International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

A REGRESSION MODEL FOR PREDICTING IMAGE SEARCH ENGINE BEHAVIOR FOR BIG DATA FILTERING

Clement H.C. Leung, United International College; Yuanxi Li, Hong Kong Baptist University

Abstract

Image search engines tend to return a large number of images which the engines consider to be relevant, and such pool of results generally is very large and may be regarded to be effectively inexhaustible. While the images are presented as relevant, it is normally true that many of them are actually irrelevant, and that the distribution of relevant images over the returned results is non-uniform. To predict the relevance for individual images is generally difficult since it only takes on binary values and therefore tends to oscillate randomly between relevance and irrelevance with little noticeable trend. Increasing the range of possible values will be necessary to enhance the ability for prediction, and it is advantageous to accumulate the aggregate relevance for larger groups of images in a sequential manner. Our approach will involve appropriately grouping the random binary sequence into non-overlapping groups and convert it into a form which makes them more amenable for prediction. In this paper, we present a regression model for predicting Image Search Engines (ISEs) behavior. We develop an empirical model and design a set of benchmark queries to measure system performance. We are able to establish a linear model which is able to give good and robust prediction of search performance. In addition, the results of this research can have a direct bearing on search engine design to provide informative guidance to users on the retrieval of relevant images, and allows the users to optimize their strategy in the recovery and discovery of images.

1. Introduction and Related Work

The pervasive deployment of digital cameras and smart phones has led to the emergence of large collections of digital images, and many raw images are constantly uploaded on to the Internet often with few meaningful text label or words. Such a situation causes confusion to search engines and users alike with many images presented as relevant, but actually are not so. Therefore, the task of image retrieval [1], [17], [14] has come into question. Retrieval effectiveness [22], [24], [26], [18] becomes one of the most important parameters to measure the performance of web image retrieval systems [23], [28], [5], [8], [29], [11], [12], [7]. As is widely accepted, the most commonly used performance measures are precision and recall, which is equivalent to positive predictive value, and the sensitivity or true positive rate respectively in ROC analysis [11], [27], [30], [25], [2], [13], [4], [21], [10], [16], [9], and to compute the sensitivity can be rather difficult as the total number of relevant images is not directly observable in such a potentially infinite repository

Many researchers have conducted studies to evaluate the retrieval effectiveness of web search engines. Ece Çakır et al. [5] describe the retrieval effectiveness of image search engines based on various query topics, and different image search engines are good at different topics. Fuat Uluç et al. [30] describe the impact of the number of query words on image search engines, and suggest that the performances of image search engines will become worse when the number of query words increases. However, none of these studies describe how to estimate the total number of relevant images for the image search engines. All of them only view the first two pages of returned results. In the study by Sprink and Jansen [2], data collected from Dogpile was analyzed and one of the findings was that the percentages of users that viewed only the first page and those that viewed only the first two pages of document search results were about 71% and 15.8% respectively. Although many studies use recall as the measure to evaluate the image search engines, not many papers work on the estimation of the number of relevant images in an inexhaustible pool. An algorithm called sample-resample is presented in by Si and Callan [25]; in environments containing resource descriptions already created by query-based sampling, the sample-resample method uses several additional queries to provide an estimate of the database size.

Here, it is obvious that the more effective the system is, the more it will offer satisfaction to the users. Since the use of image search engines such as Google, Yahoo, and MSN is becoming increasingly widespread, the need for a performance evaluation of web image search engines will be of great benefit to users. We present an empirical model for predicting image search engines behavior, and we are able to establish a linear model which is able to provide good and robust prediction of search performance.

In the next section, queries design, the regression model and evaluation measurements will be introduced. The experimental results, the validation of the models and comparison of the performance of major image search engines will be discussed in section 3. Finally, we summarize our findings, and present some directions of future work in the last section.

2. Methodology and Queries Design

Image search engines tend to return a large number of images which the engines consider to be relevant, and such pool of results generally is very large and may be regarded to be effectively inexhaustible. While the images are presented as relevant, it is normally true that many of them are actually irrelevant, and that the distribution of relevant images over the returned results is non-uniform. Here, we define an image relevance indicator random variable I_k to signify the relevance of an image,

$$\text{where } I_k = \begin{cases} 0, \text{if the } k^{th} \text{ image is irrelevant} \\ 1, \text{if the } k^{th} \text{ image is relevant} \end{cases}$$

Fig. 1 shows a particular realization of the relevance indicator random variable.

