by 0.1 for every fifty words, up to 500 words. In other words, the values fall within the range of zero to one. The normalization was deemed necessary to incorporate binary voting (Hopfgartner and Jose, 2007).
- CP refers to the copy-paste action. It is set to a zero by default, and increases by one whenever the user highlights a word or a block of text and chooses to copy-paste it (i.e., when the action is performed)

For a retrieved document that has not received any implicit feedback from the user, the Ranked score will be same as the original score, which is BM25. In other words, the second half of Equation 1 will not be executed if no feedback is gathered.

The weights given to each of the feedback techniques were based on the binary voting mechanism. The mechanism allows weighting terms and ranking them, and different weights can be provided for different implicit actions (Hopfgartner and Jose, 2007). For example, Hopfgartner and Jose introduced a model of six implicit interactions for interpreting a user's actions with an interactive video retrieval interface. A normalization of the features was needed which guarantees that the user feedback weights fall between 0.0 and 1.0. Therefore, the authors divided the video playing duration into 0 – 10 time cycles (i.e., 0.1 weight for each cycle), with each cycle having duration of five seconds. Similarly, White, Jose and Ruthven (2006) used binary voting mechanism to select terms implicitly for query modification. The authors assigned different weights for each part of a document, that is, 0.1 for title, 0.2 for top-ranking sentences, 0.3 for Summary, 0.2 for Summary Sentence and 0.2 for Sentence in Context. The weights were claimed to be defined for experimental purposes and were based on the typical length of a document representation.

In this study, for techniques like frequency and copy-paste, binary measurement was used (i.e., Yes=1, No=0). When a user performs these two techniques, the scores for each of them are incremented by 1. As for the length, normalization was necessary to ensure that the weights fall between 0.0 and 1.0 (Hopfgartner and Jose, 2007; White *et al.*, 2006). The study found the maximum length of the abstract in the test collection to be close to 500 words, therefore the length of the highlighted text was divided into a scale of 50 words, with an increasing weight (i.e., 0 – 50 = 0.1; 51 – 100 = 0.2; 101 – 150 = 0.3 and so on). In other

words, the weights were assigned based on our assumption that the longer the highlights, the more relevant it is, hence a higher weight. Therefore, a document containing 60 highlighted words, would be assigned a weight of 0.2 (i.e., parameter LEN) compared to a document with 50 words with weight of 0.1.

Evaluation

Test collection

The Communications of the ACM test collection contains bibliographic information (e.g., title, authors, abstract etc.) of articles published in the journal between 1958 and 1979. It contains 3204 scholarly documents, 64 sample queries, 429 stop words and also relevance judgments generated by computer scientists. The test collection also contains structured subfields, including author names, word stems from the title and abstract sections, direct references between articles and number of co-citations for each pair of articles. These structured subfields help researchers to understand the documents easily and to manage them without much hassle. The collection is small in size (approximately 2.2 Megabytes), hence it can be easily installed and tested in a very short span of time (Manning *et al.*, 2009). However, the size of collection is small compared to the size of real Web documents or the Text REtrieval Conference (TREC) document collection (Manning *et al.*, 2009). As a result, the reliability of experimental results may be lowered due to the small collection size. However, the CACM collection has been used frequently in small-scale experiments to test the performance of information retrieval systems (Balakrishnan and Zhang, 2014; Drias, 2011; Tsatsaronis, 2011), because of the limitations of time and resources for developing large-scale experiments. As our experiment is considered to be small-scale, this collection was deemed to be appropriate.

Queries

Search results can often be unsatisfying as multiple words may have similar meanings, or more than one meaning, causing results to be different than expected (Jansen, Spink, and Saracevic, 2000). For example, a user who is interested in the Apple iPhone would be laden with results referring to the fruit, phone and every other Apple product if he/she submits a query reading *Apple*. However, the search results will only contain references to the phone if the query is *Apple phone*. Query formulation is important and studies have reported that longer queries yield more accurate results in controlled experiment settings, but shorter queries are more pervasive in interactive systems, such as the Web search engines (Jansen *et al.*, 2000; Spink *et al.*, 2000). Therefore, queries were formulated based on their length to assess their impacts on retrieval relevance using

the enhanced model.

Studies particularly focusing on query lengths used automatic query expansion techniques (Crabtree *et al.*, 2007; Walker *et al.*, 1998), or encouraging users to enter longer queries in the first instance (Belkin *et al.*, 2004) or asking the users to describe their information need problems (Belkin *et al.*, 2004). As the main aim of the current study is to improve retrieval relevance, the queries used were randomly selected from the collection, and then manipulated to accommodate the various query lengths. In other words, all the queries were fixed and provided to the users. Thirty queries were selected from the document collection, targeting three different lengths, that is, single term, double terms and triple terms. The idea was to mimic user behaviour in the real Web environment, therefore the maximum length was set to three (Spink *et al.*, 2000). There were ten queries for each length, totalling to thirty final queries that were used in the experiment. Although the number may seem to be small (Voorhees, 2008, 2009), numerous other studies have reported results with fewer queries in information retrieval. For instance, Zemerli (2012) used five queries, Balakrishnan and Zhang (2014) used fifteen queries, and Belkin *et al.* (2004) used eight search topics and four search topic types.

Table 1 illustrates the queries used in our evaluations.

Number	Single	Double	Triple
1	Abstract	Abstract database	Abstract data type
2	Database	Database package	Database management system
3	English	Hashing English	Hashing English spelling
4	Hashing	Hashing method	Hashing index method
5	File	File search	Inverted file search
6	Parallel	Parallel computation	Parallel computation algorithm
7	Parallel	Parallel language	Parallel computation language
8	Prime	Prime number	Prime number computation
9	Shape	Shape description	Shape analysis reception
10	Surface	Surface algorithm	Surface algorithm implementation

Table 1: Queries used in the evaluations

Search engine prototype

The search engine prototype was developed using Visual Basic and MySQL. The interface mimics that of an online database containing scholarly documents. The following screen grabs provide illustrations of how the enhanced model works in improving the document relevance for a query.

For example, assume a user intends to perform a search for information retrieval. The initial results are ranked by BM25, and are as depicted in Figure 2. We will look at the top three documents to illustrate our example. Each of these documents had an original score of 8.99, however these weights are not displayed to the end user.

