Evaluating the effectiveness of information retrieval systems using effort-based relevance judgment

Author(s): Prabha Rajagopal, (University of Malaya, Kuala Lumpur, Malaysia)

Abstract:

Purpose

The effort in addition to relevance is a major factor for satisfaction and utility of the document to the actual user. The purpose of this paper is to propose a method in generating relevance judgments that incorporate effort without human judges’ involvement. Then the study determines the variation in system rankings due to low effort relevance judgment in evaluating retrieval systems at different depth of evaluation.

Design/methodology/approach

Effort-based relevance judgments are generated using a proposed boxplot approach for simple document features, HTML features and readability features. The boxplot approach is a simple yet repeatable approach in classifying documents’ effort while ensuring outlier scores do not skew the grading of the entire set of documents.

Findings

The retrieval systems evaluation using low effort relevance judgments has a stronger influence on shallow depth of
evaluation compared to deeper depth. It is proved that
difference in the system rankings is due to low effort
documents and not the number of relevant documents.

Hence, it is crucial to evaluate retrieval systems at shallow
depth using low effort relevance judgments.

Keywords: Information system, Information retrieval, TREC, Large-scale experimentation, Relevance judgments, System-oriented evaluation

Type: Research paper

Publisher: Emerald Publishing Limited

Received: 19 April 2018

Revised: 18 July 2018, 02 September 2018

Accepted: 13 September 2018

Acknowledgments: This research was supported by UMRG Programme RP059A-17SBS from Universiti Malaya and the Ministry of Higher Education, Malaysia.

Copyright: © Emerald Publishing Limited 2019

Published by Emerald Publishing Limited

Licensed re-use rights only


Downloads: The fulltext of this document has been downloaded 474 times since 2019

Article

1. Introduction

There are two categories of information retrieval evaluation, the system-oriented evaluation, and user-oriented evaluation. One of the important aspects of the system-oriented evaluation is the relevance judgment of a test collection containing information on the relevancy of documents concerning queries. In the TREC environment, topic experts judge the documents in regards to each
query. In user-oriented evaluations, the interaction of the actual users with the retrieval systems is measured. Nonetheless, one of the disagreements between both the information retrieval evaluation categories is the consideration of relevance by the expert judges and the utility of the documents to actual users (Yilmaz et al., 2014).

System-oriented evaluation has always prioritized relevance for user satisfaction. Therefore, relevance is used as a way of measuring the effectiveness of retrieval systems. However, recent studies highlighted that effort in retrieving relevant documents is equally important for user satisfaction (Verma et al., 2016; Yilmaz et al., 2014). The effort in this context is referring to the amount of work needed by the user to find and identify the relevant content in the document. The effort can be classified as low effort or high effort. Low effort indicates less amount of work needed by the user to identify the relevant context within a document. Meanwhile, the high effort requires more work by the user to identify the relevant content within a document.

Real users easily give up and do not put in as much effort as expert judges while identifying relevancy in a document (Villa and Halvey, 2013; Yilmaz et al., 2014). Therefore, a retrieval system incorporating retrieval of low effort documents is preferred by the user as it requires less effort in identifying relevancy of the documents compared to a retrieval system retrieving high effort documents. The effort in addition to relevance is a major factor for satisfaction and utility of the document to the actual user. Consequently, it is vital to evaluate the retrieval systems based on the amount of effort needed to identify the relevancy of documents to ensure user satisfaction.

The importance of effort is measured in various ways in previous studies (Verma et al., 2016; Villa and Halvey, 2013) but limited depth of evaluation and retrieval systems (top ten only) were used to show the differences in system rankings due to low effort relevance judgments. The differences in system rankings using original and low effort relevance judgments beyond evaluation depth 10 within a test collection are unknown. Possibilities exist that the differences in system rankings may vary widely as a result of larger numbers of relevant documents found with deeper depth of evaluation. Therefore, this study questions how the system rankings change when evaluated beyond evaluation depth 10 using low effort relevance judgment. There are likelihoods for real users, depending on the type of user using the search engine, to look beyond ten or as many retrieved documents to fulfill their query (Sanderson, 2010). Hence, it is necessary to evaluate the retrieval systems beyond evaluation depth 10. Consequently, this study attempts to explore deeper depth of evaluation (up to 1,000) as opposed to a previous study (Verma et al., 2016) which only evaluated at depth of 10 using effectiveness metrics P@10. Besides, this study also questions if any change in the system rankings arise from the low effort relevance judgments. Changes in
system rankings could also be due to number of relevant documents. Thus, it is needed to determine the cause of system rankings changes, if any.