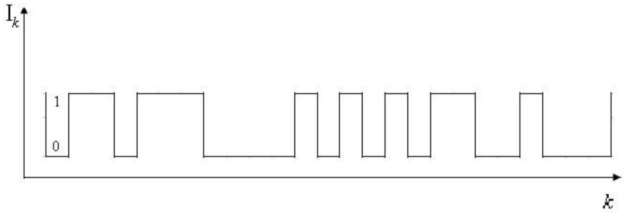


Fig. 1. Realization of the indicator random variable

To predict the relevance for individual images is generally difficult since it only takes on binary values and therefore tends to oscillate randomly between relevance and irrelevance with little noticeable trend. Increasing the range of possible values will be necessary to enhance the ability for prediction, and it is advantageous to accumulate the aggregate relevance for larger groups of images in a sequential manner. Our approach will involve appropriately grouping the random binary sequence into nonoverlapping groups and convert it into a form which makes them more amenable for prediction. We form the image results into cells as:

$$Z_k(N) = \sum_{j=(k-1)N+1}^{kN} I_j,\ k = 1, 2, \ldots. \tag{1}$$

This will produce an induced series $Z_k(N)$ based on cumulative cell relevance. We expect that, for competent search engines,

$$Z_1(N) \geq Z_2(N) \geq \ldots \geq Z_k(N) \geq \ldots \tag{2}$$

In addition, the manner of partitioning into cells will influence performance behavior and we also expect that $Z_k(N)$, in general, will be an increasing function of N. We define the yield for the k^{th} cell to be

$$\eta_k = \frac{Z_k(N)}{N}, \tag{3}$$

with $\eta_k \leq 1$, and it signifies the fractional relevance for N returned results. For competent image search engines, we expect $\eta_k \sim\leq 1$ for small k, and that for some K, we have decreasing yield $\eta_k \geq \eta_{k+1} \geq \eta_{k+2} \geq \ldots \geq \ldots$ whenever $k > K$. Fig.2 shows a linear representation of the yield.

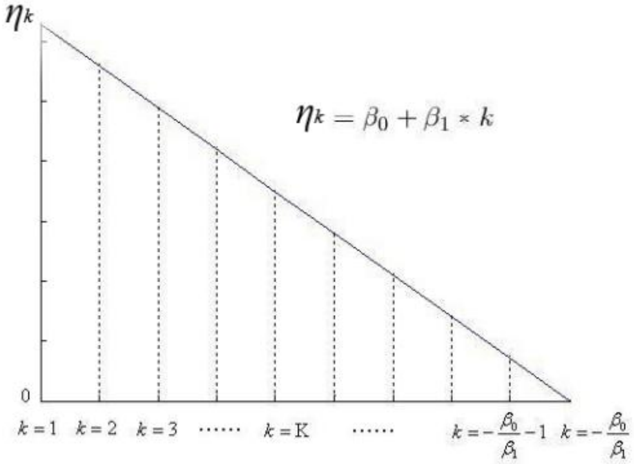


Fig. 2. Linear representation of yield

2.1. Query Design and Benchmarking

Since queries are critical to our experimentation, instead of subjectively and arbitrarily creating queries for experimental measurements, we use a systematic procedure to ensure scientific rigor and design a set of objective benchmark queries [22], [15], [18], [20] to gauge the performance of different image search engines. We make use of the Visual Dictionary [31], from which a total of 150 benchmark queries are randomly extracted, with relevance judgement based on the sample images provided in the Visual Dictionary. These cover the following topics: (i) Plants and Gardening, (ii) Astronomy, (iii) Earth, (iv) Animal Kingdom, (v) House, (vi) Transport and Machinery, (vii) Clothing and Articles, (viii) Communications, (ix) Science, (x) Human Being, (xi) Sports an Games, (xii) Society, (xiii) Arts and Architecture, (xiv) Energy, and (xv) Food and Kitchen. Ten queries are extracted randomly for each topic, and the set of queries is applied uniformly across all image search engines to calibrate and compare their performance. The full list of queries is shown in Table 1. The image search components of Google (www.google.com), Yahoo (www.yahoo.com) and MSN (www.bing.com) are selected as the search engines for experimental evaluation. These are chosen because from [6], the sum total market share of Google, Yahoo and MSN is 90.2%, with respective shares of 64%, 16.3%, and 9.9%. The measured empirical values $Z_k(N)$ are averaged over all the benchmark queries. It is expected that the characteristic parameters will be different for different image search engines, and we make use of regression models in time series analysis to evaluate the behavioral pattern of search performance. We analyze $Z_k(N)$ which is taken to be the dependent variable Y, with respect to the index k as the independent variable X, from which we determine the linear form $Y = f(X)$. Different values of N = 20, 30, 40 are measured, with N fixed for each regression curve.

2.2. Regression Model and Validation Measures

Regression model is a powerful tool to see if there is a relationship between the variables and for predicting the value of one variable based on another variable. A linear regression model assumes that the relationship between the dependent variable Y and the independent variable X is approximately linear. In our experiments, we let Y denote the ratio of the cumulative number of relevant images out of the total number of returned images observed in the respective returned pages, and let X denote the cell index. The purpose of our experiment is to investigate whether the number of relevant images in different results page follows the regression model, and how the performance behavior is influenced by the different manner of cell partitioning. Therefore, the relationship between the number of relevant images in a returned page and the cell index is taken to be linear, and we may model the corresponding regression model as:

$$Y = \beta_0 + \beta_1 X, \tag{4}$$

where Y = the ratio of the cumulative number of relevant images out of all returned images in the observed pages;

X=page number;

β_0=intercept (the value of Y when X=0);

β_1=slope of regression line.

