Development of Relevance Feedback System using Regression Predictive Model and TF-IDF Algorithm

Akuma, S. S. C. & Iqbal, R.

Original citation:
Akuma, SSC & Iqbal, R 2018, 'Development of Relevance Feedback System using Regression Predictive Model and TF-IDF Algorithm', Internatinoal journl of Education and Management Engineering, vol. 8, no. 4, pp. 31.
https://dx.doi.org/10.5815/ijeme.2018.04.04

DOI 10.5815/ijeme.2018.04.04
ISSN 2305-3623
ESSN 2305-8463

Publisher: MECS Press

I.J. Education and Management Engineering, **2018, 4, 31-49**
Published Online July 2018 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2018.04.04

Available online at http://www.mecs-press.net/ijeme

Development of Relevance Feedback System using Regression Predictive Model and TF-IDF Algorithm

Stephen Akuma [a], Rahat Iqbal [b]

[a] *Department of Mathematics and Computer Science, Benue State University, PMB 102119, Nigeria*
[b]*Department of Computing, Coventry University,Priory Street, CV1 5FB Coventry, UK*

Received: 10 January 2018; Accepted: 20 March 2018; Published: 08 July 2018

Abstract

Domain-specific retrieval systems developed for a homogenous group of users can potentially optimise the recommendation of relevant web documents in minimal time as compared to generic systems built for a heterogeneous group of users. Domain-specific retrieval systems are normally developed by learning from users' past interactions, as a group or individual, with an information system. This paper focuses on the recommendation of relevant web documents to a cohort of users based on their search behaviour. Simulated task situations were used to group users of the same domain. The motivation behind this work is to help a cohort of users find relevant documents that will satisfy their information needs effectively. An aggregated implicit predictive model derived from correlating implicit and explicit feedback parameters was integrated with the traditional term frequency/inverse document frequency (tf-idf) algorithm to improve the relevancy of retrieval results. The aggregated model system was evaluated in terms of recall and precision (Mean Average Precision) by comparing it with self-designed retrieval system and a generic system. The performance of the three systems was measured based on the relevant documents returned. The results showed that the aggregated domain-specific system performed better in returning relevant documents as compared to the other two systems.

Index Terms: Recommender System, Implicit feedback system, Domain-specific retrieval, information retrieval, search engine.

1. Introduction

As the volume of information on the internet explodes leading to information overload [1], finding and retrieving relevant web documents is a challenge. This problem has led to vast research on information retrieval.

* Corresponding author.
E-mail address:

The traditional IR system retrieves results based on user query input and other heuristics. The process of retrieving information in a traditional IR system begins when a user enters a query consisting of text (terms) in a search engine. The search engine measures the similarity between the terms used for the queries and the terms contained in the documents and then returns a Search Engine Result Page (SERP) relevant to the query. Considering that most users of the web usually enter queries that insufficiently capture their problem statement, additional measures are required to understand user's information needs and interest to optimise the search results.

Different relevance feedback based approaches have been explored over the years to improve search performance. The most common and consistent approach captures users' interest by asking them to explicitly suggest to the system what they think about an information that relates to their current needs [2], [3]. Such explicit statements of users' interest can be done through their preference information [4]. Although the explicit rating is the most used and consistent approach to personalization, it alters reading and browsing patterns [2], [3], [5]. Users of such systems also struggle to read more than they can rate due to the additional cognitive load of rating.

To build a robust non-intrusive feedback system, a user's perceived interest can be captured implicitly through their sequence of actions as they browse; this removes the cost of rating by the users [6]. Humans dwell and focus more on the things that are interesting, useful or relevant to their current situation [7], [8]. Information relating to user interest can be obtained from such dwelling activities [9]. Although the explicit rating is commonly used and trusted by many, it is not always reliable as presumed [2]. Thus the solution is to unobtrusively obtain users' suggestions especially in the context of learning. The advantages of the implicit feedback approach over the explicit feedback approach also include: A large amount of data can be collected unobtrusively; the user interaction with the system can be captured at any time; with implicit feedback, users need not examine and rate items, and bias in rating is eliminated through implicit method.

Previous research has studied several implicit measures for capturing user's interest [2], [3], [7], [10]-[13]. The indicators mostly studied are time spent on a document, also called reading time or dwell time, mouse movement, mouse distance, mouse clicks, the amount of scroll movement, copy and paste, printing, highlighting, emailing and bookmarking. When these implicit indicators are studied alone, they may not capture users' interest (perception of relevance) compared to when they are studied in combination [14].

This study focuses on improving document relevancy ranking for users of a particular domain by augmenting their queries with more than one implicit feedback parameters. A prototype implicit feedback system was developed and evaluated with real users. The performance of the system was compared with our Solr-indexed system (without implicit feedback) and Google (generic search system). Our findings show that our prototype system with implicit relevance feedback performed better in returning relevant documents. Although previous research has been conducted in combining implicit feedback parameters to improve post-retrieval document relevance [14], this study differs from previous work in the following ways:

1) We propose a prototype implicit feedback system to improve relevancy ranking of web documents based on aggregating weights generated from implicit feedback parameters and weights derived from the traditional tf-idf algorithm.
2) The previous studies used heuristics to model implicit feedback parameters whereas our research has used a regression predictive model which is derived from experimentation with real users.

Our hypothesis was that users will view documents that they find interesting. Their degree of interest can then be estimated by collecting and analysing their behaviour on the visited documents. This paper discusses the results of our wider endeavour towards the development of task specific search utilities [5], [15]. The remaining sections of this paper are structured as follows: Section 2 is the related work; section 3 explains the structure of the proposed system. Section 4 presents the evaluation of the prototype recommender system and section 5 is the conclusion and future work.

2. Related Work

The internet is becoming the main source of users' information [16], and millions of documents are uploaded every day on the internet. The general search engines are designed to serve all users without taking into consideration the context or domain of the users. Search engines normally crawl web documents via their Meta tags and store them in a database. When a user enters a query in the search engine text box, it returns URLs for documents matching the query [17]. Queries are however not enough to capture users' interest because the 'all purpose' search engines like Google use only keywords (query words) to rank documents [17]. Measures like visitor polarity have been employed by some user-controlled search engines to improve the relevancy ranking [17], but software has been designed by website operators to automatically increase the number of hits on their sites. There is, therefore, a need to supplement users' queries with additional sources of information obtained from their previous interaction with the system [18], [19]. Such supplementary information is explicitly obtained from users by asking them to rate documents perceived to be relevant or implicitly by their post-click behaviour. The explicit approach alters users' browsing behaviour and places a cognitive load on the user [2]. The implicit approach removes the stress of rating and can be obtained at any time. Implicit behaviour (amount of time users spent on the document, the amount of copy and so on) can be used as evidence of interest to optimize recommendation of relevant documents to users [14], [20].

Previous Adaptive Hypermedia Systems focused on producing a 'browsing agent' that will recommend relevant web resources to users through a content feedback approach. Letizia [21] was developed to track users' browsing behaviour and recommend web pages that were perceived as relevant to users based on the previous links they visit. WebACE [22] extended the operation of Letizia by capturing and building user profile with previous documents visited and the time the user spent viewing the documents. Other adaptive systems like WebMate [23] are based on explicit feedback, which is intrusive. WebMate contains a proxy that observes user's interaction with the system. It allows users to explicitly state some examples of links they are interested in, and the system learns from them. It was used for newspaper recommendation. Chandrakala et al [24] also worked on news document retrieval by using keyphrase extraction approach to optimize recommendation. WebWatcher [25] is similar to WebMate. Users of the system are asked to enter certain keywords to represent their interest and the system learns from these keywords. It also has a function for users to evaluate whether a link was useful or not, which is then used as feedback for future recommendation. LIRA [26] is another adaptive hypermedia system that explicitly seeks users' current interest and recommends to them documents relevant to their interest the next day.