Attempts to reduce the amount of work needed for judging relevance had been an attention of the research community. Any advancement that reduces the workload on relevance judgments without jeopardizing the quality of evaluation is an added advantage (Guiver et al., 2009). Nevertheless, inquiring the amount of effort needed for relevance judgment from the judges (Verma et al., 2016; Yilmaz et al., 2014) may cause variation in judgments (Carterette and Soboroff, 2010; Chandar et al., 2013; Scholer et al., 2011; Webber et al., 2012) and the effort needed. Possible advancements in overcoming these drawbacks are minimizing or eliminating the involvement of human judges in obtaining effort information. Therefore, this study attempts to create the relevance judgment without human participation which is different from previous studies (Verma et al., 2016; Yilmaz et al., 2014) involving human assessors.

This study aims to propose a method in generating relevance judgments that incorporate effort without involvement of human judges. The study also aims to determine the variation in system rankings due to low effort relevance judgment in evaluating retrieval systems at different depth of evaluation. The major contribution of this study is to highlight the importance of low effort relevant documents in retrieval system evaluation at different depth of evaluation.

Section 2 is the research background discussing the importance of effort in user satisfaction and the effect of relevance to system performance. Section 3 is the effort feature classification and Section 4 addresses document grades classification using the boxplot approach. The results and discussion section are included next. Finally, the conclusion is drawn and the future work is proposed.

2.1 The importance of effort in addition to relevance

Between the user-oriented and system-oriented information retrieval evaluation, preference has always been for the latter by researchers because of repeatability, short experimental turnaround, and low cost. In the system-oriented evaluation, relevance had always been a priority in measuring the performance of the systems using effectiveness metrics. Relevance has been thought to be the utility a user gains when viewing a ranked document. Therefore, a system that ranks relevant documents earlier in the ranking list is considered a more effective system (Moffat et al., 2012). Hence, an effectiveness metric should capture the gist of ranked relevant documents to score the retrieval systems.
However, there exist cases of non-correlation between the effectiveness metrics and the real user experience (Moffat et al., 2012). Furthermore, the outcome of system-based evaluation does not agree with real user satisfaction (Hersh et al., 2000; Turpin and Hersh, 2001). It is due to the factors such as effort, system and user effectiveness, and user characteristics (Al-Maskari and Sanderson, 2010) influencing user satisfaction. The effort could be divided into findability, readability and understandability factors (Verma et al., 2016). Recently, studies measured effort needed for judging relevant documents (Villa and Halvey, 2013), and identified effort as an important aspect of user satisfaction in addition to relevance (Yilmaz et al., 2014).

Experimentation on the effort needed for judging relevant documents showed that increased document size and the degree of relevance of a document indicating “relevant” requires leads to more effort in judging (Villa and Halvey, 2013). The experimentation was conducted using TREC HARD 2005 track and AQUANT test collections. During and after judging, the NASA TLX was used to gather the judges’ action and perception of the task effort. The online study assumed that a user would look through the ranked list in ascending order. Additionally, behavioral data through the time taken to make judgments and number of topic view clicks were measured. Besides the amount of effort needed for judging, the study also showed accuracy is not affected by document length but by the degree of relevance (Villa and Halvey, 2013).

Experimentation was conducted to measure the effort needed by users to identify and consume the information from a document with regard to time (Yilmaz et al., 2014). When compared, they found a mismatch between relevance judgments from real users and expert judges. The experimentation utilized dwell time and click counts to obtain real users judgments. They argue that utility of a document to actual users is dependable in the effort needed to consume the relevant information (Yilmaz et al., 2014).

Another study focused on measuring readability, findability and understandability effort in obtaining relevance judgment and stated that effort should be a part of relevance judgment if user satisfaction is prioritized (Verma et al., 2016). They even indicated user satisfaction is a function of relevance and effort. High effort documents are harder to consume. Thus it is less likely for the users to read it (Yilmaz et al., 2014). Experimentation also resulted in a high correlation of satisfaction and findability with user preferences compared to readability and understandability (Verma et al., 2016). It was previously known that readability affects assessor disagreement, whereby documents that are easy to read causes disagreements (Chandar et al., 2013) Also, Collins-Thompson (2011) advocated that reading difficulties of documents are a major factor in large-scale analyses of personalized information retrieval systems.
The logistic regression models are used to predict the relevance assessment using initial user input on relevance and effort (Chandar et al., 2013; Verma et al., 2016; Yilmaz et al., 2014). With the incorporation of effort into relevance judgments, Kendall’s \( \tau \) correlation coefficient shows the performance of the retrieval systems are different compared to when the effort is not used in the relevance judgments (Verma et al., 2016). The result highlights the importance of effort together with relevance in satisfying user’s information needs.