Based on the sample data, the values of intercept and slope can be calculated as:

$$\beta_1 = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}, \tag{5}$$

$$\beta_0 = \bar{Y} - \beta_1\bar{X}. \tag{6}$$

After determining the parameters of the regression model, we would use such a model to estimate the number of relevant images page by page. A useful measure is the correlation coefficient r, which gives an indication of how well a regression model fits a particular set of data:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y}, \tag{7}$$

where $x'_i s$ and $y'_i s$ are the sample values; $\bar{x}$ and $\bar{y}$ are the sample means of X and Y; s_x and s_y are the sample standard deviations of X and Y, which indicates the strength of the linear relationship between the variables. We shall also gauge performance by examining the residuals

$$e_i = y_i - \widehat{y_i}, \tag{8}$$

Where $y'_i s$ and $\hat{y}'_i s$ are respectively the observed and estimated y values.

Testing the model for significance enables one to determine if the values are meaningful. We do this by performing a statistical hypothesis testing [19], [3] with a significance level α=0.05.

2.3. Performance Measures

In image retrieval, the most commonly used performance measures are precision and recall. Precision gives the ratio of the number of relevant images retrieved to the total number of irrelevant and relevant images retrieved.
number of relevant image retrieved

$$precision = \frac{number\ of\ relevant\ image\ retrieved}{total\ number\ of\ image\ returned} \tag{9}$$

Here, precision ratios will be calculated at various cut-off points [8] (e.g., for the first 2, 4, 6, 8, 10 and 12 cell index returned) for each image search engine. Hence, the precision at different cut-off points can be used to roughly see how the scores of relevant images are distributed over their ranks.

The cumulative precision ratio may be computed as follows:

$$precision = \frac{\widehat{Z_1(N)} + \widehat{Z_2(N)} + \ldots + \widehat{Z_K(N)}}{K * N}, \tag{10}$$

where K is the cell index;

N is the cumulative cell size; and

$\widehat{Z_\iota(N)}$ is the observed number of relevant images in the cell.

Recall gives the ratio of the number of relevant images retrieved to the total number of relevant images in the database.

$$recall = \frac{number\ of\ relevant\ image\ retrieved}{total\ number\ of\ relevant\ image\ in\ DB} \tag{11}$$

Measuring the recall presents a challenge because the denominator equation (11) is hard to determine.

From Fig. 2, with the cell index increasing, the number of relevant images of a returned page will decrease and finally drops to zero, and the cell index k when the number of relevant images drops to zero can be calculated as

$$k = \lceil -\frac{\beta_0}{\beta_1} \rceil, \tag{12}$$

when $Z_k(N) = 0$. In such case, we could regard the database of the search engine as a finite database and thus we will be able to determine the total number of relevant images for each image search engine. We let R_n be the relevant images relating to Image Search Engine ISE n, with $|R_n|$ denoting the total number of relevant images in the ISE n:

$$\begin{aligned} |R_n| &= \widehat{Z_1(N)} + \widehat{Z_2(N)} + \ldots + \widehat{Z_{\lceil -\frac{\beta_0}{\beta_1} \rceil}(N)} \\ &= \sum_{k=1}^{\lceil -\frac{\beta_0}{\beta_1} \rceil} \widehat{Z_k(N)}, \end{aligned} \tag{13}$$

Most image search engines like Google and Yahoo return 20 images in a page by default, so we include $N = 20$ in our experiments. But some image search engines like MSN display the returned images in the form of scroll down, therefore, we also take N to be 30 and 40 and try to find how various N will influence the performance behavior.

We also let $|r_n|$ denote the number of relevant images returned by ISE n at a certain cell index K, which can be estimated as

$$\begin{aligned} |r_n| &= \widehat{Z_1(N)} + \widehat{Z_2(N)} + \ldots \\ &= \sum_{k=1}^{\infty} \widehat{Z_k(N)}, \end{aligned} \tag{14}$$

3. Experimental Results and Regression Equations

In our experiments, a total of 150 benchmark queries are submitted to the selected image search engines individually, and the retrieval outputs of the search engines are recorded. The images are evaluated in binary relevance judgement. Based on Spink and Jansen's study [2], evaluating the images in the first two pages is normally enough and such a finding seems useful for the users who only want to find less than forty images, since most of the major image search engines generally display about 20 images in a result page by default. However, it is unable to satisfy the needs of the users who want to search for more and more relevant images. Therefore, based on the sample data we have recorded, the model should not be so limited. In more general studies, it is possible to have varying degrees of relevance for a given image, as measured by a number in the interval [0, 1]. In this study, we adopt the binary scale 0, 1 for measuring the relevance of an image.

3. Regression Model for Image Search Engines

In our experiments, we average the records of all the 150 benchmark queries when N is equal to 20, 30 and 40 respectively for each image search engine. The sample data are scattered and suitable models are plotted to fit the sample data for the image search engines individually. In the following, we will present the experimental measurements and compare the models for the major image search engines.

