ORIGINAL RESEARCH

# Re-ranking Google search returned web documents using document classification scores

Suthira Plansangket*, John Q Gan

*School of Computer Science and Electronic Engineering, University of Essex, United Kingdom*

**Received:** August 14, 2016 **Accepted:** November 2, 2016 **Online Published:** November 13, 2016
**DOI:** 10.5430/air.v6n1p59 **URL:** http://dx.doi.org/10.5430/air.v6n1p59

## ABSTRACT

Web document ranking is a very challenging issue for search engines because about 80% of the search engine users are usually interested in the top three returned search results only. This paper proposes an effective method for re-ranking Google search returned web documents/pages based on document classification. This method downgrades some web documents/pages that have lower classification scores or been classified into categories irrelevant to the query. The experimental results show that the re-ranking of Google search returned web documents using document classification scores can significantly improve the ranking performance in terms of the integrated evaluation result using three criteria: MAP, nDCG, and P@20. It is evident that the proposed re-ranking method can meet the user's information need better.

**Key Words:** Web document ranking, Document classification, Text mining

## 1. INTRODUCTION

Internet search engines play the most important role in finding information from the web. One of the great challenges faced by search engines is to understand precisely users' information need, since users usually submit very short and imprecise queries.[1] Almost all the query processing and returned results ranking in search engines are done by indexing and search algorithms without fully accessing the source of documents.[2] It often occurs in search engines that top-ranked returned web documents may not contain information relevant to users' search intent, and on the other hand relevant fresh web pages may not get high ranks.[3] He and Ounis[4] have found an indicator for measuring the relevance between the user query and the web returned results. Their experimental results illustrate that only the top five documents returned from search engines are usually very highly relevant whilst the relevance of the remaining returned documents drops dramatically. From the previous research,[5] almost 80% of search engine users are interested only in the top three returned web pages. Pan[6] also reported that only the top ranked web pages get high clickthrough rates. Therefore, improving the quality of web document ranking is a very important issue although it is a challenging task. One of the solutions to these problems is to automatically organise documents into user's interesting topic groups. Two issues that will be addressed in this paper are classification and ranking of search engine returned web documents.

Document classification techniques have been applied to many areas such as spam filtering,[7] email routing,[8] and genre classification.[9] Widely used classifiers include k-nearest neighbours (kNN),[10,11] support vector machine (SVM),[10] and linear discriminant analysis (LDA).[12,13] In this paper, LDA classifier is used to classify search engine returned documents into relevant topic categories and re-rank

*Correspondence: Suthira Plansangket; Email: psuthira@gmail.com; Address: School of Computer Science and Electronic Engineering, University of Essex, United Kingdom.

the documents using classification scores. For document representation, the class specific document frequency (CSDF) weighting method is adopted, which has been demonstrated to effectively improve the performance of document classification in comparison with other widely used vector space model (VSM) based document representations.[14]

This paper proposes a new ranking method called GCrank that combines the original Google ranking scores and the LDA classification scores of the Google search returned web documents to improve ranking performance, which is demonstrated by experimental results in terms of several widely used ranking performance criteria.

## 2. WEBPAGE RANKING

### 2.1 Content-based ranking

Content-based ranking technologies were developed for retrieving web pages for specific queries and similarity page queries. Their algorithms usually work by matching queries with keywords or features in web documents and user's web logs. Traditional document content representation methods include VSM based on term presence or term frequency and inverse document frequency (TF-IDF). There are new document representation methods and similarity measures proposed in recent years including learning to rank and personalisation-based ranking. For example, Du and Hai[15] proposed a method for measuring webpage similarity based on formal concepts analysis (FCA).

Learning to rank has emerged in the past decade. It is the application of machine learning, typically supervised, semi-supervised, or reinforcement learning, in the construction of ranking models for information retrieval systems. Xiang *et al.*[16] developed different ranking principles for different types of contexts, which were integrated into a state-of-the-art ranking model by encoding the context information as features of the model using a learning-to-rank approach. Context information includes previous queries and the search results clicked on or skipped by users. Derhami *et al.*[17] represented two new ranking algorithms using reinforcement learning concepts and a new hybrid approach using a combination of BM25 (best matching 25)[13] and their machine learning method.

Personalisation-based ranking has been investigated in recent years. Lu *et al.*[18] proposed a user model based ranking method, in which the user model is mainly used to capture and record the user's interests. Wang *et al.*[19] proposed a general ranking model adaptation framework for personalised search using a user-independent ranking model and the number of adaptation queries from individual users.

Semantic web search also includes content-based ranking. Research on semantic search ranking[20–22] aims to improve traditional information search and retrieval methods by using ontologies. However, there are some problems such as heterogeneity or overlapping domains.

### 2.2 Hyperlink-based ranking or connectivity-based ranking

Hyperlink-based ranking[5] is the early ranking method that focuses on the number of hyperlinks that point to a webpage or the incoming links. Links carry information that can be used to evaluate the importance of webpages and the relevance of webpages to the user's query to some extent. The examples of the well-known hyperlink-based ranking methods are HITS and PageRank.

HITS[5, 15, 23] stands for hypertext induced topic search. The webpages' hyperlink structures in the web graph induced are defined to authorities and hubs. A good hub represents a webpage that points to many other webpages, and a good authority represents a webpage that is linked by many different hubs. It is a query-dependent method; however, the repeated web results and topic diffusion are the drawbacks of this method. In addition, in real applications, the HITS algorithm also produces some problems, such as the time and space costs of constructing the subgraph of the search topic are high, and it is not suitable for specific queries.