These systems (Letizia, WebACE, WebMate, WebWatcher and LIRA) are generic in nature, attempting to fit all domains of interest. They thereby limit efficient recommendation of relevant documents to users of a particular domain. Contextualizing information retrieval potentially helps users to find relevant and accurate information within a minimal timeframe [27]. Context sensitive systems have been developed to improve web search. INQUIRIS2 [28] was developed as a metasearch system that asked users to explicitly state their context of interest in a given category of context. It uses the desired context along with the user query to find relevant documents in general search engines. The system proposed in this work uses queries along with implicit evidence of interest to improve the retrieval of relevant documents for a community of users. Whereas unobtrusive systems like POIROT [24], [29] uses keywords obtained from users browsing history to supplement their queries and re-rank search engine results, the proposed system uses an aggregation of implicit indicators to supplement the user query.

Kumar and Ashraf [30] proposed a framework to personalize web search based on a dynamic user profile, query expansion, user search history and collaborative filtering. They found that personalisation of web search is more efficient than a generic search engine. Researchers have worked on aggregating implicit feedback parameters from user's post-click behaviour to improve the results of search engines. Guo and Agichtein [31] studied how users interact with the Search Engine Result Page (SERP). They estimated document relevance through user scrolling and cursor activities and they found that a combination of scrolling and cursor

movements predict documents relevance more effectively than using only dwell time. In a natural setting, Buscher et al [32] used large-scale behaviour log data to examine user interactive behaviour on SERP. They clustered users based on their scrolling, clicks, cursor movement, and text highlighting behaviour. Núñez-Valdéz et al [33] reported that most of these implicit indicators can be used to improve the recommendation of electronic books.

The retrieval algorithm used by most search engines to evaluate the relevance of a web page is the Vector Space Model [34]. It retrieves information based on term similarity between the query vector and document vector. Efforts have been made to improve information retrieval by augmenting query input with user's previous interaction with the system. Zhu et al [16] applied user implicit data as a surrogate of user interest to develop a personalized information retrieval system. They used a combination of selected implicit parameters (saving, printing, favourite, viewing, click-through) to estimate user interest on documents and integrated it with the traditional search engines. Their findings suggest an improvement in terms of precision and recall of information retrieval. Some of the indicators of interest employed by Zhu et al. (2010) are not frequently used by online users. A similar method was employed by Balakrishnan and Zhang [14] to improve document search results relevancy. Balakrishnan and Zhang [14] used previous users post-click behaviour (Dwell time, click-through, text selection, page review) as an additional information source to re-rank SERP. The integrated model proposed by them was based on heuristics. Bhandari et al [35] used Quine-Mccluskey algorithm to extract knowledge from web data. They were able to discover frequent patterns but they could not find infrequent patterns. An intrusive explicit feedback study to improve retrieval relevancy was conducted by Balakrishnan et al [3]. They developed a model by integrating three explicit feedback parameters (Comment, Rating and Referral) and their findings indicate that search retrieval relevancy can be improved when users' explicit feedback is aggregated.

Prior studies have proved that consistency in user behaviour is a pre-condition for the development of such search facilities[36]. The goal of the proposed system to improve the recommendations of relevant web documents to users of a particular domain by relying on implicit user behaviour. A study was conducted and a predictive model to estimate users' interest in web documents was derived from a set of implicit indicators [5]. The predictive model was integrated with the traditional vector space model (tf-itf). Whereas previous research [3], [14] used a heuristic to assign weight to implicit and explicit indicators, this work uses a predictive model derived from the correlation of implicit and explicit feedback parameters to estimate documents relevancy and improve query result re-ranking.

3. System Structure

The structure of the proposed system for optimizing the recommendation of relevant web documents to users of a particular domain is depicted in Fig 2. The system has the following structure: Data collection, Interest scoring, Document filtration, Document re-ranking and Display results. The proposed system recommends relevant documents to users based on implicit feedback.

3.1. Data Collection

Explicit data (relevance rating, document difficulty and familiarity rating) and implicit data such as the mouse clicks, amount of copy, the mouse movement along X and Y axes, the dwell time, the mouse distance and the mouse duration count were collected unobtrusively from 77 users of computer science domain through an injected plugin in Firefox browser. The participants of the study were asked to perform a searching task for 45 minutes and their implicit and explicit feedback parameters were captured and logged. The users were asked to visit documents on the web and solve the given task (Appendix A shows the task given to the students). They explicitly rated the documents visited according to relevance. The ratings were on a 6-point relevance scale, ranging from 0 to 5 [5]. The implicit and explicit feedback parameters were correlated and an implicit predictive model was developed.

3.2. Interest scoring

A number of parameters (users' dwell time, mouse movement, mouse distance, mouse velocity, mouse clicks, amount of scroll, keystroke and amount of copy, explicit ratings) were entered into a stepwise regression for analysis. It returned an implicit predictive model comprising dwell time and amount of copy from correlating implicit and explicit feedback parameters [5]. This model (as stated below) is then used to calculate user interest level on each document and weight is assigned to the documents based on the user interest [5]. The implicit predictive model is an aggregation of dwell time and amount of copy obtained from the stepwise regression, and it estimates a user's interest level in documents. A 10-fold cross-validation was carried out to prevent overfitting.

The model is given as:

$$CIW = 2.978 + 0.281(\text{Amount of Copy}) + 0.002(\text{Dwell Time}),$$

where Amount of Copy is the number of times text that is copied to the clipboard from a document, and Dwell Time is the accumulated time in seconds spent by a user on an active page during browsing.

3.3. Document Filtration

Apache Solr technology was used to filter documents matching inputted queries. It implements the Vector Space Model (VSM) functionality of indexing, term weighing, similarity matching and scoring. Solr is an open source enterprise search platform; As part of the Apache Lucene project, it communicates with other applications through a REST-like HTTP request. The major features of Solr include real-time indexing, full-text search, faceted search, dynamic clustering, hit highlighting, rich document handling and database integration.

The vector space model (VSM) algorithm ranks similarity between documents by comparing the user query with document keyword scores. The score determines how relevant a keyword is to a document. The procedure for VSM is divided into three phases. The first phase is document indexing where the document is split into units called tokens. In the second phase, each term of the document is given a weight to enhance retrieval of relevant documents. The third phase ranks the document in relation to the query based on the similarity measure.

The vector space model uses two factors to give weight to terms in a document. The factors are Term Frequency (TF) and Inverse Document Frequency (IDF) [37]. The term frequency is a number of times a term occurs in a document. Weights are assigned to terms in a document based on the number of times they occur in the document. The TF efficiency is affected by common words like “is”, “the”, “a”, though this limitation can be overcome by the IDF, which calculates the number of documents that contain each term and reduces the weight of terms that occur in many documents. The TF-IDF scheme gives high weight to terms that occur often within a document but do not commonly occur in the collection of documents [34]. It is given as:

$$wd_t = tfd_t \times idf_t \quad (1)$$

Where wd_t is term weight t found in document d

tfd_t is the frequency factor for the term t in document d
idf_t is the IDF for term ttfd;$_t$ is Term frequency factor for t in document d

$$idf_t = \log(D/docFreq_t) \quad (2)$$

Where D is the total number of documents
idf_t is term t inverse document frequency
$docFreq_t$ is all the documents that have term t

3.3.1. How VSM determines relevant document

After documents are indexed by a search system, the documents are then ranked based on their similarity with a query. The vector space model calculates the similarity by comparing the angle of deviation between each document vector and a query vector (the query vector is the same type of vector with the documents) to rank documents according to the angle they make with the query. Practically, the cosine similarity between two vectors is calculated. The cosine coefficient calculates the angle between the query vector and the document vector by multiplying the weight of each term from the vectors and dividing each by the length of the vectors [37]. Fig. 1 is an illustration of document and query relationship on the vector space model.