The various studies highlight different pieces to effort needed in identifying relevant documents involving human to generate the relevance judgments. Based on the studies, there exist mismatches between real users and expert judges in identifying relevance. Meanwhile, document size, degree of relevance, readability, understandability and findability are important aspects of effort.

### 2.2 The effect of relevance on system performance

Recently, a study (Jones et al., 2014) showed that the number of relevant documents is not an effect of evaluation inconsistencies. Evaluation inconsistencies exist with different sub-collections formed based on document source, but the number of relevant documents was uniform in each sub-collection. Nevertheless, the study did not state the effects causing the inconsistencies in the evaluation.

Previously, Voorhees (2001) mentioned the change in effectiveness could be due to small numbers of relevant documents causing the retrieval measures to be unstable. Queries with fewer than five relevant documents could even cause the average precision measure to become unstable causing variation to the system performance and rankings (Voorhees, 2000). Yet with a constant number of relevant documents, the evaluation results are inconsistent (Jones et al., 2014). A possible cause of inconsistency could be the changes in document ranking. Since small changes in document ranking can produce large variations in the evaluation scores (Voorhees, 2000).

### 3. Classifying effort features and the boxplot approach for document grading

Many factors influence user satisfaction (Al-Maskari and Sanderson, 2010) and these factors can be broken down into measurable features. Simple document features, reading level features and metarank feature have been used to predict assessor disagreement (Chandar et al., 2013). Meanwhile, readability level of the document, document length and location of query terms in the document was used to measure effort needed in judging a document (Yilmaz et al., 2014).

Below are the effort features experimented in this study and the possible effort grades for a document implemented using boxplot or standardized grading. These features were selected because some of them showed an impact on user
satisfaction (Verma et al., 2016) and relevance disagreements (Chandar et al., 2013):

1. Simple document features:
   - Number of sentences in a document (nsent) – calculates the number of sentences in the document, where sentences end with a period, question mark or exclamation mark followed by a space. A low number of sentences indicate low effort.
   - Number of words in a document (nwords) – calculates the number of words in a document. The number of words is counted by each space (Smith and Senter, 1967). A low number of words indicate low effort.
   - Number of characters in a document (nchar) – calculates the number of characters in a document. It is calculated by counting the number of alphabetic, numeric and punctuation characters. A low number of characters indicate low effort.

2. HTML features:
   - Number of images (nimg) – calculates the number of images within a document. Zero image is taken as high effort while the remaining number of images classification will follow the boxplot approach. Images are useful in spotting information in a document, and more images are better (Verma et al., 2016). At the same time, too many images may not be suitable to capture the relevant content of the document.
   - Number of tables (ntab) – calculates the number of tables within a document. Zero table is taken as high effort while the remaining number of tables classification will follow the boxplot approach. Similar to images, tables are helpful in identifying information but too many tables make it difficult to find the information needed (Verma et al., 2016).

3. Readability features:
   There are three different readability features considered for the study. The classification of these features could have many possible variations. However, for simplicity, only a single variation or classification was considered for this study:
   - ARI – Automated Readability Index (ARI) measures the readability of a document and can be defined by the following formula to obtain the grade level (GL) of a document (see Table I):

     \[
     GL = 0.5 \left( \frac{\text{number of words}}{\text{number of sentences}} \right) + 4.71 \left( \frac{\text{number of characters}}{\text{number of words}} \right)
     \]

     The number of sentences can be tabulated when ending with a period, question mark or exclamation mark and followed by space (Smith and
The number of words is counted each time the space bar is depressed. Characters or strokes are advanced one count for each printing-key on the keyboard which includes all the letters, numbers and punctuation marks (Smith and Senter, 1967). Low-ARI level or GL of a document requires low effort. The following classification will be made for each of the grading and ARI levels in this experimentation:

- CLI – Coleman-Liau (CLI) is also a readability feature and is defined by the following formula:

\[
CLI = 0.059L - 0.296S - 15.8,
\]

where, L is the average number of letters per 100 words and S is the average number of sentences per 100 words. The CLI is similar to ARI. A low CLI level indicates low effort. The following effort classifications are made for each CLI level (see Table II):

- LIX – LIX is a readability measure to calculate the difficulty of reading a document. The number of periods is defined as period (full stop), colon or capital first letter. Words with six or more letters are considered as long words. It can be computed using the following formula:

\[
LIX = \frac{\text{number of words}}{\text{number of periods}} + \frac{(\text{number of long words} \times 100)}{\text{number of words}}.
\]

A low LIX score indicates low effort. The following effort classifications are made for each LIX scores (see Table III):

- LIX for sentences with query terms (LIXsent) – calculates the LIX on sentences which have query terms. Query terms are terms that are from the title-only query (topic) which appears in sentences within a document. The classification of LIXsent effort feature is similar to LIX.