PageRank[5] is the best known method because of Google. Suppose that a webpage $a$ is pointed by webpages $p_1$ to $p_n$ on the web graph, and a user jumps to webpage $a$ with probability $q$ or follows one of the hyperlinks of webpage $a$ with probability *1-q*. The PageRank of webpage $a$ is given by the probability $PR(a)$ of finding the user in webpage $a$, which is defined as follows:

$$PR(a) = \frac{q}{T} + (1-q)\sum_{i=1}^{n}\frac{PR(p_i)}{L(p_i)} \quad (1)$$

where $T$ is the total number of webpages on the web graph, $PR(p_i)$ is the PageRank of webpage $p_i$, $L(p_i)$ is the number of outgoing links of webpage $p_i$, and $n = L(a)$. This method may have a problem because the real web graph contains dead ends where webpages have no link or self-link. The solution to this problem is the jumping criteria of the Markov chain. In addition, Alkhalifa[24] reported that the adjacency matrix used as a basis for PageRank may have biased spaces that need to be taken into consideration.

### 2.3 Hyperlink-content-based ranking

Hyperlink-content-based ranking[15] is about finding an appropriate balance between the relevance and the popularity of webpages. Generally, search engines use a combination

of hyperlink-based and content-based algorithms. The priority value or ranking score of a webpage is computed by a combination of a score related to its hyperlinks and another score related to its content. For example, the combination of BM25 and PageRank can be the baseline for evaluating new ranking methods.

Google search engine[2] adopts a complex hyperlink-content-based ranking approach. It makes use of the link structure of the web to calculate a quality ranking for each webpage (PageRank), which forms a probability distribution over webpages. It also makes use of other content-based features of webpages. The ranking systems are separated into two categories. Firstly, for a single word query, Google considers a hit list of the query word in each webpage, such as title, anchor, URL, and large font for that word. Each of these has its score. These scores are combined with PageRank score to give a final rank. Secondly, for a multi-word query, multiple hit lists are generated. A proximity score is computed based on how far apart the hits are in a webpage, which is then combined with the scores for individual single word query. Google employs a number of techniques to improve search quality, including PageRank, anchor text, and proximity information, etc.

Although Google ranking is well recognised as the best webpage ranking method, there is still room for improvement. This paper investigates whether re-ranking Google search returned web documents by using document classification scores is able to improve ranking performance in terms of widely used performance evaluation criteria. Therefore, Google ranking is used as a baseline ranking method for comparison to evaluate the proposed method in this paper.

## 3. THE PROPOSED RANKING METHOD: GCRANK

### 3.1 CSDF for web document representation

Class specific document frequency (CSDF) was proposed recently as an effective feature for document representation.[14] In this paper, the effectiveness of CSDF will be further investigated for web document representation for classification purposes. The basic idea of CSDF is that a term in a document is very important for classifying documents if it is more frequent inside the document and other documents belonging to the same class as well but less frequent in documents belonging to different classes. The CSDF value of term $i$ for class $k$ is calculated as follows:

$$CSDF_{ik} = \begin{cases} \frac{DF_{ik}/N_k}{(DF_i - DF_{ik})/(N - N_k)+1} , & \text{if term } i \text{ is present} \\ 0 , & \text{otherwise} \end{cases} \quad (2)$$

where $DF_{ik}$ is the document frequency of term $i$ based on the documents belonging to class $k$ in the training document set, $DF_i$ is the document frequency of term $i$ based on all the documents in the training document set, $N_k$ is the number of documents belonging to class $k$ in the training document set, and $N$ is the number of documents in the training document set. However, the values of $DF_{ik}$ and $N_k$ are not supposed to known in testing data. In our method, the CSDF value of term $i$ in both training and testing data is defined as the variance of the original CSDF values of term $i$ for class $k$, *i.e.*,

$$CSDF_i = var(CSDF_{ik}) \quad (3)$$

For a web document, a bag of words are extracted first and then transformed into a vector of CSDF values as the representation of the document, one value for each word, which forms the input to classifiers in the next processing stage.

### 3.2 Using LDA scores for re-ranking Google search returned web documents

Linear discriminant analysis (LDA) is a linear classification method originally developed by Fisher.[25,26] It has two distinct functions: dimensionality reduction and data classification. LDA is adopted for classification in this paper due to its simplicity and resilience to overfitting. The LDA classification score $Cscore$ is defined in this paper as follows:

$$\sum = (n_1 \sum\nolimits_1 + n_2 \sum\nolimits_2)/n \quad (4)$$

$$\mathbf{w} = \sum^{-1} (\mu_{\mathbf{1}} - \mu_{\mathbf{2}}) \quad (5)$$

$$w_0 = \mathbf{w}^T (n_1 \mu_{\mathbf{1}} + n_2 \mu_{\mathbf{2}})/n \quad (6)$$

$$Cscore = \mathbf{w}^T \mathbf{x} - w_0 \quad (7)$$

where $\sum_1$ and $\sum_2$ are the covariance matrices of the samples of class 1 and class 2 respectively, $n_1$, $n_2$ are the number of samples in class 1 and class 2 respectively, $n$ is the total number of all the samples, $\mu_{\mathbf{1}}$, $\mu_{\mathbf{2}}$ are the means of the samples of class 1 and class 2 respectively.