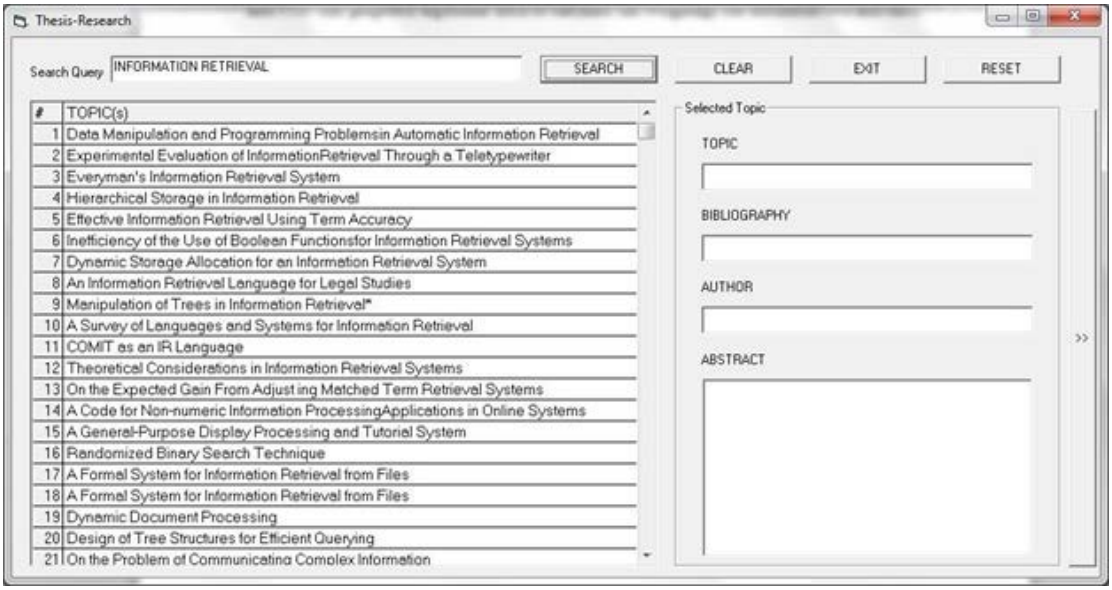

Figure 2: Initial ranked results using BM25

Also assume that the user provided no feedback for Document#1. On the other hand, Document#2 was highlighted thrice, with a total length of 11words and a single copy-paste action. Likewise, Document#3 was highlighted once with 60 words, with a single copy-paste action. Figure 3 illustrates the user's behaviour on Document#2.

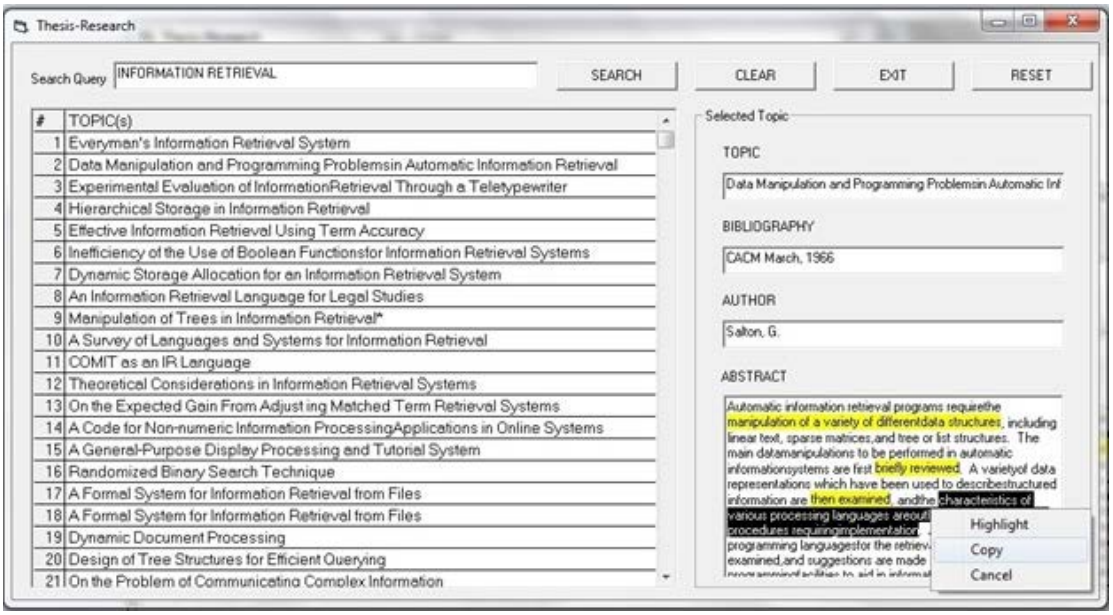

Figure 3: Sample user behaviour for Document#2

Based on the user's feedback, the binary voting produces FRE = 3, LEN = 0.1 and CP = 1 for Document#2, and FRE = 1, LEN = 0.2 and CP = 1 for Document#3. The application of Equation 1 would then produce a Ranked score of 8.99 (Document#1), 9.74 (Document#2) and 9.53 (Document#3). The documents are re-ranked based on these new weights, resulting in Document#2 to top the list, followed by Document#3 and Document#1 as show in Figure 3.

User study

Eleven students with computer science backgrounds volunteered to evaluate the proposed model. These were post-graduate students whose research work is related to information retrieval. They were identified based on their supervisors' research specialisations, and later approached by e-mail to participate in the study. Considering their technology background and field of study, these students were considered to be familiar with the concepts of information retrieval, test collections and the mechanism of the search engine.

The search engine prototype was installed in the computers in a laboratory before the experiments. The second author of this paper moderated the experiments after providing a brief demonstration on how to use the search engine. The students had no difficulties grasping the instructions, as the interface was simple to use and navigate. The set of queries was also provided to the studentsd and their behaviour was tracked as they interacted with the prototype. They were encouraged to perform the search in no particular order based on the queries provided, but they were required to perform all the searches at least once. The participants were also encouraged to repeat some of the

queries, when appropriate. For each of the queries, they were encouraged to respond to the results displayed by highlighting any portion of text that they might find relevant based on the queries executed. In addition, they were shown how relevant text can be copied and pasted into other documents, and were then asked to do the same. The students took between one and two hours to complete the tasks.