The classification of effort features is used to grade documents with the boxplot approach implementation for simple document and HTML effort features, and standardized grading approach for readability features. The readability effort features have standardized grades and will be classified in a possible equivalent manner. For both ARI and CLI, high effort constitutes to college level. The LIX, on the other hand, classifies "very difficult" as high effort with the assumption this level is equivalent to "college" level from CLI and ARI features. Therefore, all three readability effort features tend to classify the documents’ effort in a somewhat equivalent manner. Previously, it has been highlighted that emphasizing the GLs may be unfavorable of reading adults (Smith and Senter, 1967). According to the National Adult Literacy Study (National Center for Educational Statistics, 1993), an average adult in the USA reads at the seventh GL
Therefore, the grades classified as school age reading abilities does not necessarily apply to children but is relevant for adults as well. In addition, the reading level is not a continuous straight-line function above the seventh grade although the index continuous to increase. Hence, GL equivalents of high school or college is less important (Smith and Senter, 1967).

A systematic grading versatile for multiple test collections is needed. This study attempts to grade and classify the effort features in a possible simple manner to create the effort-based relevance judgments. The effort measurements are only for relevant documents since effort is a measure that can satisfy users only when the document is relevant (Verma et al., 2016). This study adopts the same by calculating effort for relevant documents only.

4. Document grades classification using the boxplot approach

The boxplot usually identifies outliers and data plots within 25 and 75 percent interquartile range which is based on the Tukey outlier detection mechanism. The suspected outliers and outliers are not removed from the evaluation but classified accordingly. The scores within 0 and the upper inner fence are divided into two to create binary relevance. The feature scores could only start from 0 and will not hold negative values, as it does not make sense to have negative scores. For example, the number of words could never be a negative value.

The boxplot approach is adapted into this experimentation to avoid skewing of the grades of the documents in test collections. In a test collection, there may be documents that could be very lengthy or contain many difficult words. These document grades may be very different compared to other documents within the test collections. It is important that these documents do not distort the evaluation of the remaining documents within the test collections. Therefore, the boxplot approach is adapted in this experimentation to overcome the skewing of document grades and allow a fair division of the document effort grades.

However, the boxplot is only implemented for simple document features and HTML features. These features do not have a standard guideline as to which level of measurement constitutes to low or high grades. In contrast, the readability features have standardized grades that could be easily suited to this study.

By using the boxplot approach, the classifications of effort are dependable on the set of documents from that particular test collection. A drawback arising from the usage of boxplot for document grading is that the test collections may not be comparable one-to-one. It is because a document graded as a low effort in one test collection may not necessarily be graded the same in another test collection. The variation occurs based on the relevant documents within the test collection and the interquartile range computation. This shortcoming is only valid for the simple document features and HTML features because there is no...
standard of classification to indicate low or high effort documents. In contrast, the readability effort features are easily comparable between test collections due to the fact there are standardized grades. Nonetheless, the correlation coefficient of system rankings can be measured reliably within the test collection and should not cause discrepancies.

The boxplot approach is shown in Figure 1. In the boxplot approach, calculate the document scores for each feature. Then sort the document scores for each feature in ascending order. Next, calculate the interquartile range of the feature scores by subtracting the lower quartile \((0.25 \times (N+1))\) from the upper quartile \((0.75 \times (N+1))\). Following this, calculate the upper inner fence score using the formula below:

\[
\text{upper inner fence} = \text{upper quartile} + (1.5 \times \text{interquartile range}).
\]

Then, divide the upper inner fence feature score into two for translation as binary relevance. The lower half indicates low effort (marked as relevant) and the upper half indicates high effort (marked as non-relevant). Any document that has a feature score that falls within the lower half will be marked as low effort relevant while the remaining is characterized as high effort non-relevant.

The outliers and suspected outliers are those document scores above the upper inner fence. The outlier and suspected outlier documents will be marked as high effort, thus non-relevant since their scores lie beyond the upper inner fence. With the grading of the documents based on their effort feature scores, a low effort relevance judgment can be created.

The boxplot approach is suitable for use in classifying effort features because it eliminates document score distortion due to outliers. Second, the simple document features and HTML features do not have any standardized grading similar to those readability features. Therefore, the approach is not stringent and appears to be suitable for classifying document effort easily.