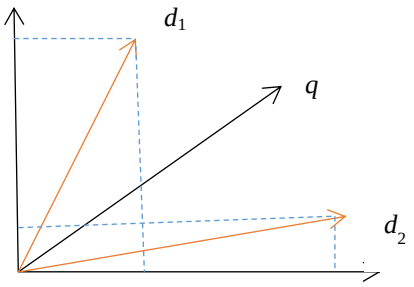

Fig.1. Illustration of Document and Query Relationship on the Vector Space Model

$$\boldsymbol{cos\theta} = \frac{\boldsymbol{d \cdot q}}{\|\boldsymbol{d}\| \, \|\boldsymbol{q}\|} \tag{3}$$

$$Sim\,(d,q) = \frac{d \cdot q}{\|d\| \, \|q\|} = \frac{\sum_{i=1}^{m} d_i q_i}{\sqrt{\sum_{i=1}^{m} d_i^2} \sqrt{\sum_{i=1}^{m} q_i^2}} \tag{4}$$

Where d is the document term vector, q is the query term vector, d.q is the intersection [38]

3.4. Aggregated Document Weight (ADW)

The aggregated document weight is the computed new weight of the documents. It combines the weight of the document derived from the predictive model and the document weight computed by the vector space model to improve document recommendation. It follows this method:

$$ADW_i = CIW_i + CVW_i \tag{5}$$

where:

CIW_i is Computed Interest Weight based on the predictive model derived as stated in section 3.2, given as:

$$CIW_i = 2.978 + 0.281(\text{Amount of Copy})_i + 0.002(\text{Dwell Time})_i \tag{6}$$

with CVW_i as the Computed Vector Weight of the original document based on TF-IDF algorithm

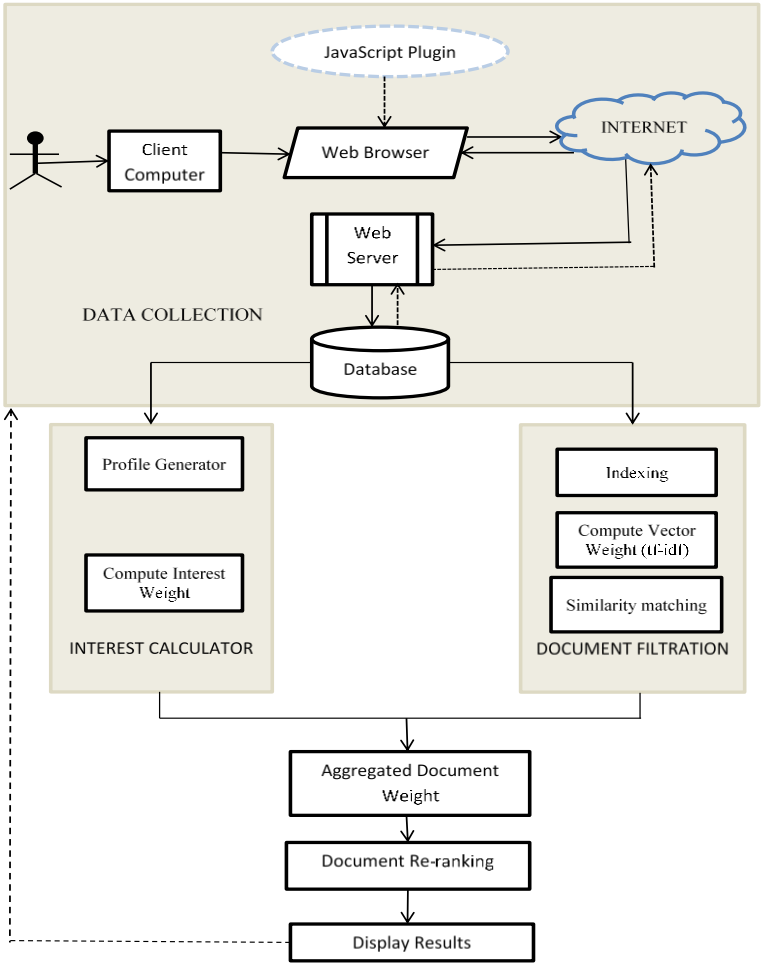

Fig.2. Conceptual Diagram Showing Aggregated Feedback System Flow

3.5. Document Re-ranking

This module sorts the documents based on the aggregated document weight in a descending order for presentation to the user. It alters the original ranking which is based on only the Computed Vector Weight (VSM score) and re-ranks the documents according to the new calculated aggregated weight. It follows this algorithm:

1: Enter user query q
2: Compute for q, the vector weight CVW_i of all documents, D in the database
3: Compute the interest weight, $CIW_i = 0.281\ (\text{Amount of Copy})_i + 0.002\ (\text{Dwell Time})_i + 2.978$ of all D
4: Considering that D = all documents, if there are common (the same) documents in the database, the mean computed interest weight (*CIW*) of the common documents is returned for the document.

5: Compute the aggregated document weight $ADW_i = CIW_i + CVW_i$ for all D
6: Re-rank original document list based on *ADW* and display result; ADW_s = (sort (ADW_i))
7: Visit current document
8: Capture implicit indicators and store in the database

3.6. Display Result

This module displays the result of the re-ranked documents. It is implemented with HTML and Laravel-5 framework of PHP and the results are presented in a descending order of relevance.

4. Evaluation of the System

This section focuses on evaluating the performance of the proposed implicit feedback system. The aim of the study was to conduct a comparative evaluation of the implicit feedback system in terms of the recall and precision. It sought to show that the quality of search results improves when queries are supplemented by users' post-click behaviour. Mean Average Precision (MAP), a popular metrics used by researchers [14], [39] in the field of information retrieval was employed for the evaluation of the systems by comparing the relevancy of documents retrieved from three systems: Google, Solr-indexed system and the aggregated system. MAP measures the efficiency of the system by computing the average number of relevant documents retrieved per query. A similar evaluation technique was used by [40], [41].

Users of the three systems were given a task brief containing instructions for the experiment and a consent form. They were given a simulated task (See Appendix A) to visit web documents and explicitly rate the documents according to how relevant they are to the given task. The ratings were on a 6-point rating scale (5 - means very relevant, 4 - means more relevant, 3 - means moderate relevant, 2 - means slightly relevant, 1 - means very low relevance, 0 - means not relevant) and were attached to the web documents via a JavaScript plugin that was embedded in a Firefox browser. The participants could enter a single query (keywords) of their choice to search for documents that are relevant to the task under consideration. Altogether, the participants entered a total number of 26 queries. They were also asked to rate the first top 10 documents on a six-point scale according to how relevant they were to the given task. They rated each document immediately after leaving the page. The six-point ratings of the users were then merged into binary form for analysis. Ratings for 0, 1 and 2 were merged as 0 and labelled as non-relevant while the rating of 3, 4 and 5 were merged as 1 and labelled as relevant.

Although Text REtrieval Conference (TREC) evaluation uses expert judges to judge the relevance of documents, such relevance ratings are inherently noisy due to the variability of the experts' behaviour [42]. Also, getting experts to judge each of the documents used for this research in relation to the task given was not feasible. This work considers relevance judgement to be subjective to the user accessing the web documents in relation to the current task [43]. The user relevance rating was used for evaluating the effectiveness of the implicit feedback system in terms of precision, recall and mean average precision.