Retrieval models

Retrieval models or algorithms need to be compared based on the same set of queries (i.e., by a matched pair experiment), therefore the thirty selected queries were used to evaluate all the models, baseline included. Most of the studies in the literature also compared proposed techniques with a baseline, which is usually a model without any feedback (Ahn, Brusilovsky, He, Grady, and Li, 2008; Balakrishnan and Zhang, 2014; Ravana, Maheri and Rajapgopal, 2015; White and Buscher, 2012;). In addition, the execution of the same set of queries on the various models ensured standardization of the experiments conducted, and thus the experiments were considered to be reliable.

The models compared in the current study are as follows:

- Okapi BM25: the baseline model with no user feedback
- Frequency (FRE): a feedback model considering only the frequency of text highlight
- Length (LEN): a feedback model considering only the length of text highlight
- Copy-paste (CP): a feedback model considering only the copy-paste action
- Enhanced model (E): the proposed model integrating all the users' behaviour

Performance evaluation of a Web search engine is usually limited to the top n positions, as users are generally interested in the top few results or the first few result pages. In this study, retrieved results were measured at three levels, that is, the top 5, 10 and 15. In addition, comparisons were also based on the query lengths.

Effectiveness

The effectiveness of the models was evaluated based on two standard metrics, that is, mean average precision (MAP) and F-score. Statistically significant differences between the models were determined using pair-wise comparisons at a confidence level of 95%.

Mean average precision

The effectiveness of an information retrieval system is usually

determined based on precision (i.e., the fraction of retrieved documents that are relevant) and recall (i.e., the fraction of relevant documents that are retrieved). Although the preferred scenario would be to have these two values maximised, in most instances, precision decreases as recall increases (Beebe, Clark, Dietrich, Ko, and Ko, 2011). Mean average precision is used as overall performance indicator for retrieval algorithms. It allows easy comparisons to be made among similar researches as it carries a single value in measuring the quality of ranked results. Mean average precision has been shown to have good discrimination and stability compared to other evaluation metrics (Bidoki *et al.*, 2010; Sakai 2006). The mean average precision value can be obtained upon computing the average precision for each query, which can be determined using Equation 2 below (Manning *et al.*, 2009):

$$\text{AveP} = \frac{\sum_{k=1}^{n}(P(k) \times \text{rel}(k))}{\text{number of relevant documents}} \quad (2)$$

where n represents the number of retrieved documents, rel(k) shows if the document is relevant or not, and P(k) defines the precision at document level k

The mean average precision is then obtained by Equation 3:

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP}(q)}{Q} \quad (3)$$

where Q refers to the number of queries.

For example, assume the results of two queries are as shown in Table 2 in which the relevance column indicates if the document is relevant or not (i.e., Yes=1; No=0).

Relevant documents for query 1 = 5				**Relevant documents for query 2 = 3**			
Rank	**Relevance**	**Precision**		**Rank**	**Relevance**	**Precision**	
1	**1**	**1 / 1**	**= 1**	1	0	0/1	= 0
2	0	1/2	= 0.5	2	**1**	**1 / 2**	**= 0.5**
3	**1**	**2 / 3**	**= 0.67**	3	0	1/3	= 0.33
4	0	2/4	= 0.5	4	0	1/4	= 0.25
5	0	2/5	= 0.4	5	**1**	**2 / 5**	**= 0.4**
6	**1**	**3 / 6**	**= 0.5**	6	0	2/6	= 0.33
7	0	3/7	= 0.43	7	**1**	**3 / 7**	**= 0.43**
8	0	3/8	= 0.38	8	0	3/8	= 0.38
			=				=

9	**1**	**4 / 9**	**0.44**	9	0	3/9	0.33
10	**1**	**5 / 10**	**= 0.5**	10	0	3/10	= 0.3

Table 2: An example of mean average precision calculation

Using Equation 2, the average precision for query one is (1.0 + 0.67 + 0.5 + 0.44 + 0.5) / 5 = 0.62, whilst for query two is (0.5 + 0.4 + 0.43) / 3 = 0.44. The MAP value is finally calculated using Equation 3, which yields a value of (0.62 + 0.44) / 2 = 0.53.

F – score

The F-score combines both precision and recall into a single value, as shown in Equation 4:

$$F(j) = 2/(\frac{1}{r(j)} + \frac{1}{p(j)}) \quad (4)$$

where r(j) is the recall and p(j) is the precision for the jth document in the ranking, respectively.

Similar to mean average precision, the function F also assumes values in the interval of 0 to 1. The harmonic mean F assumes a high value only when both recall and precision are high. According to Baeza-Yates and Ribeiro-Neto (1999), the determination of the maximum value for F can be interpreted as an attempt to find the best possible compromise between recall and precision. For instance, assume the recall and precision values for Document#1 is 0.0625 and 0.5, respectively. Similarly, for Document#2 the values are 0.1875 (recall) and 0.75 (precision). The F-score for Document#1 is hence 0.11 whereas for Document#2 is 0.3, showing that higher recall and precision values would yield a higher F-score.

Results and discussion

Performance comparisons between the models

Query length	Levels	BM25	FRE	LEN	CP	E
Single	Top 5	0.4197*	0.4690*	0.4430*	0.5027	**0.5113***
	Top 10	0.3925*	0.4284	0.4083*	0.4178	**0.4366***
	Top 15	0.3491*	0.3888	0.3809	0.3878	**0.3947***
Double	Top 5	0.4177*	0.4960	0.4960	0.4783*	**0.5283***
	Top 10	0.4136*	0.4239*	0.4239*	0.4272*	**0.4686***
	Top 15	0.3856*	0.4134	0.4174	0.4035*	**0.4396***
Triple	Top 5	0.3950*	0.3270*	0.3003*	0.4410*	**0.5427***
	Top	0.3639*	0.3124*	0.2970*	0.4068*	**0.4557***

	10					
	Top 15	0.3304*	0.2897*	0.2782*	0.3615*	**0.4103***
* Significant at p < 0.05 when compared with the enhanced model (E)						

Table 3: Overall mean average precision results

Table 3 shows the mean average precision values for all the models, clearly indicating the enhanced model to produce the best retrieval results, at all three document levels regardless of the query length. In fact, pair-wise comparisons revealed overall significant improvements for the enhanced model compared to BM25. This finding was expected as generally document relevance improve when users' feedback are available (Ahn *et al.*, 2008; Balakrishnan and Zhang, 2014; Bidoki *et al.*, 2010; Buscher *et al.*, 2012; Fox *et al.*, 2005; White and Buscher, 2012). Nevertheless, it has been shown that the three parameters, that is, frequency and length of text highlight, and users' copy-paste actions, can be incorporated into a single feedback model.