Two different test collections were selected for use in this experimentation. They are TREC-9 Web track and TREC-2001 Web track. In TREC-9 test collection, 40 retrieval system runs used title-only in their retrieval process and contains 50 topics. The TREC-2001 Web track has 77 title-only system runs and also contains 50 topics. The topics from the TREC-2001 have a similar form with TREC-9 topics, but the topic title for the prior is a real Web query taken from search engine logs (Hawking and Craswell, 2002). Both the test collections used the WT10g data set which comprises of Web data. Table IV shows details on number of relevant documents, average number of images, tables, sentences, words and characters. The table also details for both test collections and separated by the depth of documents (Top 10, 100 and 1,000).
The results and discussion section is divided into three subsections. The first subsection contains the results of correlation coefficient evaluation of individual effort features for various depths of evaluation. The second subsection details the trend of correlation coefficient with regard to depth of evaluation. The final subsection analyses the cause of changes in system rankings by percentage of agreement of relevant documents between the original and effort-based relevance judgments.

5.1 Correlation coefficient evaluation of individual effort features for different depth of evaluation

The impact on the system rankings will be evaluated with Kendall’s $\tau$ correlation coefficient to determine the variation in the system rankings. Low effort relevance judgments (eqrel) were created for each of the effort features separately and one with all features combined (labeled as AllFeature). The eqrel for AllFeature uses the feature grades of the documents from the individual effort features and deduce a final grade using mode. The performance of the retrieval systems is then measured using these low effort relevance judgments. Also, note that the effectiveness scores measured from the eqrel will always be lower than or at most equal to that of the qrel because low or high effort classifications are on relevant documents only. Nonetheless, the correlation coefficient should not be affected by the lower effectiveness scores using eqrel since Kendall’s $\tau$ uses system ranks for evaluation.

The experimentation measured Kendall’s $\tau$ correlation coefficient between the system rankings of original and low effort relevance judgments. The Kendall’s $\tau$ is commonly used in information retrieval field to measure the similarities of rank orderings. A strong correlation coefficient ($\geq 0.8$) would indicate that the system rankings are not affected by the effort feature while moderate to poor correlation coefficient ($<0.8$) would indicate the system rankings are affected by the effort feature. The intention of such observation is to recognize any changes in the correlation coefficient due to the depth of evaluation.

Figure 2 shows Kendall’s $\tau$ correlation coefficient for the various effort features evaluated at different depths for both test collections. Two different metrics, $P@k$ and $AP@k$, were used for a combination depth, $k=10, 100$ or $1,000$. The $x$-axis represents the individual features and AllFeature, and the $y$-axis represents Kendall’s $\tau$ correlation coefficient values. The dotted horizontal lines in the graph indicate the $\tau$ value of 0.8.

First, the plots for TREC-9 $P@k$ show that depth of evaluation using eqrel affect the system rankings for some features. The system rankings for features such as LIX, LIXsent and ntab are all influenced by effort for all depth of evaluation. Whereas, the system rankings for features CLI, nimg and nsent are not impacted by the evaluation depth as all $\tau$ values fall at or above 0.8. As for the remaining
features, ARI, nchar, nword and AllFeature, the effort impacts the system rankings for some depth of evaluation.

As for the assessment with TREC-9 AP@k metric, AllFeature, CLI, LIX, LIXsent and ntab features have Kendall’s $\tau$ values below 0.8 for all depth of evaluation. These features show the low effort relevance judgments have now caused changes in the systems’ performances regardless of the evaluation depth. The remaining features show varying effect on the system rankings as some correlation coefficient values appear above while others are below 0.8, depending on the depth of evaluation.

For TREC-2,001, it can be observed that all the features have caused an impact on the system rankings due to low effort relevance judgments, for both P@k and AP@k metrics. All the effort features have Kendall’s $\tau$ values below 0.8. The effort features that affect the system rankings have been identified through the correlation coefficient values. However, did the correlation coefficient become stronger or weaker with the depth of evaluation?

5.2 The trending of correlation coefficient with depth of evaluation

For TREC-9 P@k, the increase in depth from 10 to 100 or 1,000 has caused the correlation coefficient to remain equally good or stronger for ARI, CLI, LIXsent and nimg features. The increase in Kendall’s $\tau$ values with the increase in depth of evaluation shows effort does not necessarily impact system rankings with deeper depth of evaluation. Meanwhile, AllFeature, LIX, nchar, nsent, nword and ntab have all caused Kendall’s $\tau$ values to decrease with the increase in evaluation depth. A possible cause of the correlation coefficient changes could be the number of relevant documents in the different depth of evaluation.