3.1.1. Google. Figure 3 gives the experimental results for this search engine with N = 20

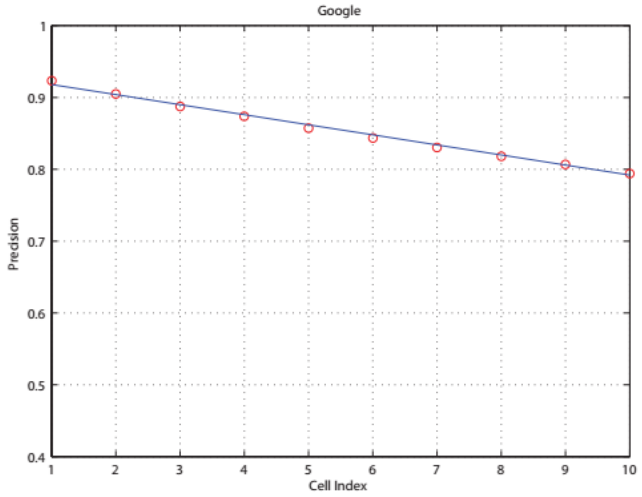


Fig. 3. Least square curve for Google when cumulative cell N=20.

The sample data are plotted in the scatter diagram, which helps to determine if a linear relationship is present. From Fig. 3, we see that the linear regression fits the data very well. We also find that the standard deviation s equals to 0.043 and the correlation coefficient r equals to -0.997, which indicates an excellent fit of the linear regression model to the measured points. Moreover, we use hypothesis testing to determine whether the linear regression model is meaningful. The model is tested using a significance level of $\alpha = 0.05$, and the result suggests that the hypothesis that $\beta_1 = 0$ should be rejected, indicating there is a statistically significant relationship at the 0.05 level. This linear regression model for the Google image search engine when the cumulative cell size 20 is given as:

$$\eta_k(20) = -0.014 * k + 0.932. \tag{15}$$

The linear regression model indicates that the precision of relevant images starts at 93.2% and decreases gradually at a rate of 0.014 per cell.

Figure 4 gives the experimental result for Google with N = 30. In Fig. 4, the linear regression model is also plotted alongside the sample data for N = 30. Here the standard deviation s equals to 0.045 and the correlation coefficient r is equal to -0.999, which implies that there is a strong relationship between the dependent variable and independent variable. Moreover, similar hypothesis testing also indicates that there is definitely a linear relationship between the number of relevant images and cell index. We find that the number of relevant images for Google starts with the precision of 92.8% and then declines steadily at a rate of 0.021 per cell. The corresponding relationship is

$$\eta_k(30) = -0.021 * k + 0.928. \tag{16}$$

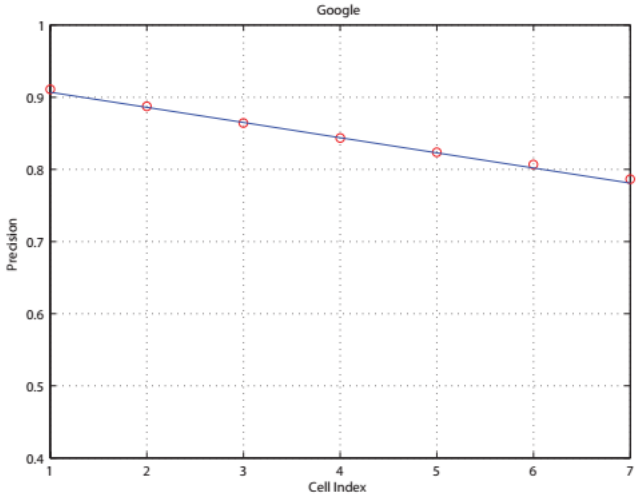


Fig. 4. Least square curve for Google when cumulative cell N=30.

Figure 5 gives the experimental results for Google with N = 40. In the scatter plot of sample data for N = 40, the pattern (Fig. 5) fits the sample data very well with s = 0.044 and r = −0.998. On the basis of hypothesis testing, we also conclude that the relationship of the number of relevant images and cell index follows a linear regression. We find that the number of relevant images begins 38, and then decreases at a rate of 2.1 for every 40 images.

$$\eta_k(40) = -2.1 * k + 38. \tag{17}$$

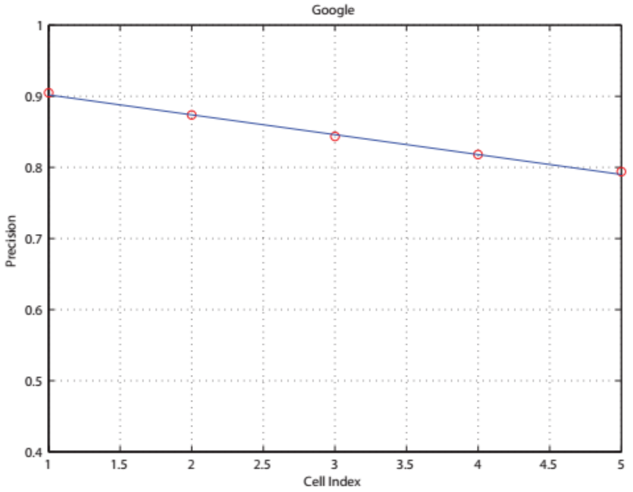


Fig. 5. Least square curve for Google when cumulative cell N=40.