A second comparison shows that the enhanced model outperformed the frequency model, although significant differences were not observed at all the levels. Therefore, it has been successfully shown that document relevance can be vastly improved with the inclusion of the length of the text highlight and users' copy-paste actions. This finding supports the notion provided by White and Buscher (2012) that further manipulations to users' text highlights might improve document relevance. In fact, our finding seems to be in line with studies that have also reported text highlight and copy-paste actions to be strong indicators of users' interests (Hauger *et al.*, 2011; Hijikata, 2004). The enhanced model also outperformed all the feedback models and thus shows that the combination of these three mechanisms can be used to infer users' interests, hence improving retrieval relevance.

A second evaluation using F-score was conducted for all the models, and the results are depicted in Table 4.

Query length	Levels	BM25	FRE	LEN	CP	E
Single	Top 5	0.2068*	0.2219	0.2298	0.2025*	**0.2312***
	Top 10	0.1574*	0.1842	0.1609	0.1649	**0.1829***
	Top 15	0.0920	0.1111	0.1093	0.1063	**0.1135***
Double	Top 5	0.2509*	0.2735	0.2832	0.2752	**0.2851***
	Top 10	0.2073*	0.2182	0.2178	0.2112	**0.2206***
	Top 15	0.1254*	0.1447	0.1438	0.1341*	**0.1516***
Triple	Top 5	0.2359*	0.2383*	0.2341*	0.2481	**0.2654***
	Top					

	10	0.1688*	0.1765*	0.1722*	0.1979	**0.2076***
	Top 15	0.1158*	0.0994*	0.0941*	0.1293	**0.1375***
* Significant at p < 0.05 when compared with the enhanced model (E)						

Table 4: Overall F-scores

Similar patterns were observed in which the enhanced model was found to produce the best scores in terms of retrieval relevance compared to the rest of the models, regardless of query length and document levels.

Overall, results produced by mean average precision and F-scores indicate that the enhanced model has better retrieval capabilities than the other models. This shows that when a user's highlighting behaviour is analysed based on the length, frequency and copy-paste action, significant improvements can be noted for the relevance of the retrieval results. Although the findings are significant, the study is limited in the sense that other similar post-click actions, such as marking (i.e., symbols, underlining etc.) or commenting, were excluded. Future studies could look into these feedback techniques and integrate them into the enhanced model. Additionally, apart from copy-paste actions, documents can also be saved and printed. It would be interesting to examine if these post-selection actions have similar effects on retrieval relevance as opposed to copy-paste actions.

Performance comparisons between query lengths

Comparisons were also made across the varying query lengths. Looking at the precision values (Table 3 and Table 4), there seems to be a pattern whereby the retrieval effectiveness peaks when the query length is two, except for the enhanced model whereby the mean average precision value has improved for the triple term. However, this observation was made only for the top five documents. The overall results generally illustrate precisions to improve when query length is two, as shown in Figures 5 and 6.