Analyzing the same for TREC-9 AP@k shows that features ARI, nchar, nword, nsent and nimg have a stronger correlation coefficient with deeper depth while the correlation coefficient gets weaker for CLI, LIX, LIXsent and ntab. The only exception is AllFeature that show the depth of 100 has weaker tau compared to depth 10, and stronger $\tau$ for depth 1,000 compared to the other two depths. The possible cause of such changes could be the low effort relevant documents or the number of relevant documents due to deeper evaluation.

For TREC-2,001, it can be observed that all the features have caused an impact on the system rankings due to low effort relevance judgments, for both P@k and AP@k metrics. All the effort features have Kendall’s $\tau$ values below 0.8. The features LIX, LIXsent, nimg and ntab have all caused the correlation coefficient to increase with depth of evaluation. The trend is same for both P@k and AP@k. Such trend indicates low effort relevance judgments may have a lesser influence on the system rankings with deeper evaluation. Also, evaluating with metric AP@k has an increasing correlation coefficient with deeper evaluation for AllFeature. However, evaluation of AllFeature using metric P@k displays decreasing.
correlation coefficient values with increased evaluation depth. The same observations appear for nchar effort feature when assessed using P@k and AP@k with the addition to nword feature for AP@k.

When assessed using P@k, the ARI and CLI features both produce consistent correlation coefficient despite the change in evaluation depth. The results indicate the system rankings are consistent or rather the effectiveness of the systems changed consistently among the different depth of evaluation such that the correlation coefficient hardly changed. As for the other features such as nsent and nword evaluated using P@k, the results show varying correlation coefficient with the increase in depth while the ARI, CLI and nsent features evaluated using AP@k also shows the same trend. The Kendall’s $\tau$ value is weaker for depth 100 compared to depth 10 but Kendall’s $\tau$ values for depth 1,000 is always stronger than that of depth 10.

Evaluating at depth 100 using low effort relevance judgment appears to have a larger influence on the system rankings. Most likely the change in relevant documents between ranks 11 to 100 using eqrel have caused this alteration in the system rankings, and thus lowering Kendall’s $\tau$ correlation coefficient.

Although the results from both the test collections do not show exact replication, it confirms influences of low effort relevance judgments to system rankings for different depth of evaluation. More times, the depth of 10 has weaker correlation coefficient than the remaining two depths experimented. The results reveal that system evaluation using low effort relevance judgments may have a lesser impact on the system rankings for deeper evaluation depth.

In the past study (Verma et al., 2016), Kendall’s $\tau$ correlation coefficient for system rankings generated from P@10 original and low effort relevance judgments show $\tau$ values between 0.53 and 0.71 for TREC Web Track ad hoc task 2012–2014 test collections. Also, only the correlation coefficient for top 10 systems was considered. When considering all systems for evaluation, the results from this study do indicate some similarities and differences with the past study for various features. It is also important to highlight that Verma et al. (2016) used a linear regression model to create the low effort relevance judgment. The current study, however, uses a boxplot approach and a standardized grading for readability features to determine the low effort relevant documents. Nonetheless, this study explored low effort relevance judgment with regard to depth of evaluation and highlighted a stronger influence of effort in shallow evaluation compared to deeper evaluation.

Analyses of Kendall’s $\tau$ results have demonstrated the variation in system rankings due to effort for different depth of evaluation. It cannot be shunned that the change in system rankings could have been due to the number of relevant documents in the relevance judgments. A total number of relevant documents
influence the performance of the systems (Voorhees, 2001) and thus their ranking as well. Nevertheless, it is also possible that the number of relevant documents is not affecting the evaluation inconsistencies (Jones et al., 2014). An analysis is necessary to reestablish the cause of changes in correlation coefficient due to evaluation depth. Further analysis attempts to determine whether a change in the system rankings is due to low effort relevant documents or the number of relevant documents in the eqrel?

5.3 Percentage of agreement of relevant documents between the original and effort-based relevance judgments

The percentage of agreement of relevant documents between qrel and eqrel can indicate the proportion of relevant documents in eqrel. The percentage of agreement between the qrel and eqrel is measured separately for each feature. Variations in the percentage of agreement could only happen due to the relevant documents. The reason for this is that low effort relevant documents in the eqrel originated from the qrel relevant documents. Meanwhile, those relevant documents graded as high effort become non-relevant in the eqrel. Hence, they would not have a relevancy match with the qrel. If the percentage of agreement is high, there are more low effort relevant documents that match the existing relevant documents from qrel. If the percentage of agreement is low, there are lesser low effort relevant documents from eqrel that matches the existing relevant documents from qrel.