The motivation of the proposed method arises from an idea that top-ranked webpages should belong to the same topic category as the one relevant to the query. That is, involving classifiers in the ranking process may improve webpage ranking performance. Classification scores of web documents usually indicate how much the webpages are relevant to the query. For the LDA classifier used in this paper, one can visualize its operation as splitting a high-dimensional feature space with a hyperplane defined by $Cscore = 0$. All points

representing web documents on one side of the hyperplane are classified into one class. If a point is far away from the hyperplane, the corresponding web document will have a high classification score and it is ensured that this web document is in that class with high confidence. Therefore, this method assumes that a web document with a high classification score should have a relatively high rank in the search engine returned results. On the other hand, if a returned web document is classified into a query irrelevant topic category, its rank should be considerably reduced. Google ranking has already been recognised as an outstanding ranking method. It would be highly desirable if Google ranking can be further improved by combining it with web document classification scores. For this purpose this paper proposes the GCrank method as described by equations (8), (9), and (10).

$$norGscore_d = \frac{1}{GoogleRank_d}, 0 \leq norGscore_d \leq 1 \quad (8)$$

$$norCscore_d = \begin{cases} \frac{Cscore_d}{MaxCscore}, & 0 \leq norCscore_d \leq 1 \\ 0, & \text{if } d \text{ is misclassified} \end{cases} \quad (9)$$

$$GCrank_d = \alpha \times norGscore_d + (1-\alpha) \times norCscore_d \quad (10)$$

where $GCrank_d$ is a combined ranking score of document $d$, $norGscore_d$ is the normalised Google ranking score of document $d$, $GoogleRank_d$ is the original Google's rank of document $d$, $norCscore_d$ is the normalised classification score of document $d$, $Cscore_d$ is the original classification score of document $d$, $MaxCscore$ is the maximum classification score of all the web documents returned by a query, and $\alpha$ is a weighting factor. In our experiment, $\alpha$ has been investigated and its value is determined by through cross-validation using a small fraction of the training data.

## 4. Experimental results

### 4.1 Experimental procedure

For collecting sample web documents for training LDA classifiers and evaluating their classification accuracy, the top 56 Google search returned web documents for each query were obtained and saved using the Google API. Each web document was pre-processed. Firstly, only the title and snippet content in each document were processed, with HTML tags discarded. Secondly, the contents were divided into tokens. Because nouns are the most discriminative terms,[5,27] only nouns were considered. Finally, stemming was conducted to convert inflected or derived words to their stem or root form. Furthermore, the CSDF method was used for representing the document as a vector of CSDF values.

**Table 1.** Ten categories of test queries

| Category | Description | Number of queries |
|---|---|---|
| 1 | Animals | 10 |
| 2 | Arts | 10 |
| 3 | Flowers | 10 |
| 4 | Food | 10 |
| 5 | Movie | 10 |
| 6 | Shopping | 10 |
| 7 | Sports | 10 |
| 8 | Travel | 10 |
| Total | | 80 |

In this experiment, there are 80 specifically designed queries that were chosen from eight popular search categories, as illustrated in Table 1. Each category has 10 queries, each composed of a couple of words that are well-known and suitable for user evaluation. It is important to know whether a search returned web document is truly relevant to the query or not in the performance evaluation. Highly relevant, mildly relevant and irrelevant returned web documents were decided by three users from University of Essex, which were used as true ranks of the returned web documents in the ranking performance evaluation. Three performance evaluation criteria: MAP, P@20, nDCG, and their integration were adopted in the experiment to decrease subjective bias and make the experimental results more reliable.

**Table 2.** Information about the collected web document dataset

| Class | No. of documents | No. of training doc | No. of testing doc |
|---|---|---|---|
| Animals | 558 | 334 | 224 |
| Arts | 560 | 335 | 225 |
| Flowers | 558 | 335 | 223 |
| Food | 555 | 335 | 220 |
| Movie | 560 | 335 | 225 |
| Shopping | 559 | 335 | 224 |
| Sports | 558 | 335 | 223 |
| Travel | 556 | 335 | 221 |
| Total | 4464 | 2679 | 1785 |

About 4,500 returned web documents were used in the experiment, which were randomly split into training data and testing data using 60% and 40% of all documents for the training set and the testing set, respectively. The training dataset was used to select features, train classifiers, and select hyper-parameter values such as $\alpha$ in equation (10) through cross-validation, and the testing dataset to evaluate the performance of the trained classifiers and the ranking performance of the GCrank method. In order to have sufficient and balanced documents for each class, only eight classes that have the reasonable number of documents were adopted in the experiment, with 2,679 documents for training data and 1,785 documents for testing data. The details are shown in Table 2.

The words kept after pre-processing were the sources of feature extraction, without using predefined term dictionary. In this dataset, more than 5,000 words as initial source of features were extracted from about 4,500 documents. Only 4,500 documents as samples cannot be well represented in a space with over 5,000 features. Therefore, feature selection is necessary in this case.[5,28] Both filter and wrapper approaches were adopted for feature selection in this paper. For filter approach, the terms that have the top 40 document frequency values in each category of documents were considered, resulting in 258 features from 320 terms with high document frequency values for the eight classes, with duplicate terms removed. For wrapper approach, the features were selected using the sequential forward floating search (SFFS) method with LDA classifier as the wrapper.[29]

### 4.2 Performance criteria

Three widely used performance criteria[30] were adopted to evaluate ranking performance in the experiment: mean average precision (MAP), normalised discounted cumulated gain (nDCG), and precision at 20 (P@20).

#### 4.2.1 *Mean average precision (MAP)*

MAP[23,31] is an average precision over various queries and rankings. MAP is based on the assumption that a lot of highly relevant documents with respect to a query may appear in the top list of returning documents. Let the position of the ith relevant document returned for query $j$ by a ranking method be $r_{ji}$. The precision of the $i$th document is defined by

$$P_{ji} = \frac{i}{r_{ji}} \tag{11}$$

The precision score will be set to 0, if it is an irrelevant document. An average precision of all documents and queries is an $MAP$ score defined as follows:

$$MAP = \frac{1}{q}\sum_{j=1}^{q}\frac{1}{Q_j}\sum_{i=1}^{Q_j}P_{ji} \tag{12}$$

where $Q_j$ is the number of relevant documents for query $j$ and $q$ is the number of queries.