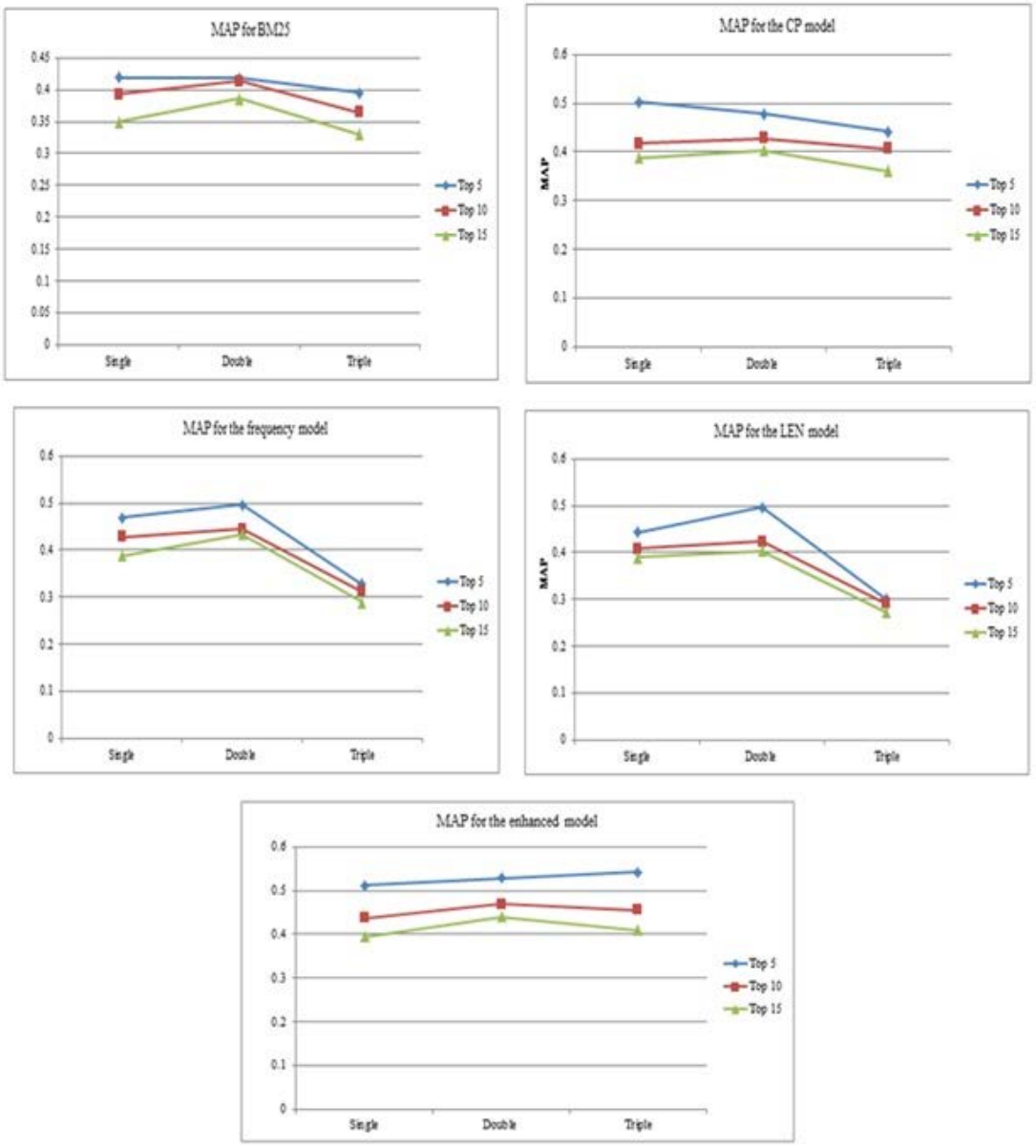

Figure 5: Mean average precision based on query lengths

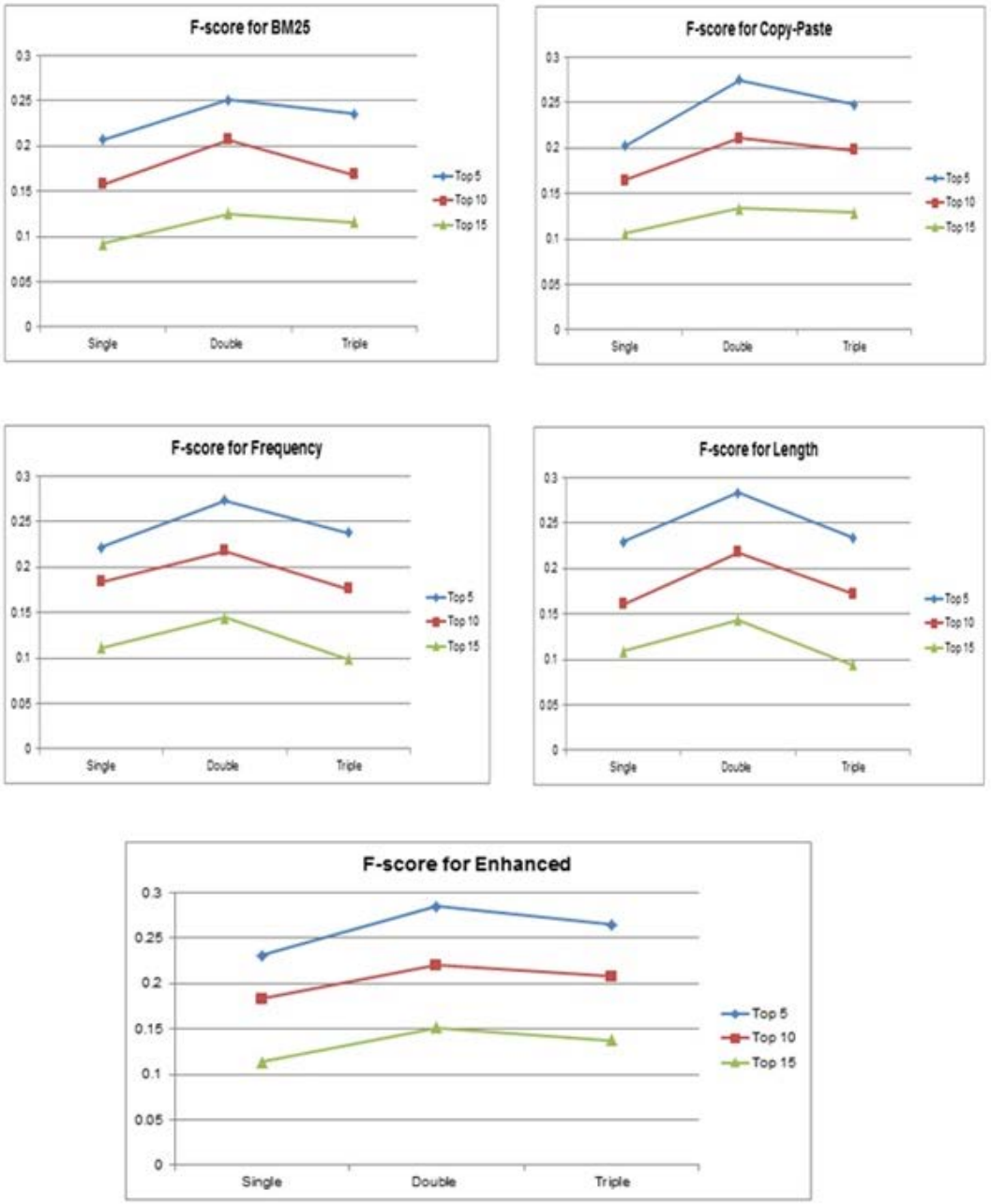

Figure 6: F-scores based on query lengths

These trends seem to indicate that the retrieval performance is the best when the query length is kept to two terms. Single term queries probably set a very strict focus, whereas longer queries tend to lose focus. However, it is noted that the current study set the maximum query length to only three. A similar finding was reflected in Hopfgartner and Jose (2007) who found their mean average precision values decreased significantly when the number of terms in the queries increased. The authors evaluated their model using 2-4-5-6 query lengths with results indicating best retrieval when the query length is two and the worst for six. It has been documented that in a traditional information retrieval environment (i.e., experiments), longer queries seem to yield better results, but this is not the case in interactive systems, such as the Web domain. In fact, the average query length has

been reported to be around 2.3 words in the Web domain (Jansen *et al.*, 2000; Spink *et al.*, 2000). This is in contrast to Belkin *et al.* (2004) who investigated the effectiveness and usability of a simple interface technique for eliciting longer queries from searchers in a Web-based information retrieval system. The authors found their users chose to enter significantly longer queries in their query elicitation system than the baseline, suggesting that when the right interface is provided to the user, s/he might enter longer queries. However, they noted that the mean query length overall was related to better performance (i.e., measured based on correctness of the answers), regardless of the interface used. In other words, longer queries were found to lead to increased performance, regardless of query elicitation mode in the interactive information retrieval systems. It is noted that as our experiment was not specifically designed to examine query length effects on retrieval relevance, our findings are to be regarded cautiously. In other words, the queries were manually tweaked and set to single, double and triple terms based on the topics provided in the ACM test collection, hence this does not really reflect the way users would perform a search.