4.1. Experimental Setup

Twenty-six students in the Faculty of Engineering, Environment and Computing at Coventry University participated in the evaluation study for a duration of 30 minutes. Two approaches were employed to conduct the study. In the first approach (Approach 1), 15 users out of the 26 participated while the remaining 11 users participated in the second approach (Approach 2). The following three systems were used for the evaluation:

1. ***Baseline System:*** Google was the baseline system because it is generic and non-domain-specific. Documents relating to user queries are returned based on Google.
2. ***Controlled system (Solr-Indexed system):*** The system was designed to return documents that were

related to the user query. The controlled system had only the solr-indexed tf-idf algorithm with no implicit feedback as shown in Fig 3.

3. ***Experimental system (Aggregated system):*** The implicit model was integrated into the system so that documents relating to the input query are re-ranked according to the degree of user interest. The degree of user interest is estimated using the implicit model derived. The experimental system re-ranks the documents according to the aggregated document weight, which is a combination of the computed interest weight (*CIW*) and computed vector weight (*CVW*).

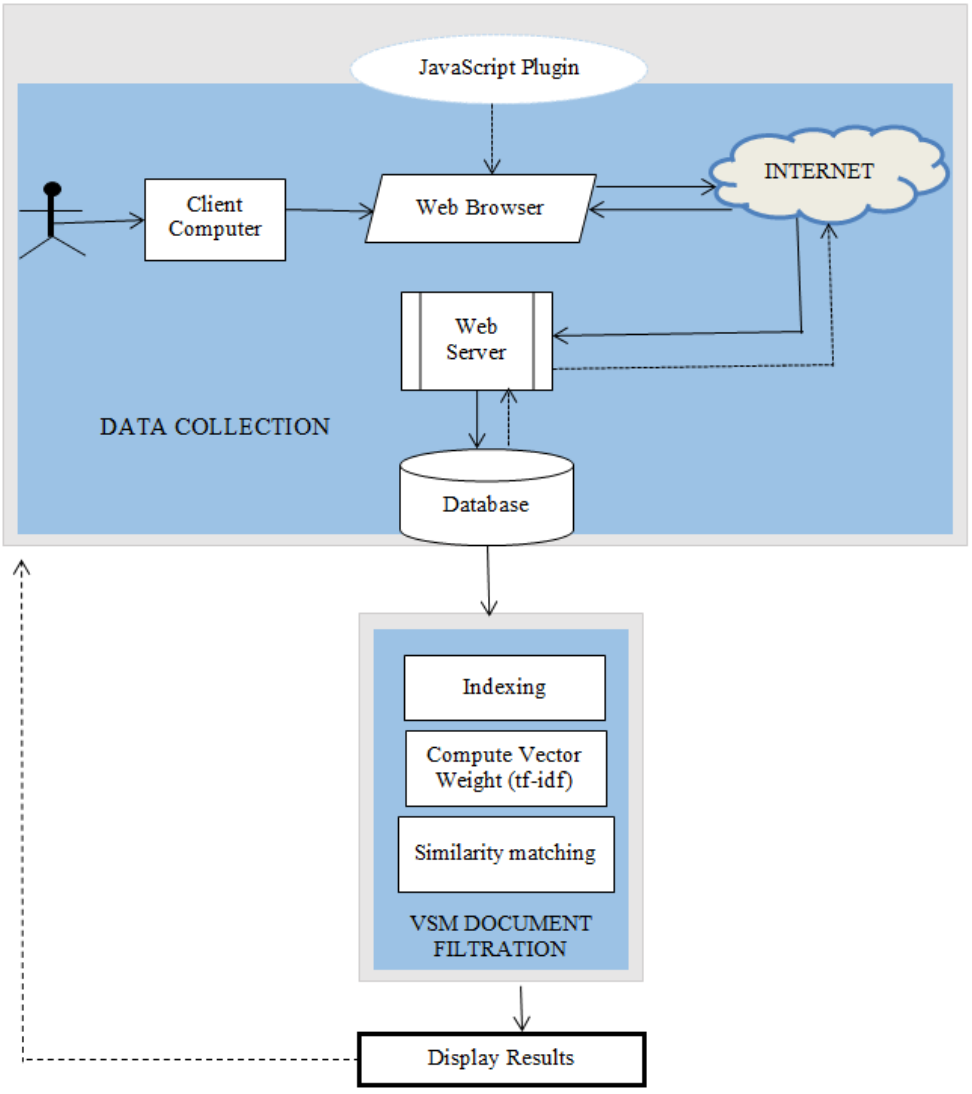

Fig.3. Conceptual Diagram Showing Solr-Indexed System Flow

Both the controlled and experimental systems had the same pool of documents obtained from an experiment [5]. Common documents in the dataset were merged and their mean interest weight was computed and presented as a single document. This reduced the size of the dataset from 343 to 140. Documents retrieved from

Google were crawled from different sources and indexed in their database. Fig. 4, 5, and 6 shows the screen for the search query "RUP vs waterfall model". Fig. 4 shows the documents returned by Google, Fig. 5 shows the original SERP returned by the Solr-indexed system, and Fig. 6 is the re-ranked result returned by the aggregated system. For example, the document, "Difference Between Waterfall Methodology and RUP" is ranked 3rd by Google as can be seen in Fig. 4, It is ranked 1st in the solr-indexed system as shown in Fig. 5 and it is ranked 2nd in the aggregated system as shown Fig. 6. The different ranked position of the document for the three systems underpins the variability of the three systems.