#### 4.2.2 *Normalised discounted cumulated gain (nDCG)*

MAP measures whether returned documents are relevant or irrelevant, which is a binary relevance assessment, whilst discounted cumulated gain (DCG) distinguishes between highly and mildly relevant documents. DCG[23,32] is an evaluation method with different graded relevance assessments. The cumulative gain (CG) of $Q_j$ documents for query $j$ is calculated as follows:

$$CG_j = w_1 + w_2 + \cdots + w_{Qj} \tag{13}$$

where $w_i$ is the relevance weighting factor of the $i$th document returned for query $j$ by a ranking method. DCG is calculated as follows by using a discount factor 1/($\log_2 i$):

$$DCG_j = w_1 + \frac{w_2}{\log_2 2} + \frac{w_3}{\log_2 3} + \cdots + \frac{w_{Qj}}{\log_2 Qj} \tag{14}$$

The $nDCG$ of query $j$ is calculated by

$$nDCG_j = \frac{DCG_j}{IDCG} \tag{15}$$

where $IDCG$ is the ideal DCG or the maximum possible DCG. The average of $nDCG_j$ over $q$ queries is defined as follows:

$$nDCG = \frac{1}{q}\sum_{j=1}^{q}nDCG_j \tag{16}$$

#### 4.2.3 *Precision at 20 (P@20)*

Precision is the amount of relevant documents divided by the total amount of irrelevant and relevant documents. Precision@20[31,33] is the precision for the top 20 returned web documents, which is defined by

$$P@20 = \frac{(amount\ of\ relevant\ documents\ among\ top\ 20)}{20} \tag{17}$$

The above three criteria emphasise different aspects of ranking performance. Only the performance of returning relevant web documents is measured by P@20, whilst the performance of both returning and ranking relevant web documents are measured by MAP and nDCG. In our experiment, they were integrated by averaging for statistical significance test.

**Table 3.** Overall classification accuracy of the LDA classifier

| Accuracy | The top 40 (258 features) | Wrapper (85 features) |
|---|---|---|
| Training | 94.25% | 88.09% |
| Testing | 92.38% | 87.28% |

**Table 4.** Classification accuracy of the LDA classifier for each category

| Category | Accuracy (%) |
|---|---|
| Animals | 88.19 |
| Arts | 95.54 |
| Flowers | 89.43 |
| Food | 99.64 |
| Movie | 90.71 |
| Shopping | 90.70 |
| Sports | 95.52 |
| Travel | 98.38 |
| Average | 93.51 |

### 4.3 Experimental results

#### 4.3.1 *Evaluation of the performance of classification and classification scores*

Table 3 shows the overall classification performance of the LDA classifier in terms of both training and testing accuracy. For filter approach, 258 features from the top 40 document frequency scores achieved over 92% accuracy. For wrapper approach, 85 features achieved over 87% accuracy. At this stage, the classification scores of all documents were determined and saved for use for web document ranking later. Furthermore, the classification performance of each category is shown in Table 4. The results illustrate that "animals" and "flowers" categories had the lowest classification accuracy (lower than 90% accuracy) while "food" category had the highest accuracy.

#### 4.3.2 *Choosing the weighting factor value ($\alpha$)*

For finding a proper value of the weighting factor $\alpha$ in equation (10), the performance of ranking the web documents returned for the 10 queries in the movie category only by the GCrank method with different $\alpha$ values was evaluated by the three performance criteria. The true ranks of the returned web documents were obtained through the feedback given by the three participants from the University of Essex. In this experiment, only the top 20 returned web documents were considered in all evaluation methods. Highly relevant documents were the top 10 documents, whose scores were multiplied by two in the nDCG evaluation method, while mildly relevant documents were the documents in 11th to 20th rank, whose scores were kept unchanged.

**Table 5.** Experimental results of using different weighting factor values

| Alpha | nDCG | | | MAP | | | P@20 | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| | P1 | P2 | P3 | P1 | P2 | P3 | P1 | P2 | P3 | |
| 0.80 | 0.8932 | 0.8482 | 0.5899 | 1.0000 | 0.8961 | 0.4467 | 0.7900 | 0.7400 | 0.4900 | 0.7438 |
| 0.85 | 0.9107 | 0.8589 | 0.6131 | 0.7814 | 0.7170 | 0.3825 | 0.8250 | 0.7650 | 0.5050 | 0.7065 |
| 0.90 | 1.0000 | 0.9464 | 0.6547 | 0.9224 | 0.8293 | 0.4293 | 1.0000 | 0.9350 | 0.5800 | 0.8108 |
| 0.95 | 0.9722 | 0.9222 | 0.6472 | 0.7290 | 0.6544 | 0.3529 | 0.9250 | 0.8650 | 0.5600 | 0.7364 |
| Average | 0.9440 | 0.8939 | 0.6262 | 0.8582 | 0.7742 | 0.4029 | 0.8850 | 0.8263 | 0.5338 | 0.7494 |
| | 0.8214 | | | 0.6784 | | | 0.7483 | | | |