Conclusion and limitations

Previous studies focusing on analysing users' text highlighting behaviour in terms of its frequencies were further extended in this study by taking highlight length and copy-paste action into consideration. The feedback gathered was assigned weights using the binary voting mechanism, and was then incorporated into a re-ranking algorithm. The traditional BM25 was used as the baseline model (i.e., without any user's feedback). Both mean average precision and F-scores showed that the enhanced model outperformed BM25 and the other feedback models, indicating that users' text highlighting behaviour can be further manipulated to improve retrieval relevances. In fact, the enhanced model also showed significant improvements over the frequency model. Furthermore, it was also observed that retrieval performances are best when the query length is two, a trend that is often reported in real interactive systems such as the Web search engines. In addition, the enhanced model was found to consistently outperform the baseline, and the rest of the feedback models at all the document levels, regardless of the query lengths. This consistent performance of the enhanced model shows that the proposed mechanism is reliable (Voorhees, 2008). Nevertheless, there is also a need for the enhanced model to be evaluated on other test collections, such as TREC in order to determine its performance in retrieving relevant results.

In summary, the study found improvements in document retrieval relevance which are both substantial and statistically

significant. One of the main outcomes of the study would be the inclusion and manipulation of users' post-click behaviour, namely text highlighting and copy-paste actions. Previous studies have explored the use of text highlighting, but none has focused on the length of the highlighted text. Future studies should look into the possibility of exploring users' post-click behaviour in more depth.

It is also noted that the results of this study are restricted to the experimental environment adopted. A few limitations exist, therefore future studies could look into addressing these issues. First, although ACM test collection is suitable for small-scale experiments, it would be interesting to evaluate the effectiveness of the proposed enhanced model based on other test collections, such as TREC. Although TREC increases the time and the difficulty of building information retrieval systems, it provides a huge testing document collection, search topics and relevance judgments which truly simulate the real search environment. Furthermore, it has been widely used to evaluate the performance of various retrieval systems (Belkin *et al.*, 2004; Lagun and Agichtein, 2011; Manning *et al.*, 2009; Xu *et al.*, 2010).

Second, previous studies focusing on the number of topics to be used in assessing a retrieval technique using test collections have recommended at least fifty topics (Voorheese, 2008, 2009). As the current study only used thirty, future studies could look into assessing the proposed model using a higher number of topics, for each query length.

Third, although the study found interesting results based on query lengths, further investigation is warranted. A more robust and proper mechanism is required to investigate the effect of query lengths on document retrieval relevance.

Finally, considering the nature of the experiment in which the participants were asked to perform certain actions (i.e., highlighting and copy-paste) based on the relevancy of the results, there is a possibility that user characteristics would have played a role in impacting their behaviour. For example, Shipman *et al.* (2003) differentiated between happy and meagre markers in their study, therefore future studies could further expand this study to include user characteristics as well.

Acknowledgement

The authors wish to thank University of Malaya for help supporting this project (RP028A-14AET)

About the authors

Vimala Balakrishnan is a Senior Lecturer in the Department

of Information System, Faculty of Computer Science and Information Technology, University of Malaya. Her areas of interest include data engineering, opinion mining, information retrieval and health informatics. She can be contacted at: vimala.balakrishnan@um.edu.my

Yasir Mehmood is a post graduate (Masters) student in the Department of Information System, Faculty of Computer Science and Information Technology, University of Malaya. His research interests are opinion mining, information retrieval and data analysis. He can be contacted at: yasir725@siswa.um.edu.my.

Yoganathan Nagappan is a post graduate (Masters) student in the Department of Information System, Faculty of Computer Science and Information Technology, University of Malaya. His research interests include information retrieval. He can be contacted at: yogacruise@yahoo.com.

References

Agichtein, E., Brill, E., Dumais, S. & Ragno, R. (2006). Learning user interaction models for predicting Web search result preferences. In *SIGIR '06 Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 3-10). New York, NY: ACM Press.

Agrawal, R., Halverson, A., Kenthapadi, K., Mishra, N. & Tsaparas, P. (2009). Generating labels from clicks. In *WSDM '09 Proceedings of the Second ACM International Conference on Web Search and Data Mining* (pp. 172 – 181). New York, NY: ACM Press.

Ahn, J., Brusilovsky, P., He, D., Grady, J. & Li, Q. (2008). Personalized Web exploration with task models. In *Proceeding of the 17th International Conference on World Wide Web 2008* (pp. 1 – 10). New York, NY: ACM.

Baeza-Yates, R & Ribeiro-Neto, B. (1999). *Modern information retrieval*. New York, NY: ACM Press.

Balakrishnan, V. & Zhang, X.Y. (2014). Implicit user behaviours to improve post-retrieval document relevancy. *Computers in Human Behavior, 33*, 104 -112.

Balakrishnan, V., Ahmadi, K. & Ravana, S.D. (2016). Improving retrieval relevance using users' explicit feedback. *Aslib Journal of Information Management, 68*(1), 76–98.

Beebe, N. L., Clark, J. G., Dietrich, G. B., Ko, M. S. & Ko, D. (2011). Post-retrieval search hit clustering to improve information retrieval effectiveness: two digital forensics case studies. *Decision Support Systems, 51*(4), 732–744.

Belkin, N. J., Cool., C., Kelly, D., Kim, G., Kim, J-Y., Lee, H-J., Muresan, G., Tang, M-C. & Yuan, X-J. (2004). Query length in interactive information retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '03)* (pp. 205-212). New York, NY: ACM Press.

Bidoki, Z. A. M., Ghodsnia, P., Yazdani, N. & Oroumchian, F. (2010). A3CRank: an adaptive ranking method based on

connectivity, content and click-through data. *Information Processing & Management, 46*(2), 159–169.

Bullock, B.N., Jäschke, R. & Hotho, A. (2011) Tagging data as implicit feedback for learning-to-rank. In *Proceedings of the ACM WebSci'11* (pp. 1 – 4). New York, NY: ACM Press.