3.1.2. Yahoo. Figure 6 gives the experimental results for Yahoo with N = 20. In Fig. 6, we plot the linear regression model alongside the scattered sample points. It indicates that the linear relationship between number of relevant images and cell index is well defined with s = 0.047 and r = −0.967. Moreover, based on hypothesis testing with significant level α = 0.05, the null hypothesis β_1 = 0 should be rejected. It means there is a relationship between the number of relevant images and cell index. Formula (18) indicates that the number of relevant images for Yahoo starts with 82.3% relevant images and then declines steadily at a rate of 0.015 with page steps.

$$\eta_k(20) = -0.015 * k + 0.823. \tag{18}$$

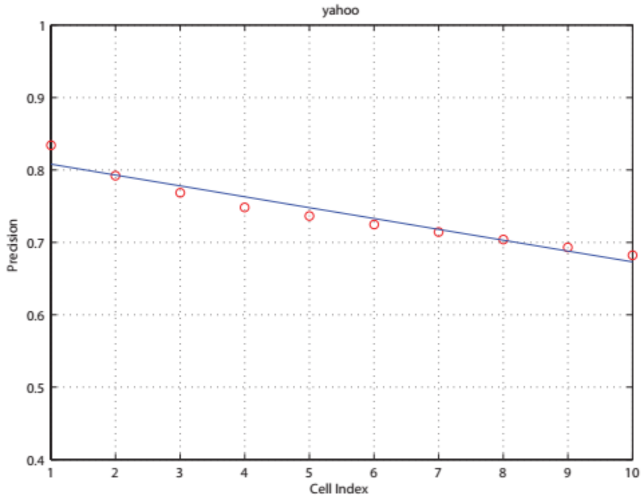


Fig. 6. Least square curve for Yahoo when cumulative cell N=20.

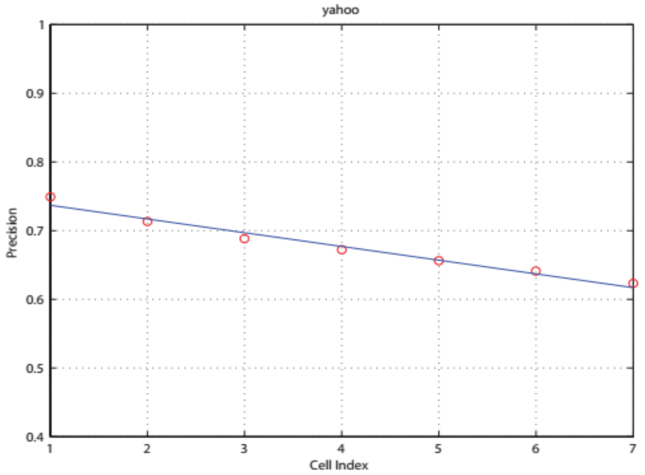


Fig. 7. Least square curve for Yahoo when cumulative cell *N*=30.

Figure 7 gives the experimental results for Yahoo with *N* = 30. From Fig. 7, the linear regression fits the data very well. We also find that the standard deviation s is 0.043 and the correlation coefficient r equals to -0.985, which all suggest that the linear regression model has a strong predictive strength. Meanwhile, we use hypothesis testing to test whether the regression model is meaningful. The model is tested using a significance level of α = 0.05, and we find that the hypothesis that $\beta_1 = 0$ should be rejected. The linear regression model for Yahoo image search engine when the cumulative cell size 30 is given as:

$$\eta_k(30) = -0.020 * k + 0.757. \qquad (19)$$

Figure 8 gives the experimental results for Yahoo with *N* = 40. In Fig. 8, the pattern shows that the linear relationship between the number of relevant images and cell index is also very good with *s* = 0.042 and correlation coefficient *r* = −0.986. For Yahoo, the linear regression model when *N* is taken as 40 is:

$$\eta_k(40) = -0.026 * k + 0.810. \qquad (20)$$

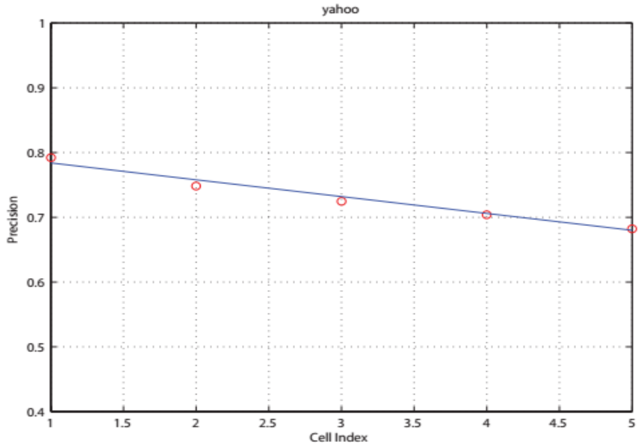


Fig. 8. Least square curve for Yahoo when cumulative cell *N*=40.