**Table 6.** Evaluation results of the original Google ranking

| No. | Category | nDCG | | | MAP | | | P@20 | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | P1 | P2 | P3 | P1 | P2 | P3 | P1 | P2 | P3 | |
| 1 | Animals | 0.9873 | 0.8924 | 0.7106 | 0.9668 | 0.8779 | 0.5981 | 0.9750 | 0.8950 | 0.6700 | 0.8415 |
| 2 | Arts | 0.9629 | 0.9424 | 0.8773 | 0.9583 | 0.9583 | 0.7708 | 0.9750 | 0.9750 | 0.8400 | 0.9178 |
| 3 | Flowers | 0.7456 | 0.6928 | 0.4559 | 0.6484 | 0.6048 | 0.3171 | 0.7550 | 0.5950 | 0.5350 | 0.5944 |
| 4 | Food | 0.9753 | 0.9763 | 0.8948 | 0.9625 | 0.9248 | 0.7919 | 0.9700 | 0.9650 | 0.8500 | 0.9234 |
| 5 | Movie | 0.9231 | 0.9424 | 0.6586 | 0.8617 | 0.9208 | 0.7417 | 0.9250 | 0.9700 | 0.5650 | 0.8343 |
| 6 | Shopping | 0.9814 | 0.9323 | 0.9203 | 0.9208 | 0.8699 | 0.8599 | 0.9650 | 0.9000 | 0.9100 | 0.9177 |
| 7 | Sports | 0.9684 | 0.9847 | 0.9399 | 0.9514 | 0.9749 | 0.8864 | 0.9800 | 0.9900 | 0.9300 | 0.9562 |
| 8 | Travel | 0.9858 | 0.9539 | 0.8559 | 0.9769 | 0.9249 | 0.8483 | 0.9900 | 0.9500 | 0.7800 | 0.9184 |
| Average | | 0.9412 | 0.9147 | 0.7892 | 0.9059 | 0.8820 | 0.7268 | 0.9419 | 0.9050 | 0.7600 | 0.8630 |
| | | 0.8817 | | | 0.8382 | | | 0.8690 | | | |

Table 5 illustrates three evaluation results of four different weighting factor values based on the true ranks given by the three participants (P1, P2, P3). A weighting factor equal to 0.9 gave the best ranking performance. Therefore, the following evaluations used $\alpha$ equal to 0.9 to evaluate the proposed method.

#### 4.3.3 *GCrank effectiveness evaluation*

In the experiment the GCrank method was used to re-rank the Google returned web documents, aiming to improve their relevance to the query. The performances of ranking the returned web documents from the 80 test queries in eight categories were evaluated by the same criteria as used in Section 4.3.2. The statistical test used in this experiment was the Wilcoxon rank sum test with the $p$ value $\leq .05$. The experimental results of the original Google method and the GCrank method are shown in Table 6 and Table 7, respectively.

**Table 7.** Evaluation results of the GCrank method

| No. | Category | nDCG | | | MAP | | | P@20 | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | P1 | P2 | P3 | P1 | P2 | P3 | P1 | P2 | P3 | |
| 1 | Animals | 0.9809 | 0.9048 | 0.7253 | 0.9437 | 0.9006 | 0.6220 | 0.9500 | 0.9050 | 0.6850 | 0.8464 |
| 2 | Arts | 1.0000 | 1.0000 | 0.8861 | 1.0000 | 1.0000 | 0.7902 | 1.0000 | 1.0000 | 0.8500 | 0.9474 |
| 3 | Flowers | 0.7865 | 0.7031 | 0.4714 | 0.7090 | 0.6176 | 0.3419 | 0.7900 | 0.6350 | 0.5700 | 0.6249 |
| 4 | Food | 0.9850 | 0.9840 | 0.9040 | 0.9917 | 0.9724 | 0.8128 | 0.9750 | 0.9850 | 0.8650 | 0.9416 |
| 5 | Movie | 0.9453 | 1.0000 | 0.6547 | 0.8987 | 1.0000 | 0.7248 | 0.9350 | 1.0000 | 0.5800 | 0.8598 |
| 6 | Shopping | 1.0000 | 0.9425 | 0.9164 | 0.9895 | 0.9175 | 0.8861 | 0.9750 | 0.9250 | 0.9300 | 0.9424 |
| 7 | Sports | 0.9964 | 1.0000 | 0.9530 | 0.9900 | 1.0000 | 0.9170 | 0.9900 | 1.0000 | 0.9500 | 0.9774 |
| 8 | Travel | 1.0000 | 0.9808 | 0.8679 | 1.0000 | 0.9568 | 0.8327 | 1.0000 | 0.9650 | 0.8150 | 0.9354 |
| Average | | 0.9618 | 0.9394 | 0.7973 | 0.9403 | 0.9206 | 0.7409 | 0.9519 | 0.9269 | 0.7806 | 0.8844 |
| | | 0.8995 | | | 0.8673 | | | 0.8865 | | | |

The average evaluation results of the GCrank method based on the true ranks of each participant were higher than those of the original Google ranking by about 2%. The best improvement was obtained in terms of the MAP evaluation criterion. The best improvement by the GCrank method was in "arts", "flowers", and "shopping" categories, while there was no obvious improvement in "animals" category. Therefore, it is clear that the ranking of the returned web documents using the GCrank method was better than that of the original Google ranking based on the true ranks from the three participants in terms of all evaluation criteria.

**Table 8.** Statistical significance test results: GCrank *vs.* Google ranking

| Number of categories | Statistical test results ($p$ value) |
|---|---|
| 8 categories | .0427 |
| 7 categories (no animals) | .0243 |
| 6 categories (no animals and flowers) | .0059 |

In order to see whether the ranking performance improvement by the GCrank method is statistically significant, the integration (average) of the MAP, nDCG, and P@20 evaluation results of the two ranking methods were compared using the Wilcoxon rank sum test. Table 8 illustrates the statistical test results, with the $p$ value for eight categories being .0427 which is less than .05. From the results in Section 4.3.1, "animals" and "flowers" categories had the lowest classification accuracy. If these two categories are ignored, the statistical difference between the two ranking methods should be bigger. Table 8 shows that the $p$ value for seven categories after "animals" category had been removed was .0243. For six categories, after "animals" and "flowers" categories had been removed, the $p$ value was .0059. This indicates that the GCrank method was significantly better than the original Google ranking and it was more significantly better if only the categories with higher classification accuracies were considered.