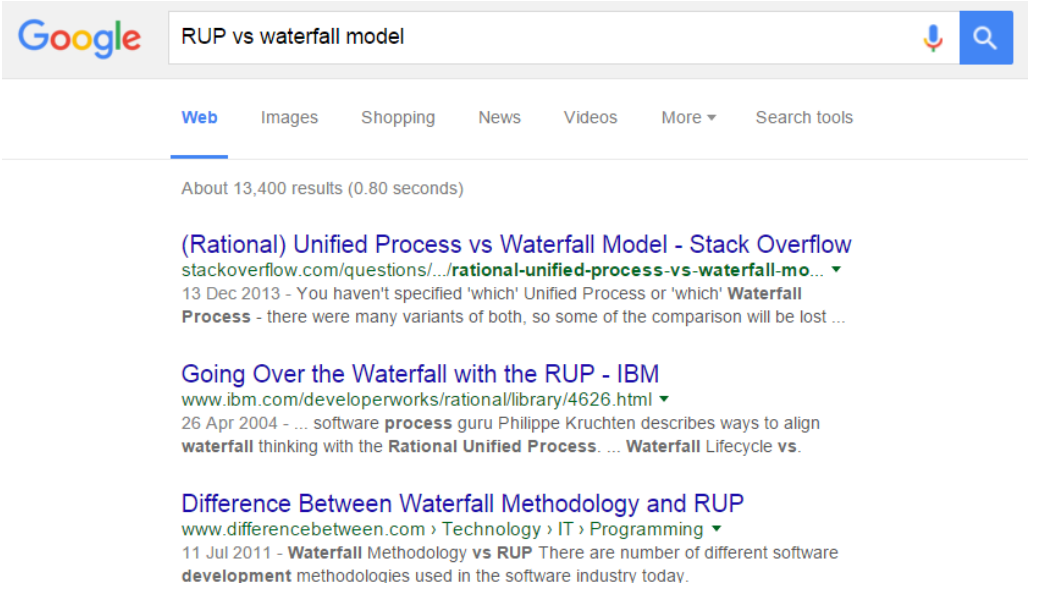

Fig.4. Sample Interface Showing Search Query and SERP for Google

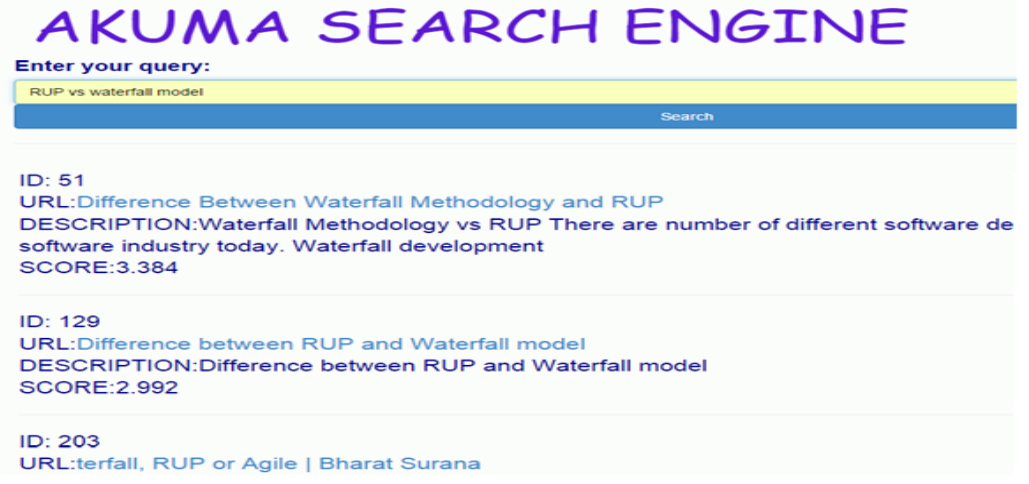

Fig.5. Sample Interface Showing Search Query and SERP for Solr-Indexed System (Note: The ID and SCORE were Hidden on the Student's Search Interface)

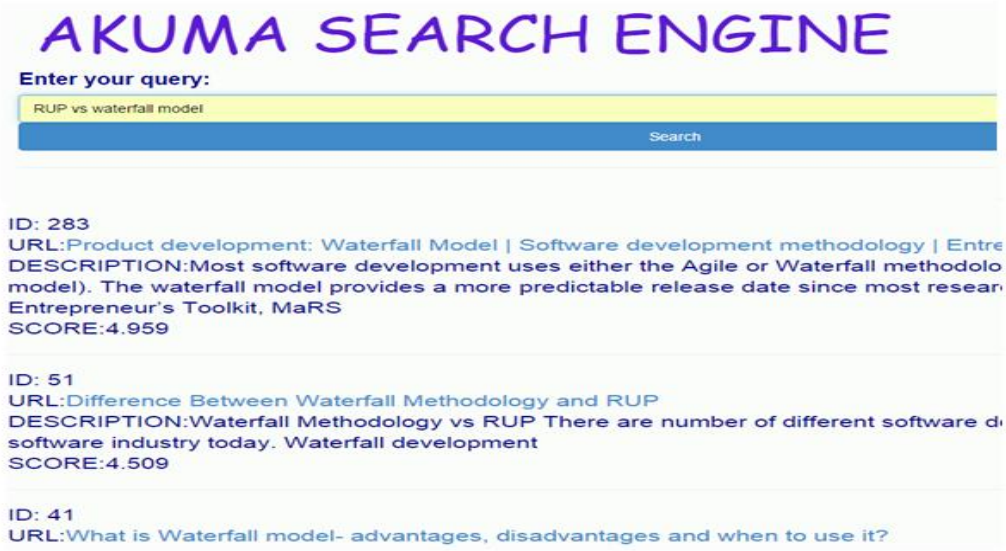

Fig.6. Sample Interface Showing Search Query and SERP for the Aggregated System (Note: The ID and SCORE were hidden on the Student's Search Interface)

The two approaches employed for the evaluation compares the performance of both the solr-indexed system and the aggregated system against the performance of the baseline Google system are explained in section 4.2 and 4.3. It should be noted that the interface presented to the users during the evaluation did not display the weight ("SCORE") and ID of each of the web document. It only showed the "URL" and the "document abstract/description". The "SCORE" is displayed on the interface presented in this paper to show how each of the web documents differs in weight.

4.2 Approach 1

In this approach, the A/B testing [44] was used. 15 users participated and they were randomly grouped into three sets, labelled A, B and C. Each set comprised 5 participants. The three groups of participants were given the same tasks and they visited different retrieval systems. Participants in group A performed the experiment with the baseline Google search engine; group B participants used the controlled system while the participants in group C performed the experiment with the experimental system.

4.3 Approach 2

In this approach, all the 11 participants were given the baseline, controlled and experimental system to use. They entered the same unique query of their choice in the three systems and rated the first top 10 results of each system according to how relevant they are to the given task. The participants were not told the difference between the systems in order to prevent bias in rating. Each of the 11 users rated up to 30 web pages.

4.4 Evaluation Metrics

The precision and recall relating to information retrieval problem were employed. The relationship between precision and recall is such that as precision increases, recall decreases and vice versa. The importance of each

depends on the context of usage. In this case, we desire that the top shown documents of a retrieval system should be more relevant than documents at the bottom of a retrieval list. Therefore, high precision is needed.

The precision of information retrieval problem measures the portion of relevant items within the total items retrieved. It involves retrieving the most relevant top-ranked documents and it is given as:

$$\textbf{Precision} = \frac{\textbf{number of relevant items retrieved}}{\textbf{total retrieved items}} \tag{7}$$

The recall for an information retrieval problem measures the portion of relevant items within the total relevant items retrieved. This involves the ability to find all relevant items in each collection. It is given as:

$$\textbf{Recall} = \frac{\textbf{number of relevant items retrieved}}{\textbf{total number of existing relevant items retrieved}} \tag{8}$$

Mean Average Precision (MAP) represents the area under the precision and recall curve. It is a single number that is used to compare the performance of retrieval algorithm. It is the average of precision values of a retrieval list at the positions where relevant documents were retrieved [45]. It is given as:

$$\textbf{MAP}(\boldsymbol{n}) = \frac{\mathbf{1}}{|\boldsymbol{n}|}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{AP}(\mathbf{i}) \tag{9}$$

where n is the number of queries used for searching and AP is the average precision. It is given as:

$$\boldsymbol{AP} = \frac{\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{P}(\boldsymbol{k})\times\boldsymbol{R}@\boldsymbol{k})}{\mathbf{Q}} \tag{10}$$

where n is the total retrieved documents, P(k) is the precision at k document level, R@k states whether the document at k is relevant or not. Q is the total relevant documents for a given query. When Q is zero, the document is zero [14], [39].

4.5 Statistical Significance Testing

Paired t-test calculated the significance between the average precision values of the baseline system against the controlled system, and the baseline system against the experimental system. Researchers [42], [46], [47] argued that paired t-test is the most reliable test for evaluating MAP values. A confidence interval of 95% and a statistical significant coefficient, $p < 0.05$, is used for analysing the dataset.

Null hypothesis indicates that there is no statistical significant relationship or association between two measured parameters. When the null hypothesis is rejected, it means there is a relationship between two parameters.

5. Results

This section presents the results of the two approaches discussed in the previous section. Section 5.1 presents the results for Approach 1 and section 5.2 presents the results for Approach 2. For the analysis of the results, some acronyms are defined in Table 1.

Table 1. Basic Acronyms used for Analysis

Acronym	Meaning
Top 5	This is the first 5 documents returned by a search system.
Top 10	This is the first 10 documents returned by a search system.
M_x	The mean of x, where x is either Google, Solr-indexed or Aggregated system.
SD_x	The Standard deviation of x, where x is either Google, Solr-indexed or Aggregated system.