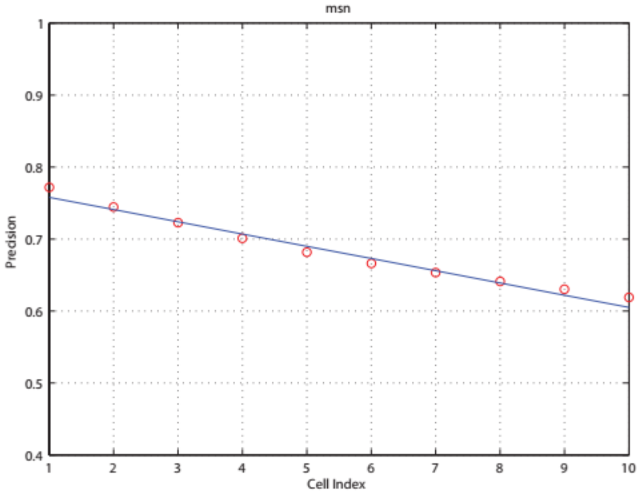


Fig. 9. Least square curve for MSN when cumulative cell *N*=20.

3.1.3. MSN. Figure 9 gives the experimental results for MSN with *N* = 20. In Fig. 9, the linear regression model is plotted alongside the sample data. Mathematically, the linear regression model fits the sample data very well in terms of the value of the standard deviation s = 0.051 and the correlation coefficient *r* = −0.987. The hypothesis testing also indicates that there is a statistically significant relationship at α = 0.05 level. The linear regression model for MSN is:

$$\eta_k(20) = -0.017 * k + 0.775. \qquad (21)$$

Formula (22) indicate that the number of relevant images decreases by one image for every two cell index steps.

Figure 10 gives the experimental results for MSN with *N* = 30, where a linear regression model is obvious, with the standard deviation s equals to 0.051 and the correlation coefficient r equals to -0.991. It comes down by 1.2 image with every cell index increase, and the relationship is:

$$\eta_k(30) = -0.023 * k + 0.768. \qquad (22)$$

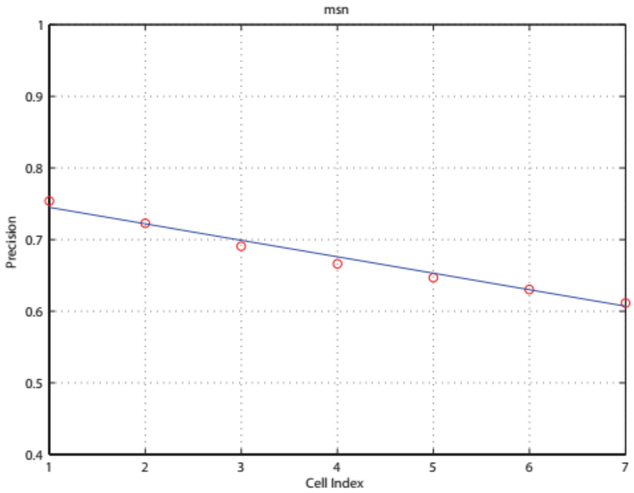


Fig. 10. Least square curve for MSN when cumulative cell *N*=30.

ISSN:2319-7900

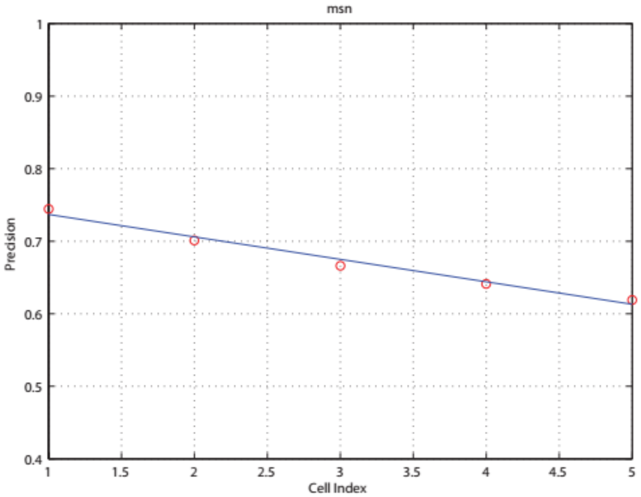


Fig. 11. Least square curve for MSN when cumulative cell *N*=40.

Figure 11 gives the experimental results for MSN with *N* = 40. Notice that *s* equals to 0.050 and *r* equals to -0.990, suggesting that the number of relevant images has a strong linear relationship to cell index. On the basis of hypothesis testing, this model is significant at the α = 0.05 level. Therefore, MSN starts the precision of number of relevant images from 76.8%, which is the smallest among all the image search engines, and it decreases at a rate of 0.031 with every cell index increase. It drops quicker than Yahoo but at the same rate as Google. The relationship is:

$$\eta_k(40) = -0.031 * k + 0.768. \tag{23}$$

Based on Formulas (16), (17), (18), (19), (20), (21), (22), (23) and (24), we arrive at an unified linear regression model which may be able to be used to investigate the distribution of the number of relevant images for general image search engines in terms of different *N*. The generic regression equations are:

$$Z_k(20) = -0.50 * k + 17. \tag{24}$$

$$Z_k(30) = -1.1 * k + 25. \tag{25}$$

$$Z_k(40) = -1.98 * k + 34. \tag{26}$$

In accordance with hypothesis testing, if these linear regression models are meaningful and significant at the α = 0.05 level. However, how well the unified linear regression model can predict the number of relevant images for all the image search engines will be analyzed using in the next section.