In general, this experiment has demonstrated that the proposed method is significantly helpful for ranking the returned web documents. Table 9 illustrates four examples of re-ranking using the GCrank method that downgrades the ranking of some documents that are not classified into the category relevant to the query and upgrades the ranking of some documents that are classified into the category relevant to the query with high scores. In the first example, the query was "avatar" which was a famous movie. A highlighted returned web document was classified into "sports" category by the LDA classifier. The GCrank method ranked this web document down to the bottom of the list from a Google rank of 49. As a matter of fact, this web document was not di-

rectly related to "avatar" movie based on the ranks of the three participants. The second example shows some results returned by the query "frozen", which was also a famous movie. The GCrank method upgraded a highlighted web document about the review of this movie from 55th to 41st position due to that it was classified to the movie category with a high score. In the third example, the query was "taj mahal" from "travel" category. A highlighted returned web document was not directly related to "taj mahal" because it was actually a restaurant name. It was classified by the LDA classifier into "food" category, and the GCrank method ranked this document down to 53rd from 44th position. In the final example, the query was "great wall", which was a famous attraction in China. The proposed method increased the rank of a highlighted web document about the history of this place from the bottom of the list up to 23rd position due to the high classification score. These four examples indicate that the GCrank method can produce interpretable results.

**Table 9.** Some examples of re-ranking using GCrank

| Query | Original ranking | Re-ranking using GCrank |
|---|---|---|
| Avatar | 46 5 38.767 46 Avatar Secrets: An Interactive Documentary for t<br>47 5 42.261 47 Avatar & Aliens are the same movie - The O<br>48 5 39.332 48 Avatar \| Film \| The Guardian Working on live-actio<br>49 7 0 49 Requests are closed This is a simple blog dedicated to A<br>50 5 42.261 50 Avatar & Aliens are the same movie - The O<br>51 5 38.234 51 FaceYourManga: Home Download and Print your avatar<br>52 5 43.293 52 Avatar Movie Review (2009) \| Plugged In Plugged<br>53 5 38.767 53 AvatarHD - Android Apps on Google Play Trong thế giới <br>54 5 38.655 54 Xbox Avatars – Windows Apps on Microsoft Store Use the X<br>55 5 38.767 55 VUDU - Avatar From Academy Award(R) winning<br>56 5 43.51 56 Avatar Fortress Fight 2 \| 1000 Free Flash Games | 46 5 39.332 48 Avatar \| Film \| The Guardian Working on live-a<br>47 5 38.726 45 Avatar Vectors, Photos and PSD files \| Free Do<br>48 5 38.767 46 Avatar Secrets: An Interactive Documentary fo<br>49 5 37.737 44 Avatar ... Woman's Invisible Jet) – Hugo<br>50 5 36.953 41 2045 Initiative The Dalai Lama Supports 2045's <b<br>51 5 38.767 53 AvatarHD - Android Apps on Google Play Trong thế giới <<br>52 5 38.234 51 FaceYourManga: Home Download and Print your ava<br>53 5 38.655 54 Xbox Avatars – Windows Apps on Microsoft Store Use th<br>54 5 38.767 55 VUDU - Avatar From Academy Award(R) winnir<br>55 5 5.5026 42 What is avatar? A Webopedia Definition (1) A v<br>56 7 0 49 Requests are closed This is a simple blog dedicated to <b |
| Frozen | 46 5 4.5183 46 Frozen Games Frozen Games, Play t<br>47 5 3.4454 47 Frozen Food and Power Outages: When to S<br>48 5 10.085 48 Frozen (Widescreen) - Walmart.com Bring h<br>49 5 3.4454 49 Frozen Food and Power Outages: When to S<br>50 5 19.951 50 Huffy 16" Girls' Disney Frozen Bik<br>51 5 10.085 51 Frozen (Widescreen) - Walmart.com Bring h<br>52 5 4.5183 52 Save 50% on Frozen Cortex on Steam A hard<br>53 5 19.309 53 Frozen Toys, Costumes, Gifts & Mercha<br>54 5 23.728 54 Shop for Disney Frozen & Licensed Cha<br>55 5 9.2977 55 Frozen \| Film \| The Guardian Film blog Let it<br>56 5 19.143 56 Frozen / Disney - TV Tropes Frozen | 41 5 9.2977 55 Frozen \| Film \| The Guardian Film bl<br>42 5 3.6486 27 'Frozen' director crushes a<br>43 5 3.661 28 'Frozen' director on Tarzan<br>44 7 0 21 Frozen \| zulily Frozen has be<br>45 5 4.5183 38 Frozen Synapse: A Simultaneous Tur<br>46 5 3.661 34 'Frozen' director on Tarzan<br>47 5 3.6486 35 'Frozen' Director Finally Cl<br>48 5 3.6486 40 frozen - Wiktionary frozen (<br>49 5 4.5183 46 Frozen Games Frozen Games<br>50 5 4.5183 52 Save 50% on Frozen Cortex on Steam<br>51 5 3.4454 47 Frozen Food and Power Outages: Wh |
| Taj mahal | 44 4 0 44 Taj Mahal Indian Restaurant The<br>45 8 60.782 45 India Agra Taj Mahal - YouTube A<br>46 8 60.538 46 Taj Mahal NewTaj.EOL<br>47 8 106.1 47 Taj Mahal needs nine-year mud pa<br>48 8 62.437 48 taj mahal: Latest News, Videos an<br>49 8 60.538 49 Rockport Music ‎€" Taj Mahal Aug<br>50 8 59.86 50 Taj Mahal Gardens Found to Align<br>51 8 61.849 51 Taj Mahal - A Tribute to Beauty -<br>52 8 67.372 52 Tourist reportedly dies at Taj Mahal<br>53 7 0 53 Tajmahal: The True Story - The H<br>54 8 59.121 54 Tourist falls to death while posing for selfie at | 44 8 62.437 48 taj mahal: Latest News, Videos<br>45 8 60.782 45 India Agra Taj Mahal - YouTub<br>46 8 58.715 42 Taj Mahal \| Biography, Albums<br>47 8 60.538 46 Taj Mahal NewTaj.EOL<br>48 8 61.849 51 Taj Mahal - A Tribute to Beaut<br>49 8 60.538 49 Rockport Music ‎€" Taj Mahal<br>50 8 59.86 50 Taj Mahal Gardens Found to Al<br>51 8 59.121 54 Tourist falls to death while posing for selfi<br>52 1 0 27 Tajmahal AR ... Sat - 5 pm to<br>53 4 0 44 Taj Mahal Indian Restaurant T<br>54 7 0 53 Tajmahal: The True Story - Th |
| Great wall | 46 8 55.034 46 Great Wall Szechuan House - Chinese Re<br>47 7 0 47 GW Supermarket ... Store Locator; \|; Employmer<br>48 8 41.448 48 The Great Wall: From Beginning to End:<br>49 8 55.181 49 Great Wall at Mutianyu (Beijing, China):<br>50 8 59.358 50 Great Wall Chinese Restaurant - Order O<br>51 8 55.063 51 GREAT WALL CHINESE RESTAURANT-FAR<br>52 8 54.285 52 BrainPOP \| Social Studies \| Learn about Grea<br>53 8 50.992 53 Great Wall - Order Online - Prince Georg<br>54 8 48.724 54 Great Wall - Order Online - Palm Coast -<br>55 8 40.494 55 Great Wall The first emperor of the Qin<br>56 8 73.987 56 Ancient China for Kids: The Great Wall - | 23 8 73.987 56 Ancient China for Kids: The Great Wall<br>24 8 54.949 25 Great Wall Marathon Annual maratho<br>25 8 50.992 22 Great Wall - Order Online - Danville -<br>26 8 58.163 32 Hiking China's Great Wall @ Nati<br>27 8 47.22 24 GREAT WALL - greatwall37.com GR<br>28 8 53.532 34 The Great Wall, China - Lonely Planet<br>29 8 57.581 43 Great Wall Chinese Restaurant - Orde<br>30 8 54.949 37 Great Wall of China - Enchanted Learn<br>31 8 59.358 50 Great Wall Chinese Restaurant - Orde<br>32 8 57.177 44 Great Wall Hiking Tours: Hike WILD <b<br>33 8 54.661 39 Off the Great Wall - YouTube Get you |