The analysis determines if the change in the correlation coefficient (due to system rankings changes using eqrel) with regard to evaluation depth is due to the influence of low effort relevant documents instead of just the number of relevant documents. If Kendall’s $\tau$ value increases with the percentage of agreement, the change in $\tau$ could have occurred due to the increase in the number of relevant documents in the eqrel. In contrast, if Kendall’s $\tau$ value increases while the percentage of agreement remains same, the number of relevant documents is unlikely the cause. Similarly, if the percentage of agreement increases but Kendall’s $\tau$ value remains consistent, the effect is not due to the increase in the number of relevant documents.

The results of the analysis are shown in Figure 3 for TREC-9 and Figure 4 for TREC-2,001 respectively. The x-axis represents Kendall’s $\tau$ values, and the y-axis represents the percentage of agreement (relevant documents). The graphs are divided into two rows; the top row is for metric $P@k$ and the bottom row for $AP@k$. The columns represent each effort feature experimented and are labeled accordingly along the top of each row.

From Figure 3, features ARI, CLI, LIX, nchar, nimg, nsent and nword have a similar tabulation of plots for $P@k$. The position of the plots appears almost horizontal and close to each other. Horizontal positioning implies minimal changes in the percentage of agreement that causes shifts in Kendall’s $\tau$ value. It means the
The correlation coefficient of the system rankings is changing despite the constant percentage of relevant documents. The same is seen for AP@k for features ARI, CLI, LIX, LIXsent and nimg. Therefore, the change in the system rankings is not due to number of relevant documents.

The features nchar, nsent and nword for AP@k have a decreasing percentage of agreement for deeper evaluation compared to shallow depth. However, the declining percentage of agreement results in stronger Kendall’s $\tau$ value. Decreasing percentage of agreement means there are lesser relevant document matches between eqrel and qrel but still yields strong correlation coefficient. The analysis suggests the number of relevant documents is not causing the change in system rankings. However, due to a lower percentage of agreement, the effectiveness scores could be much lower for the systems as well. The change in effectiveness could be due to small numbers of relevant documents causing the retrieval measures to be unstable (Voorhees, 2001). Nonetheless, these features actually have approximately 80 percent of agreement with qrel. Also, deeper depth has a lower percentage of agreement but stronger Kendall’s $\tau$. Hence, the system rankings change is likely due to the low effort relevant documents.

The LIXsent feature for P@k has an almost vertical plot. There are small changes in the percentage of agreement, but Kendall’s $\tau$ value remains consistent. In fact, the percentage of agreement decreases slightly with the depth of evaluation. It could mean the similarity of ranks between the qrel and eqrel did not change with the depth. Despite variations in the number of relevant documents, the system rankings are not impacted. In this case, the change in system rankings is due to low effort relevant documents instead of just the number of relevant documents.

As for ntab feature, for both P@k and AP@k, a higher percentage of agreement yields better Kendall’s $\tau$ value. In fact, the shallow depth evaluation appears to have a better percentage of agreement compared to deeper evaluations. Such scenario could indicate the number of relevant documents is the cause positive change in system rankings. AllFeature has a similar observation. Although the changes in Kendall’s $\tau$ appears to have occurred due to the number of relevant documents, it is important to note, the grading of AllFeature was contributed by all the individual effort features. Almost all the features, except ntab, demonstrated the effect of low effort relevant documents to Kendall’s $\tau$. Hence, it is acceptable to state the change in system rankings for the AllFeature is also due to the low effort relevant documents and not just the number of relevant documents.

The results for TREC-2001 (Figure 4) show most features (AllFeature, LIX, LIXsent, nchar, nsent, ntab, and nword for P@k, and AllFeature, ARI, CLI, LIX, LIXsent, nchar, nsent, ntab and nword for AP@k) have their plots positioned horizontally. The horizontal positioning means small changes in the percentage of agreement
while Kendall’s r correlation coefficient has significant changes. As stated earlier, such results did not occur due to the number of relevant documents but most likely due to the low effort relevant documents. Similarly, low effort relevant documents contribute to the observations in P@k for features ARI and CLI. These features have their plots positioned vertically. Even though the percentage of agreement decreases with the depth of evaluation, Kendall’s r value remains consistent across the depths. Therefore, the number of relevant document did not cause the changes in system rankings.

The only feature that could suggest the role of relevant documents’ number to the variation in Kendall’s r is the nimg feature. The percentage of agreement increases with the depth, which could mean more relevant documents in eqrel. Also, these higher percentage of agreement appears to have a positive influence to Kendall’s r correlation coefficient. However, the AllFeature has its plots positioned horizontally. It is different compared to TREC-9 observation. Horizontal positioning is a clear indication that the number of relevant documents is not the cause of changes in system rankings. The analysis directly demonstrates the system rankings change is due to low effort relevant document.