Buscher, G., Elst, L. V. & Dengel, A. (2009). Segment-level display time as implicit feedback : a comparison to eye tracking. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in information retrieval* (pp. 67–74). New York, NY: ACM Press.

Buscher, G., White, R. W., Dumais, S. & Huang, J. (2012). Large-scale analysis of individual and task differences in search result page examination strategies. In *Proceedings of the fifth ACM International Conference on Web Search and Data Mining* (pp. 373 – 383). New York, NY: ACM Press.

Carterette, B. & Jones, R. (2008). Evaluating search engines by modeling the relationship between relevance and clicks. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)* (pp. 1-8). Cambridge, MA: The MIT Press.

Chatterjee, P., Hoffman, D. L. & Novak, T. P. (2003). Modeling the clickstream: implications for web-based advertising efforts. *Marketing Science, 22*(4), 520-541.

Chuklin, A., Markov, I. & Rijke, M. D. (2015). Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services, 7*(3), 1-115.

Clay, B. and Esparza, S. (2012). *Search engine optimization all in one for dummies*, 2nd Edition, Hoboken. NJ: John Wiley & Sons, Inc.

Claypool, M., Le, P., Wased, M. & Brown, D. (2001). Implicit interest indicators. In *Proceedings of the 6th International Conference on Intelligent User Interfaces* (pp. 33-40). New York, NY: ACM Press.

Crabtree, D., Andreae, P. & Gao, X. (2007). Exploiting underrepresented query aspects for automatic query expansion. In *Proceedings of the 13th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining* (pp.191–200). New York, NY: ACM Press.

Dang, V., Bendersky, M. & Croft, W. B. (2013) Two-Stage Learning to Rank for Information Retrieval. *Advances in Information Retrieval* (pp. 423–434). Berlin, Heidelberg: Springer-Verlag.

Drias, H. (2011). Web information retrieval using particle swarm optimization based approaches. In *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (pp. 36-39). Piscataway, NJ: IEEE Press.

Dupret, G. & Liao, C. (2010). A model to estimate intrinsic document relevance from the clickthrough logs of a Web search engine. In *Proceedings of the third ACM International Conference on Web Search and Data Mining* (pp. 181-190). New York, NY: ACM Press.

Feimin, Z., Dong, W., Weizhu, C., Yuchen, Z., Zheng, C. & Haixun, W. (2010). Incorporating post-click behaviors into a

click model. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR* (pp. 303-312). New York, NY: ACM Press.

Fox, S., Karnawat, K., Mydland, M., Dumais, S. & White, T. (2005). Evaluating implicit measures to improve Web search. *ACM Transactions on Information Systems, 23*(2), 147-168.

Graepel, T., Candela, J.Q., Borchert, T. & Herbrich, R. (2010). Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine. In *Proceedings of ICML* (pp. 13-20). New York, NY: ACM Press.

Guo, Q & Agichtein, E. (2012). Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior, In *Proceedings of the 21st International Conference on World Wide Web* (pp. 569–578). New York, NY: ACM Press.

Guo, Q. & Agichtein, E. (2010). Towards predicting Web searcher gaze position from mouse movements. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (pp. 3601-3606). New York, NY: ACM Press.

Hassan, A., Jones, R. & Klinkner.K.L. (2010). Beyond DCG: user behavior as a predictor of a successful. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (pp. 221-230). New York, NY: ACM Press.

Hauger, D., Paramythis, A. & Weibelzahl, S. (2011). Using browser interaction data to determine page reading behavior. In *Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization* (pp.147-158). Berlin, Heidelberg: Springer-Verlag.

Hijikata, Y. (2004). Implicit user profiling for on demand relevance feedback. In *Proceedings of the 9th International Conference on Intelligent User Interfaces* (pp.198-205). New York, NY: ACM Press.

Hopfgartner, F. & Jose, J. (2007). Evaluating the implicit feedback models for adaptive video retrieval. In *Proceedings of the International Workshop on Multimedia Information Retrieval* (pp. 323-331). New York, NY: ACM Press.

Huahuan, Cao, H-H., Daxin, J., Jian, P., Qi, H., Zhen, L., Enhong, C. & Hang, L. (2008). Context-aware query suggestion by mining click-through and session data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 875-883). New York, NY: ACM Press.

Huang, J., White, R. W. & Buscher, G. (2012). User see, user point: gaze and cursor alignment in Web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1341 – 1350). New York, NY: ACM Press.

Huang, J., White, R. W. & Dumais, S. (2011). No clicks , no problem: using cursor movements to understand and improve search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1225-1234).

New York, NY: ACM Press.

Jansen, B. J. , Spink, A. & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the Web. *Information Processes Management, 36*(2), 207-227.

Joachims, T., Granka, L., Pan, B., Hembrooke, H. & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 154- 161). New York, NY: ACM Press.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F. & Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM Transactions on Information Systems, 25*(2).

Jordan, J., Simone, D.J.B, Thomas, S. & Alexander, S. (2010). Improving the search for user interface design patterns through typed relationships. In Peter Forbrig, Fabio Paternó & Annelise Mark Pejtersen, (Eds.). *Human-Computer Interaction Second IFIP TC 13 Symposium, HCIS 2010, Held as Part of WCC 2010, Brisbane, Australia, September 20-23, 2010. Proceedings* (pp. 3-14). Berlin: Springer-Verlag. (IFIP Advances in Information and Communication Technology, Vol. 332)

Jung, S., Herlocker, J. L. & Webster, J. (2007). Click data as implicit relevance feedback in Web search. *Information Processing & Management, 43*(3), 791–807.

Koumpouri, A. & Simaki, V. (2012). Queries without clicks: evaluating retrieval effectiveness based on user feedback. In *Proceeding SIGIR '12 Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1133-1134). New York, NY: ACM Press.

Lagun, D. and Agichtein, E. (2011). ViewSer: enabling large-scale remote user studies of Web search examination and interaction categories and subject descriptors. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 365-374). New York, NY: ACM Press.

Liu, J.J. & Belkin, N. J. (2010). Personalizing information retrieval for multi-session tasks: the roles of task stage and task type. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 26-33). New York, NY: ACM Press.