## 5. CONCLUSIONS

This paper proposes an effective web document ranking method using LDA classification scores to re-rank Google search returned webpages. This method downgrades web documents that have low classification scores or whose classes are not in the same category as the one related to the query, and on the other hand increases the ranks of web documents that have high classification scores. The experimental results show that the ranking of the returned web documents by the GCrank method was significantly better than the original Google ranking in terms of the ranking performance criteria described in Section 4, as indicated in Table 8. There is also evidence showing that the GCrank method can rank web documents more specific to user's information need. Therefore, our hypothesis about the LDA hyperplane has been successfully tested by the experiment, which states that if a point representing a web document is far away from the LDA hyperplane this web document should have a relatively high rank among the search engine returned web documents.

However, this paper focuses on improving the original Google ranking only, without comparing with other ranking methods. Subjective bias in the performance evaluation is another main limitation. For example, the performance evaluation usually depends on the queries used in the experiment and the judgment on the relevance of web documents with the original queries. We adopted multiple evaluation criteria from different perspectives to ensure a fair comparison and evaluation. However, further work should be conducted to overcome the limitations in this aspect of performance. It is noteworthy that with a limited number of topic categories and limited size of web documents tested in the experiment, this paper presents preliminary but promising results of re-ranking Google search returned web documents using classification scores. Deeper investigation and more extensive testing would be required in future research.

## REFERENCES


[1] Fonseca BM, Golgher PB, Moura ES, et al. Using association rules to discover search engines related queries. In the First Latin American Web Congress. USA; 2003. http://dx.doi.org/10.1109/laweb.2003.1250284

[2] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Google [Internet]. Available from: http://infolab.stanford.edu/backrub/google.html [Accessed 11 March 2016].

[3] Zhuang Z, Cucerzan S. Re-ranking search results using query logs. In International Conference on Information and Knowledge Management (CIKM). Virginia, USA; 2006. http://dx.doi.org/10.1145/1183614.1183767

[4] He B, Ounis I. Studying query expansion effectiveness. In European Conference on Information Retreival Research on Advances in Information Retreival (ECIR), Toulouse, France; 2009. http://dx.doi.org/10.1007/978-3-642-00958-7_57

[5] Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval: the concepts and technology behind search. England: Person Education Limited; 2011.

[6] Pan B. The power of search engine ranking for tourist destinations. Tourism Management. 2015; 47: 79-87. http://dx.doi.org/10.1016/j.tourman.2014.08.015

[7] Bratko A, Cormack GV, Filipic B, et al. Spam filtering using statistical data compression models. Machine Learning Research. 2006; 7: 2673-98.