5.1 Approach 1 Results

This section compares the results of the participants in the baseline system with the controlled system and experimental system in terms of the Mean Average Precision. Among the 15 participants, 5 users visited the baseline system, 5 others visited the controlled system and the remaining 5 students visited the experimental system. The precision was calculated for each of the 15 queries and the precisions at ranks where the documents were relevant for each query was summed and averaged. The mean average precision was then computed and the results from Google showed that at top 10, the MAP was 0.51 and at top 5, the MAP was 0.54. The MAP for the Solr-indexed system at top 10 was 0.77 and at top 5 was 0.84. The aggregated system produced an improved result. The MAP of the aggregated system for top 10 was 0.86 and for top 5 was 0.91. The paired T-test of the average precisions between the baseline system and solr-indexed system for the top 10 documents was statistically significant, it shows the mean of the solr-indexed system to be 0.26 higher than the baseline system ($p = 0.015$, $M_{google} = 0.51$, $SD_{google} = 0.096$, $M_{solr\text{-}indexed} = 0.77$, $SD_{solr\text{-}indexed} = 0.21$), there was a higher mean difference of 0.35 when the paired T-test was run between the baseline and the aggregated system ($p = 0.007$, $M_{google} = 0.51$, $SD_{google} = 0.096$, $M_{aggregated\_system} = 0.86$, $SD_{aggregated\_system} = 0.14$).

The result of the top 5 also showed a significant improvement of the solr-indexed system and the aggregated system over the baseline Google system. Google vs solr-indexed system produced ($p = 0.019$, $M_{google} = 0.54$, $SD_{google} = 0.11$, $M_{solr\text{-}indexed} = 0.84$, $SD_{solr\text{-}indexed} = 0.2$) and Google vs aggregated system produced ($p = 0.006$, $M_{google} = 0.54$, $SD_{google} = 0.11$, $M_{aggregated\_system} = 0.91$, $SD_{aggregated\_system} = 0.12$). Fig. 7 shows the mean average precision of the three systems in the two measured ranks.

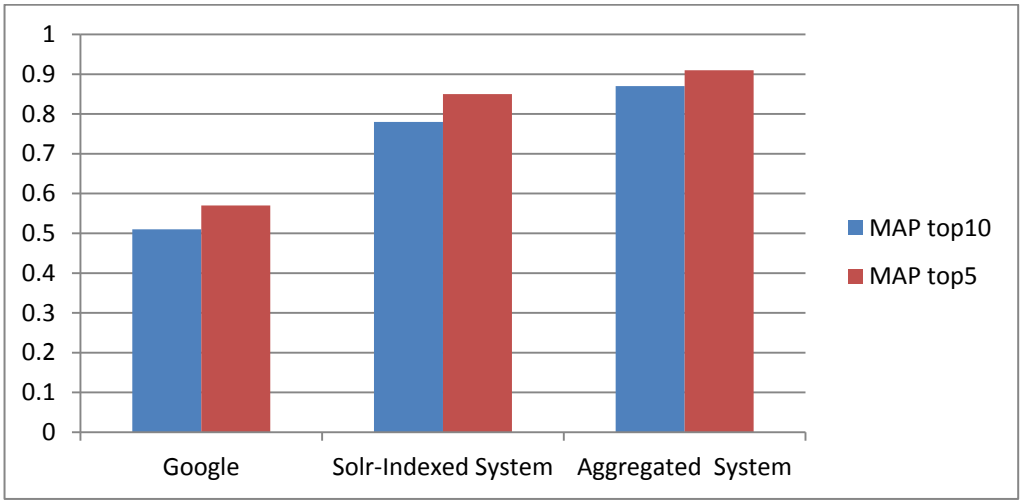

Fig.7. Approach 2 MAP Histograms Comparing Google, Solr-Indexed and the Aggregated System

5.2 Approach 2 Results

In this approach, 11 participants accessed the three systems and each user entered the same query in the three systems and rated the 10 top results according to relevance. Their ratings for the baseline, controlled and the experimental system was captured. The mean average precision for three systems was computed and the result shows that Google at top 10 was 0.51 and at top 5 was 0.57 while the MAP of the Solr-indexed system at top 10 was 0.78 and at top 5 was 0.85. The aggregated system in this approach also produced an improved result in terms of the MAP. The MAP of the aggregated system at top 10 was 0.87 while that of top 5 was 0.91 as shown in Figure 8. The paired T-test of the MAP of the systems also showed that the aggregated system has a statistically significant improvement than the baseline and Solr-indexed system.

The paired T-test between Google and solr-indexed system for the top 10 documents was significant with the solr-indexed system performing better by 0.27 MAP (p = 0.004, M_{google} = 0.51, SD_{google} = 0.083, $M_{solr\text{-}indexed}$ = 0.78, $SD_{solr\text{-}indexed}$ = 0.22) and the top 5 produced a significant improvement in MAP of the Sol-indexed system by 0.28 (p = 0.031, M_{google} = 0.57, SD_{google} = 0.14, $M_{solr\text{-}indexed}$ = 0.85, $SD_{solr\text{-}indexed}$ = 0.29). The aggregated system has a significant higher MAP over Google. For the top 10 documents, it is higher by 0.36 (p = 0.000, M_{google} = 0.51, SD_{google} = 0.083, $M_{aggregated\_system}$ = 0.87, $SD_{aggregated\_system}$ = 0.18) and for the top 5 documents, it is higher by 0.34 (p = 0.002, M_{google} = 0.57, SD_{google} = 0.14, $M_{aggregated\_system}$ = 0.91, $SD_{aggregated\_system}$ = 0.19).

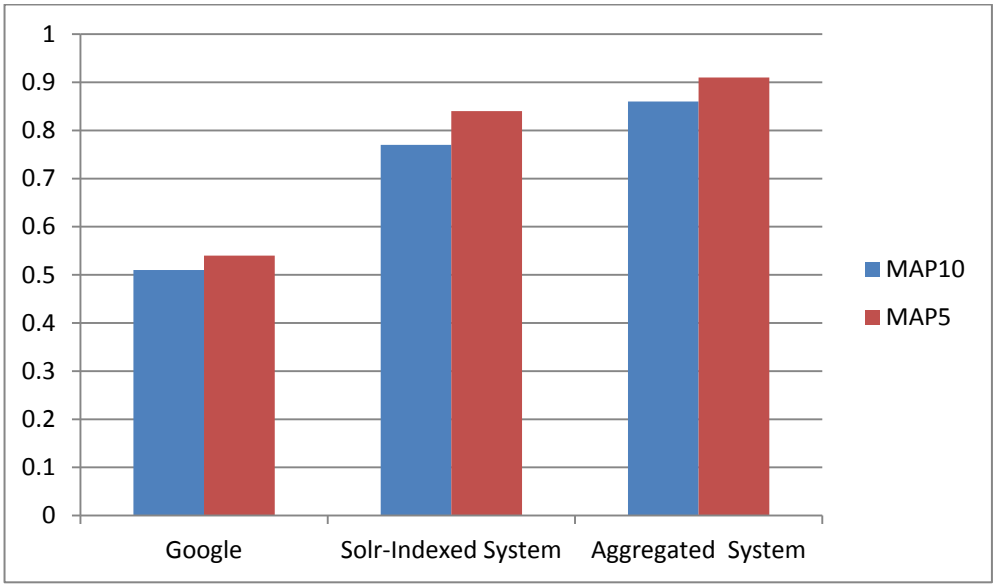

Fig.8. Approach 1 MAP Histograms Comparing Google, Solr-Indexed and the Aggregated System

As shown in the two approaches, the MAP values decrease as the number of documents increases from top 5 to top 10. This shows the trade-off between precision and recall. The aggregated system still performed better than the Solr-indexed system and the baseline system at the two computed document levels. It indicates an improvement of document relevance when queries are supplemented with user post-click behaviour. This result is similar to the results reported by previous studies [12], [14], [31], [41], [48]-[50].