The experimental results reveal that low effort relevance judgments cause changes to the system rankings at different depths of evaluation. Deeper depth of evaluation also tends to have a lesser impact on the system rankings due to low effort relevance judgments compared to shallow depth evaluation. The analysis confirmed the correlation coefficient changes were caused by low effort relevant document instead of just the number of relevant documents in the eqrel. Sometimes, a small number of relevant documents produced better or consistent Kendall’s r. The position of the low effort relevant document could have impacted the system rankings. Another possibility is the magnitude of difference in the effectiveness scores. Due to lower numbers of relevant documents in the eqrel, the effectiveness scores of the systems are lower than those from original qrel. Therefore, the difference in scores could have changed consistently, such that the system ranks did not vary between using qrel and eqrel.

6. Conclusions and future work

This study generated and evaluated retrieval systems with low effort relevance judgments at different depths of evaluation using the proposed boxplot and standardized grading approaches. The retrieval systems evaluation using low effort relevance judgments has a stronger influence on shallow depth of evaluation compared to deeper depth. It is proved that difference in the system rankings is not due to just the number of relevant documents. At times, low numbers of relevant documents have shown strong correlation coefficient results compared to the higher number of relevant documents.
It can be concluded that low effort relevance judgment is the cause of poor correlation coefficient between the system rankings from original and low effort relevance judgments. The rank position of these low effort relevant documents has an influence on the variation of the system rankings. Therefore, it is crucial to evaluate retrieval systems at shallow depth using low effort relevance judgments. Besides, it is necessary to use low effort relevance judgments in evaluating retrieval systems to determine retrieval systems that satisfy users.

However, this study mainly focuses on retrieval systems similar to the web. It is unknown if similar outcomes can be observed on other type of test collections. Nevertheless, the results obtained from one web test collection are consistent to that of another web test collection. Hence, the conclusions can be generalized to other web test collections due to similar characteristics of the test collections. The experimentation on two test collections also ensures repeatability and consistency of the results. Thus, the outcomes and conclusions drawn from the experimentation are considered reliable.

This study could be extended to future work by including other effort features such as summary features, document specific features, outlink-oriented features, query specific features and query term window specific features (Verma et al., 2016) in evaluating retrieval systems. Other than that, retrieval of documents could be personalized to the user’s reading capabilities (Collins-Thompson, 2011) with the incorporation of reading effort of a document.
Figure 3 TREC-9 – Percentage of relevant documents agreement vs Kendall’s $r$ correlation coefficient for different depth of evaluation

Figure 4 TREC-2001 – Percentage of relevant documents agreement vs Kendall’s $r$ correlation coefficient for different depth of evaluation

Table I Effort classification for ARI feature

Table II Effort classification for CLI feature

Table III Effort classification for LIx feature

Table IV Number of total relevant documents, HTML features and simple document features

References


Corresponding Author

Sri Devi Ravana can be contacted at: sdevi@um.edu.my

Related articles

Ranking retrieval systems using pseudo relevance judgments
Sri Devi Ravana et al., Aslib Journal of Information Management

A library’s information retrieval system (in)effectiveness: case study
Robert Marijan et al., Library Hi Tech

Evaluation of semantic retrieval systems on the semantic web
Luis Morato Jorge et al., Library Hi Tech

A machine learning approach to query generation in plagiarism source retrieval
Lei-lei Kong et al., Frontiers of Information Technology & Electronic Engineering

Tumor clonality and resistance mechanisms in EGFR mutation-positive non-small-cell lung cancer
Future Oncology

Audio Retrieval Based on Manifold Ranking and Relevance Feedback
Jing Qin et al., Tsinghua Science and Technology
<table>
<thead>
<tr>
<th>Link</th>
<th>Page Name</th>
<th>Company Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>About Us</td>
<td></td>
<td>Emerald Publishing</td>
</tr>
<tr>
<td>Company Information</td>
<td></td>
<td>Emerald Group</td>
</tr>
<tr>
<td>Working for Emerald</td>
<td></td>
<td>50th Anniversary</td>
</tr>
<tr>
<td>Contact Us</td>
<td></td>
<td>Emerald Bookstore</td>
</tr>
<tr>
<td>How to Find Us</td>
<td></td>
<td>Emerald Careers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The Emerald Foundation</td>
</tr>
<tr>
<td>Cookie Policy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Privacy Policy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Copyright Policy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industry Standards</td>
<td></td>
<td></td>
</tr>
<tr>
<td>End User Terms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digital Preservation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accessibility</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text and Data Mining Licence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modern Slavery Act transparency statement</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>