Liu, C., Gwizdka, J. & Liu, J. (2010). Helping identify when users find useful documents: examination of query reformulation intervals. In *Proceedings of the Third Symposium on Information Interaction in Context* (pp. 215–224). New York, NY: ACM Press.

Liu, Y., Miao, J., Zhang, M., Ma, S. & Ru, L. (2011). How do users describe their information need: query recommendation based on snippet click model. *Expert Systems with Applications, 38*(11), 13847–13856.

Manning, C.D, Raghavan, R. & Schutze, H. (2009). *An introduction to information retrieval*. Cambridge: Cambridge University Press.

Núñez-Valdéz, E. R., Cueva Lovelle, J. M., Sanjuán Martínez, O., García-Díaz, V., Ordoñez de Pablos, P. & Montenegro Marín, C. E. (2012). Implicit feedback techniques on recommender systems applied to electronic books. *Computers in Human Behavior, 28*(4), 1186–1193.

Oard, D. W. & Kim, J. (1998). Implicit feedback for recommender systems, In *Proceedings of the AAAI Workshop on Recommender Systems* (pp. 81-83). Palo Alto, CA: AAAI Press.

Obendorf, H., Weinreich, H., Herder, E. & Mayer, M. (2007). Web page revisitation revisited: implications of a long-term click-stream study of browser usage. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 597-606). New York, NY: ACM Press.

Ravana, S.D., Taheri, M. & Rajagopal, P. (2015). Document-based approach to improve the accuracy of pairwise comparison in evaluating information retrieval systems. *Aslib Journal of Information Management, 67*(4), 408 – 421.

Robertson, S. & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond, *Foundations and Trends in Information Retrieval, 3*(4), 339 – 389

Sakai, T. (2006). On the reliability of information retrieval metrics based on graded relevance. *Information Processing and Management, 43*(2), 531–548.

Sarwar, B., Konstan, J., Borchers, A., Herlocker, J., Miller, B. & Reidl, J. (1998). Using filtering agents to improve prediction quality in the Grouplens research collaborative filtering system. In *Proceedings of ACM Conference on Computer Supported Collaborative Work (CSCW)* (pp. 345-354). New York, NY: ACM Press.

Shipman, F., Price, M., Marshall, C., Golovchinsky, G. & Schilit, B. (2003). Identifying useful passages in documents based on annotation patterns. In *Proceedings of European Conference on Digital Libraries* (pp. 101-112). Berlin, Heidelberg: Springer-Verlag.

Spink, A., Jansen, B.J. & Ozmultu, H.C. (2000). Use of query reformulation and relevance feedback by Excite users. *Internet Research, 10*(4), 317-328

Tsatsaronis, G. (2011). An experimental study on syntactic and semantic annotation in text retrieval. In *Proceedings of the fourth workshop on Exploiting Semantic Annotations in Information Retrieval* (pp. 27-28). New York, NY: ACM Press.

Tyler, S. K. & Teevan, J. (2010). Large scale query log analysis of re-finding. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (pp. 191 – 200). New York, NY: ACM Press.

Tyler, S. K., Wang, J. & Zhang, Y. (2010). Utilizing re-finding for personalized information retrieval. In *Proceedings of the 19th ACM International Conference on Information and*

Knowledge Management (pp. 1469 – 1472). New York, NY: ACM Press.

Varathan, K.D., Tengku Sembok, T.M., Abdul Kadir, R. & Omar, N. (2014). Semantic indexing in question answering systems. *Malaysian Journal of Computer Science, 27*(4), 261-274.

Voorhees, E. M. (2008) On test collections for adaptive information retrieval. *Information Process & Management, 44*(6), 1879-1885.

Voorhees, E. M. (2009). Topic set size redux. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 806-807). New York, NY: ACM Press.

Walker, S., Robertson, S.E., Boughanem, M., Jones, G.J.F., Sparck Jones, K. (1998). Okapi at TREC-6 automatic ad hoc, VLC, routing, filtering and QSDR. In E.M. Voorhees & D.K. Harmands, (Eds.). *The Sixth Text REtrieval Conference TREC-6* (pp. 125-136). Gaithersburg, MD: National Institute for Science and Technology.

White, R. W. & Buscher, G. (2012). Text selections as implicit relevance feedback. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1151 – 1152). New York, NY: ACM Press.

White, R. W., Jose, J.M. & Ruthven, I. (2006). An implicit feedback approach for interactive information retrieval. *Information Processing and Management, 42*(1), 166–190

Xu, J.F., Chen, Chuanliang, Xu, G., Li, H. & Elbio, R.T.A. (2010). Improving quality of training data for learning to rank using click-through data. In *Proceedings of the third ACM International Conference on Web Search and Data Mining* (pp. 171-180). New York, NY: ACM Press.

Yu, J.X, Lu, L., Sun, S. & Zhang, F. (2012). Search results evaluation based on user behavior, In Yuyu Yuan, Xu Wu & Yueming Lu, (Eds.). *Trustworthy Computing and Services: International Conference, ISCTCS 2012, Beijing, China, May 28 – June 2, 2012, Revised Selected Papers* (pp. 397–403). Berlin, Heidelberg: Springer-Verlag. (Communications in Computer and Information Science, Vol. 320)

Zemirli, N. (2012). WebCap: inferring the user's interests based on a real-time implicit feedback. In *Seventh International Conference on Digital Information Management (ICDIM)* (pp. 62-67). Piscataway, NJ: IEEE Press.

Zhou, D., Liu, J. & Zhang, S. (2013). Query generation techniques for patent prior-art search in multiple languages. In Guodong Zhou, Juanzi Li, Dongyan Zhao & Yansong Feng, (Eds.). *Natural Language Processing and Chinese Computing: Second CCF Conference, NLPCC 2013, Chongqing, China, November 15-19, 2013, Proceedings* (pp. 310-321). Berlin, Heidelberg: Springer-Verlag. (Communications in Computer and Information Science, Vol. 400).

Zhu, Y., He, L. & Wang, X. (2012). User interest modeling and self-adaptive update using relevance feedback technology. *Procedia Engineering, 29*, 721–725.

---

106 Last updated: 1 October, 2016

Contents | Author index | Subject index | Search | Home