Agichtein, Brill and Dumais (2006) used only a single indicator (click-through) to re-rank documents, Guo and Agichtein (2012) used scrolling and cursor movements to estimate relevance and Balakrishnan and Zhang (2014) used a heuristic to aggregate implicit indicators. In this work, a model obtained from the experimental analysis was used to derive a predictive function that estimates document relevance and the model was integrated with the traditional vector space model to improve the relevancy of retrieved documents.

6. Discussion

The result of the evaluation indicates that when implicit feedback is added to a retrieval system designed for users of a common domain, it improves document recommendation. In the experiment, we compared three systems; the generic Google system, the self-designed retrieval system using Solr technology without implicit feedback and the proposed system with implicit feedback. We argue that since Google is generic, a domain-specific retrieval system designed for a community of users through capturing and sharing relevant documents visited by them is needed to optimise document recommendation. Simulated task situation limited users to a particular domain for the study; this, however, limited the researchers from comparing the approach used in this work with similar approaches because the prototype differs from other systems in the way data is collected. Also, the researchers did not have access to the database of related systems and so could not populate them with documents related to the simulated task. We, therefore decided to follow a similar evaluation approach used by Balakrishnan and Zhang [14] and Alhindi et al [27], where a self-designed system was used for the evaluation. In as much as the similar type of systems were not used for the evaluation, the self-designed systems were carefully designed without bias to achieve the objective of the research.

This work proves that we can gather common relevant documents from users of the same domain (like students studying the same module) and use the information to optimise recommendation by re-ranking the Search Engine Result Page (SERP) based on integrating the traditional tf-idf and the implicit regression model. We show that even though domain-specific systems perform better than generic systems, there is much improvement in terms of document recommendation when the system is designed with implicit feedback than when they are not. This work gives a clearer pointer in understanding implicit feedback systems in the context of document recommendation.

With respect to the usability analysis of the proposed system, we will apply a user-centred design approaches to evaluate the system from usability perspectives [51], [52]. This work is limited in the sample size of the participants. The data used for the evaluation was relatively small when compared to TREC and Cranfield. The research attempted to show that documents used by previous learners of a common domain can be utilised to improve recommendation. For this reason, the proposed system is compared with a generic system (Google) and self-designed domain-specific system (without implicit feedback). Also, comparative evaluation with related work was not carried out because domain-specific tasks situations used for this work are not the same with those of the previous studies.

Future work should include a longitudinal study to collect a large amount of data in order to build user profiles for hybrid (content and collaborative) recommendation system. The evaluation will also be designed such that similar systems/approaches will be compared with the proposed system.

7. Conclusion

This paper has discussed the proposed domain-specific implicit feedback system to improve the retrieval of the relevant document. The paper describes the Vector Space Model which the system uses to match queries to related documents. Apache Solr technology was used to implement the Vector Space Model functionality of indexing, term weighing, similarity matching and scoring. The system integrates interest weight (*CIW*) obtained from a predictive model based on implicit indicators and classical TF-IDF algorithm based on Vector Space Model. An enhanced algorithm is developed to demonstrate the working principle of the system (for assigning scores to documents and re-ranking retrieval documents). Two approaches were used to evaluate the system. The results in both approaches show that the aggregated system performed better than the baseline and Solr-indexed system in terms of the mean average precision. This indicates that when users' queries are supplemented with their post-click behaviour, it improves the relevancy ranking of retrieval results. This further validates that personalisation is key in solving the information overload problem.

References

[1] D. Gurung, U. K. Chakraborty and P. Sharma, "An analysis of the Intelligent Predictive String Search Algorithm: A Probabilistic Approach," *International Journal of Information Technology and Computer Science(IJITCS)*, vol. 9, 2, pp. 66-75, 2017.

[2] M. Claypool et al, "Implicit interest indicators," in International Conference on Intelligent User Interfaces, Proceedings IUI, 2001, pp. 33-40.

[3] V. Balakrishnan, K. Ahmadi and S. D. Ravana, "Improving retrieval relevance using users' explicit feedback," *Aslib J. Inf. Manage.*, vol. 68, *(1)*, pp. 76-98, 2016.

[4] K. Takano and K. F. Li, "An adaptive personalized recommender based on web-browsing behavior learning," in *Proceedings - International Conference on Advanced Information Networking and Applications, AINA*, 2009, pp. 654-660.

[5] S. Akuma *et al*, "Comparative analysis of relevance feedback methods based on two user studies," *Comput. Hum. Behav.*, vol. 60, pp. 138-146, 7, 2016.

[6] B. Zhang et al, "Survey of user behaviors as implicit feedback," in 2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering, CMCE 2010, 2010, pp. 345-348.

[7] G. Buscher *et al*, "Attentive Documents: Eye Tracking as Implicit Feedback for Information Retrieval and Beyond," *ACM Transactions on Interactive Intelligent Systems*, vol. 2, *(1)*, pp. 1-30, 2012.

[8] A. F. M. Nazmul *et al*, "Identifying emotion by keystroke dynamics and text pattern analysis," *Behaviour & Information Technology*, vol. 33 (9), 2014.

[9] S. Akuma, "Investigating the Effect of Implicit Browsing Behaviour on Students' Performance in a Task Specific Context," *International Journal of Information Technology and Computer Science(IJITCS)*, vol. 6, *(5)*, pp. 11-17, 2014.

[10] G. Jawaheer, P. Weller and P. Kostkova, "Modeling User Preferences in Recommender Systems: A Classification Framework for Explicit and Implicit User Feedback," *ACM Transactions on Interactive Intelligent Systems*, vol. 4, pp. 1-26, 2014.

[11] S. Akuma *et al*, "Implicit predictive indicators: Mouse activity and dwell time," in *10th IFIP WG 12.5 International Conference, AIAI 2014*, Rhodes, Greece, 2014, pp. 162-171.

[12] S. Fox *et al*, "Evaluating implicit measures to improve Web search," *ACM Transactions on Information Systems*, vol. 23, *(2)*, pp. 147-168, 2005.

[13] S. Akuma *et al*, "Inferring users' interest on web documents through their implicit behaviour," *Commun. Comput. Info. Sci.*, vol. 517, pp. 315-324, 2015.

[14] V. Balakrishnan and X. Zhang, "Implicit user behaviours to improve post-retrieval document relevancy," *Comput. Hum. Behav.*, vol. 33, pp. 104-112, 2014.

[15] A. Grzywaczewski and R. Iqbal, "Task-Specific Information Retrieval Systems for Software Engineers," *Journal of Computer and System Sciences, Elsevier*, vol. 78, *(4)*, pp. 1204-1218, 2012.

[16] Z. Zhu et al, "User interest modeling based on access behavior and its application in personalized information retrieval," in Proceedings - 3rd International Conference on Information Management, Innovation Management and Industrial Engineering, ICIII 2010, 2010, pp. 266-270.

[17] M. Busby, *Learn Google.* Plano, Texas: Wordware Publishing Inc, 2003.

[18] R. Iqbal *et al*, "Design implications for task-specific search utilities for retrieval and reengineering of code," *Enterprise Information Systems*, pp. 1751-7575, 2015.

[19] R. W. White and D. Kelly, "A study on the effects of personalization and task information on implicit feedback performance," in *International Conference on Information and Knowledge Management, Proceedings*, 2006, pp. 297-306.