3.2. Accuracy Test for Regression Model

To compare across engines and to compare the effect of different cell sizes, a total of 15 benchmark queries, one from each of the 15 topics is randomly selected from Table 1, and they are used to test the forecasting accuracy of the model. The Mean Absolute Error (*MAE*) is used to measure the accuracy of a model and it is a common measure of forecast error in time series analysis. *MAE* is a quantity used to measure how close forecasts or predictions are to the eventual outcomes:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|, \tag{27}$$

where f_i is the predicted value, and y_i is the true value.
Since the cell size would have a bearing on the error range, some normalization is necessary for comparison purpose. In particular for *N* = 20, the maximum error range is [0, 20], and for *N* = 30, the error range is [0, 30]. In general, the error range is [0, *N*] for a cell size of *N*. Thus, for meaningful comparison, we normalize the error range to [0, 1] so that we replace e_i by $\frac{e_i}{N}$. Effectively, this is using:

$$MAE_{norm} = \frac{MAE}{N}, \tag{28}$$

where *N* is the cumulative cell size. As we know, the smaller the *MAE* the better the model is.

Table 2 shows the accuracy of the prediction results which exhibit good performance in predicting the image search engine behavior. These experimental results indicate that the linear regression model could estimate the number of relevant images for Google better than Yahoo and MSN, no matter using individual linear regression model or the unified linear regression model; meanwhile, the performances of Yahoo and MSN are similar. The values of MAE_{norm} are 0.0032, 0.0021 and 0.0016 for Google for *N* = 20, *N* = 30 and *N* = 40 respectively. The corresponding errors for Yahoo are 0.0065, 0.0051 and 0.0033, and those for MSN are 0.0083, 0.0055 and 0.0042. From these observations, we see that *N* = 40 yields the lowest error. Using *N* = 40, we compute MAE_{norm} for the three search engines and the results are shown in the last three columns of Table 2. According to Figure 11, we see that using the unified linear regression model, the values of MAE_{norm} for MSN has improved from 0.042 to 0.038, while the value of MAE_{norm} for Yahoo improves slightly from 0.033 to 0.031. However, the value of MAE_{norm} for Google sees a slight increase. Thus, we can conclude that individual linear regression model will be better for predicting the performance of each image search engine, and such a model is able to give good and robust prediction of image search performance. Nevertheless, it is still convenient to use a unified linear regression model to predict all the image search engines performance since the values of MAE_{norm} are quite acceptable.

3.3. Comparison of the Predictive Behavior for Images Search Engines

Figure 12 compares different image search engine performance $N = 40$. For instance, the precision of Google is computed as follows, when the cut-off point is 6 and the cumulative cell size is 40,

$$\begin{aligned} precision &= \frac{\widehat{Z_1(N)} + \widehat{Z_2(N)} + \ldots + \widehat{Z_6(N)}}{6 * 40} \\ &= 77\% \end{aligned}$$

Google retrieves the greatest number of relevant images at all cut-off points; its best precision rate is 87% at cut-off point 2 and decrease gradually to 61% at cut-off point 12. Yahoo comes second. It retrieves more relevant images than MSN at all cut-off points and it declines at a rate of 1.8, which is smaller than Google and MSN. The precision ratio of MSN is the lowest at all cut-off points among all tested ISEs; however, the precision is still good enough with 70% at cut-off point 2 and over 43% at cut-off point 12.

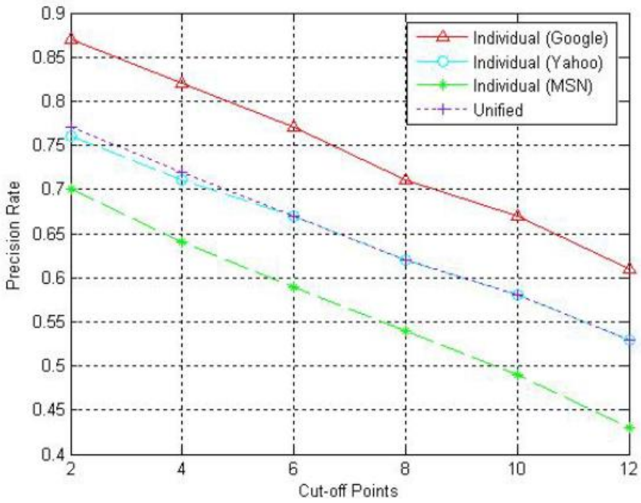


Fig. 12. Precision for ISEs at various cut-off points

3.4. Recall Performance

It is also useful to examine the recall behavior. The performance of different image search engines is compared in terms of recall for 400 returned images. For example, for N = 40, the recall for Google is computed below,