[8] Busemann S, Schmeier S, Arens RG. Message classification in the cell center. In Applied Natural Language Processing Conference (ANLP). Seattle, Washington, USA; 2000. http://dx.doi.org/10.3115/974147.974169

[9] Santini M, Rosso M. Testing a genre-enabled application: a preliminary assessment. In the BCS IRSG Symposium: Future Directions in Information Access. London, UK; 2008.

[10] Haris BS, Guru DS, Manjunath S. Representation and classification of text documents: a brief review. International Journal of Computer Applications, Special Issue on Recent Trends in Image Processing and Pattern Recognition (RTIPPR). 2010; 2(2): 110-9.

[11] Trstenjaka B, Mikac S, Donko D. KNN with TF-IDF based framework for text categorization. Procedia Engineering. 2014; 69: 1356-64. http://dx.doi.org/10.1016/j.proeng.2014.03.129

[12] Mclachlan G. Discriminant analysis and statistical pattern recognition. Wiley Interscience; 2004.

[13] Kalina J, Tebbens JD. Algorithm for regularized linear discriminant analysis. In Biomedical Engineering Systems and Technologies. Lisbon, Portugal; 2015. http://dx.doi.org/10.5220/0005234901280133

[14] Plansangket S, Gan JQ. A new term weighting scheme based on CSDF for document representation and classification. In Computer Science and Electronic Engineering Conference (CEEC). Essex, UK; 2015.

[15] Du Y, Hai Y. Semantic ranking of web pages based on formal concept analysis. Journal of Systems and Software. 2013; 86(1): 187-97. http://dx.doi.org/10.1016/j.jss.2012.07.040

[16] Xiang B, Jiang D, Pei J, et al. Contex-aware ranking in web search. In International Conference on Research and Development in Information Retrieval (SIGIR). Geneva, Switzerland; 2010.

[17] Derhami V, Khodadadian E, Ghasemzadeh M, et al. Applying reinforcement learning for web pages ranking algorithms. Applied Soft Computing. 2013; 13(4): 1686-92. http://dx.doi.org/10.1016/j.asoc.2012.12.023

[18] Lu Y, Li Y, Xu M, et al. A user model based ranking method of query results of meta-search engines. In: International Conference on Network and Information Systems for Computers (ICNISC). Wuhan, China; 2015. http://dx.doi.org/10.1109/icnisc.2015.123

[19] Wang H, He X, Chang MW, et al. Personalized ranking model adaptation for web search. In: International Conference on Research and Development in Information Retrieval (SIGIR). New York, USA; 2013. http://dx.doi.org/10.1145/2484028.2484068

[20] Jindal V, Bawa S, Batra S. A review of ranking approaches for semantic searach on web. Information Processing and Management. 2014; 50(2): 416-25. http://dx.doi.org/10.1016/j.ipm.2013.10.004

[21] Garcia JM, Junghans M, Ruiz D, et al. Integrating semantic web services ranking mechanisms using a common preference model. Knowledge-Based Systems. 2013; 49: 22-36. http://dx.doi.org/10.1016/j.knosys.2013.04.007

[22] Lee J, Min JK, Oh A, et al. Effective ranking and search techniques for web resource considering semantic relationships. Information Processing and Management. 2014; 50(1): 132-55. http://dx.doi.org/10.1016/j.ipm.2013.08.007

[23] Manning CD, Raghavan P, Schutze H. Introductin to information retrieval. Cambridge, England: Cambridge University Press; 2008. http://dx.doi.org/10.1017/CBO9780511809071

[24] Alkhalifa E. Investigating bias in the page ranking approach. In International Conference on Information and Communication Technology Research (ICTRC). Abu Dhabi, the United Arab Emirates; 2015. http://dx.doi.org/10.1109/ictrc.2015.7156480

[25] Balakrishnama S, Ganapathiraju A. Linear discriminant analysis - a brief tutorial. In: International Symposium on Information Processing; 1998.

[26] Fisher RA. The user of multiple measurements in taxonomic problems. Annals of Eugenics. 1936; 7(2): 179-88. http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x

[27] Bordag S. A comparison of co-occurrence and similarity measures as simulations of context. In: International Conference on Computational Linguistics and Intelligent Text Processing. Springer-Verlage Berlin, Heidelberg; 2008. http://dx.doi.org/10.1007/978-3-540-78135-6_5

[28] Powell WB. Approximate dynamic programming: solving the curses of dimensionality, Wiley-Interscience; 2007. http://dx.doi.org/10.1002/9780470182963

[29] Gan JQ, Hasan BAS, Tsui CSL. A filter-dominating hybrid sequential forward floating search method for feature subset selection in high-dimensional space. International Journal of Machine Learning and Cybernetics. 2014; 5(3): 413-23. http://dx.doi.org/10.1007/s13042-012-0139-z

[30] Plansangket S, Gan JQ. A query suggestion method combining TF-IDF and Jaccard coefficient for interactive web search. Artificial Intelligence Research. 2015; 4(2): 119. http://dx.doi.org/10.5430/air.v4n2p119

[31] Otegi A, Arregi X, Agirre E. Query expansion for IR using knowledge-based relatedness. In International Joint Conference on NLP. ChangMai, Thailand; 2011.

[32] Jurafsky D, Martin JH. Speech and language processing: an introduction to natural language processing. Prentice Hall; 2008.

[33] Okabe M, Yamada S. Semisupervised query expansion with minimal feedback. IEEE Transactions on Knowledge and Data Engineering. 2007; 19(11): 1585-9. http://dx.doi.org/10.1109/TKDE.2007.190646