[20] L. A. Leiva and J. Huang, "Building a better mousetrap: Compressing mouse cursor activity for web analytics," *Information Processing & Management*, vol. 51, *(2)*, pp. 114-129, 3, 2015.

[21] H. Lieberman, "Autonomous interface agents," in *Conference on Human Factors in Computing Systems - Proceedings*, 1997, pp. 67-74.
[22] E. Han *et al*, "WebACE: A web agent for document categorization and exploration," in *Proceedings of the International Conference on Autonomous Agents*, 1998, pp. 408-415.
[23] L. Chen and K. Sycara, "WebMate - a personal agent for searching and browsing," in *Proceedings of the 2nd International Conference on Autonomous Agents*, 1998, .
[24] A. Chandrakala and k. Sanjay Dwivedi, "Keyphrase Extraction of News Web Pages ," *International Journal of Education and Management Engineering(IJEME)*, vol. 8, 1, pp. 48-58, 2018.
[25] T. Joachims, D. Freitag and T. Mitchell, "WebWatcher: A tour guide for the world wide web," in *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, 1997, pp. 770-775.
[26] M. Balabanovic, Y. Shoham and Y. Yun, "An Adaptive Agent for Automated Web Browsing," *Journal of Image Representation and Visual Communication*, vol. 6(5), 1995.
[27] A. Alhindi *et al*, "Profile-Based Summarisation for Web Site Navigation," *ACM Transactions on Information Systems*, vol. 33, *(1)*, pp. 1-40, 2015.
[28] E. J. Glover *et al*, "Web Search - Your Way: Improving Web searching with user preferences," *Commun ACM*, vol. 44, *(12)*, pp. 97-102, 2001.
[29] J. M. Ram ́ırez, J. Donadeu and F. J. Neves, "Poirot: A relevance-based web search agent," 2000.
[30] A. Kumar and M. Ashraf, "Efficient technique for personalized web search using users browsing history," in *International Conference on Computing, Communication and Automation, ICCCA 2015*, 2015, pp. 919-923.
[31] Q. Guo and E. Agichtein, "Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior," in *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web*, 2012, pp. 569-578.
[32] G. Buscher *et al*, "Large-scale analysis of individual and task differences in search result page examination strategies," in *WSDM 2012 - Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, 2012, pp. 373-382.
[33] E. R. Núñez-Valdéz *et al*, "Implicit feedback techniques on recommender systems applied to electronic books," *Comput. Hum. Behav.*, vol. 28, *(4)*, pp. 1186-1193, 2012.
[34] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, vol. 24, *(5)*, pp. 513-523, 1988.
[35] B. Bina, R. H. Goudar and K. Kaushal, "Quine-McCluskey: A Novel Concept for Mining the Frequency Patterns from Web Data," *International Journal of Education and Management Engineering(IJEME)*, vol. Vol.8, No.1, pp. 40-47, 2018.
[36] A. Grzywaczewski *et al*, "An Investigation of User Behaviour Consistency for Context-Aware Information Retrieval Systems ," *International Journal of Advanced Pervasive and Ubiquitous Computing (IJAPUC)*, vol. 1(4), pp. 69-90, 2009.
[37] G. Salton, A. Wong and C. S. Yang, "VECTOR SPACE MODEL FOR AUTOMATIC INDEXING." *Commun ACM*, vol. 18, *(11)*, pp. 613-620, 1975.
[38] G. Salton and C. Buckley, "Improving Retrieval Performance by Relevance Feedback," *Journal of the American Society for Information Science.*, vol. 44(4), pp. 288-297, 1990.
[39] D. Kelly, "Methods for evaluating interactive information retrieval systems with users," *Foundations and Trends in Information Retrieval*, vol. 3, *(1-2)*, pp. 1-224, 2009.
[40] R. W. White and G. Buscher, "Text selections as implicit relevance feedback," in SIGIR'12 - Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2012, pp. 1151-1152.
[41] E. Agichtein, E. Brill and S. Dumais, "Improving web search ranking by incorporating user behavior information," in Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006, pp. 19-26.
[42] M. D. Smucker, J. Allan and B. Carterette, "A comparison of statistical significance tests for

information retrieval evaluation," in *In Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM '07,* New York, NY, USA, 2007, pp. 623-632.

[43] B. P. Knijnenburg *et al*, "Explaining the user experience of recommender systems," *User Modeling and User-Adapted Interaction*, vol. 22, pp. 441-504, 2012.

[44] D. Manning C., P. Raghavan and H. Schütze, *Introduction to Information Retrieval.* York, NY, USA: Cambridge University Press, 2008.

[45] *Evaluation 12: mean average precision*. Available: https://www.youtube.com/watch?v=pM6DJ0ZZee0.

[46] M. Sanderson and J. Zobel, "Information retrieval system evaluation: Effort, sensitivity, and reliability," in In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05, New York, NY, USA, 2005, pp. 162-169.

[47] V. Cormack G and T. Lynam R., "Validity and power of t-test for comparing map and gmap ," in In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, New York, NY, USA, 2007, pp. 753-754.

[48] S. Jung, J. L. Herlocker and J. Webster, "Click data as implicit relevance feedback in web search," *Inf. Process. Manage.*, vol. 43, *(3),* pp. 791-807, 2007.

[49] T. Joachims *et al*, "Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search," *ACM Trans. Inf. Syst.*, vol. 25, *(2),* 2007.

[50] J. Huang, R. W. White and S. Dumais, "No clicks, no problem: Using cursor movements to understand and improve search," in *Conference on Human Factors in Computing Systems - Proceedings*, 2011, pp. 1225-1234.

[51] R. Iqbal *et al*, "ARREST: From Work Practices to Redesign for Usability," *The International Journal of Expert Systems with Applications, Elsevier,* vol. 38(2), pp. 1182-1192, 2011.

[52] R. Iqbal et al, "User-centred design and evaluation of ubiquitous services," in Proceedings of the 23rd Annual International Conference on Design of Communication: Documenting and Designing for Pervasive Information, ACM SIGDOC, 2005, pp. 138-145.

Authors' Profiles

Dr. Stephen Akuma is currently a Lecturer of Computer Science at Benue State University. He holds a PhD in Computing and a Masters degree in Software Development (Distinction) from Coventry University, United Kingdom. Stephen also holds a Bachelor's degree in Computer Science (2.1) from Benue State University. His research area is informational retrieval and personalisation. He has 9 years of Teaching Experience and has published papers in International Conferences and Journals.

Dr. Iqbal serves on programme committee and an advisory committee of many international conferences and journals. He has published more 100 papers in international journals, conferences, book chapters and workshops. He has supervised to completion 12 PhD students in information retrieval, health shock prediction, disaster management, emotions modelling, industrial automation and fault detection. He is currently a Reader in Human-Centred Technology in the School of Computing, Electronics and Mathematics at Coventry University. His main research interests lie in Big Data Analytics and Information Retrieval.

APPENDIX A

Simulated Work Task Situation

GIG Software Development company employed you as a consultant to provide a solution to the Company's pressing problem of developing a customised software within a minimal time frame. Some professional software developers achieved this by using the Rational Unified Process while others used the waterfall model.

Indicative Request

Which of the approaches would you consider for a small project of few lines of code (LOC) and what stage of the software lifecycle do you consider to be the most important? State the reason for your answer in your report.

How to cite this paper: Stephen Akuma, Rahat Iqbal,"Development of Relevance Feedback System using Regression Predictive Model and TF-IDF Algorithm", International Journal of Education and Management Engineering(IJEME), Vol.8, No.4, pp.31-49, 2018.DOI: 10.5815/ijeme.2018.04.04