$$\begin{aligned} k &= \lceil -\frac{\beta_0}{\beta_1} \rceil = 18 \\ |r_{Google}| &= \widehat{Z_1(40)} + \widehat{Z_2(40)} + \ldots + \widehat{Z_{10}(40)} \\ &= 265, \\ |R_{Google}| &= \widehat{Z_1(40)} + \widehat{Z_2(40)} + \ldots + \widehat{Z_{\lceil -\frac{\beta_0}{\beta_1} \rceil}(40)} \\ &= \widehat{Z_1(40)} + \widehat{Z_2(40)} + \ldots + \widehat{Z_{18}(40)} \\ &= 325, \\ recall_{Google} &= \frac{|r_{Google}|}{|R_{Google}|} \\ &= \frac{265}{325} \\ &= 81.4\%, \end{aligned}$$

where $\widehat{Z_{\lceil -\frac{\beta_0}{\beta_1} \rceil}(N)}$ is the first page that has zero relevant images.

For $N = 40$, the recall for Yahoo is computed below,

$$\begin{aligned} k &= \lceil -\frac{\beta_0}{\beta_1} \rceil = 18 \\ |r_{Yahoo}| &= \widehat{Z_1(40)} + \widehat{Z_2(40)} + \ldots + \widehat{Z_{10}(40)} \\ &= 231, \\ |R_{Yahoo}| &= \widehat{Z_1(40)} + \widehat{Z_2(40)} + \ldots + \widehat{Z_{\lceil -\frac{\beta_0}{\beta_1} \rceil}(40)} \\ &= \widehat{Z_1(40)} + \widehat{Z_2(40)} + \ldots + \widehat{Z_{18}(40)} \\ &= 286, \\ recall_{Yahoo} &= \frac{|r_{Yahoo}|}{|R_{Yahoo}|} \\ &= \frac{231}{386} \\ &= 80.7\%, \end{aligned}$$

For N = 40, the recall for MSN is computed below,

$$\begin{aligned} k &= \lceil -\frac{\beta_0}{\beta_1} \rceil = 14 \\ |r_{MSN}| &= \widehat{Z_1(40)} + \widehat{Z_2(40)} + \ldots + \widehat{Z_{10}(40)} \\ &= 195, \\ |R_{MSN}| &= \widehat{Z_1(40)} + \widehat{Z_2(40)} + \ldots + \widehat{Z_{\lceil -\frac{\beta_0}{\beta_1} \rceil}(40)} \\ &= \widehat{Z_1(40)} + \widehat{Z_2(40)} + \ldots + \widehat{Z_{14}(40)} \\ &= 214, \\ recall_{MSN} &= \frac{|r_{MSN}|}{|R_{MSN}|} \\ &= \frac{195}{214} \\ &= 91.1\%, \end{aligned}$$

Using the unified model, we have:

$$\begin{aligned} k &= \lceil -\frac{\beta_0}{\beta_1} \rceil = 17 \\ |r_{Unified}| &= \overline{Z_1(40)} + \overline{Z_2(40)} + \ldots + \overline{Z_{10}(40)} \\ &= 230, \\ |R_{Unified}| &= \overline{Z_1(40)} + \overline{Z_2(40)} + \ldots + \overline{Z_{\lceil -\frac{\beta_0}{\beta_1} \rceil}(40)} \\ &= \overline{Z_1(40)} + \overline{Z_2(40)} + \ldots + \overline{Z_{17}(40)} \\ &= 273, \\ recall_{Unified} &= \frac{|r_{Unified}|}{|R_{Unified}|} \\ &= \frac{230}{273} \\ &= 84.2\%, \end{aligned}$$

where $\overline{Z_i(N)}$ signifies the average value of $\widehat{Z_\iota(N)}$ for the three search engines. From these, we see that MSN provides significantly better overall recall compared with the other

ISSN:2319-7900

two engines, while the overall recall rates of Google and Yahoo exhibit a slight difference of less than one percent. They are plotted in Figure 13.

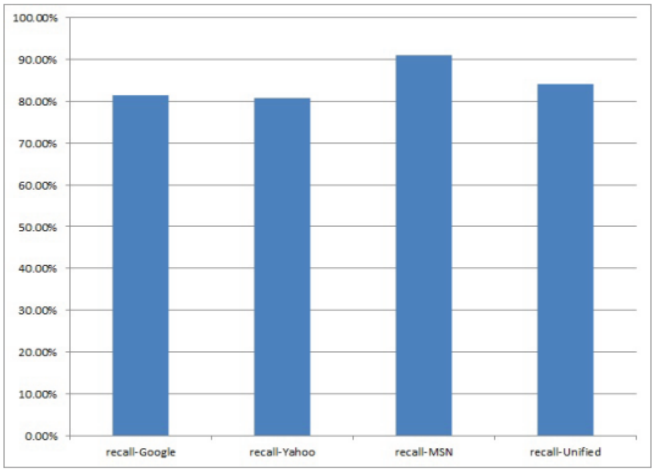


Fig. 13. Recall for ISEs with 400 returned images

It is interesting to examine ρ_k, the marginal recall at k, and observe how it changes with different cell index,

$$\rho_k(N) = \frac{\widehat{Z_k(N)}}{|R|}$$

where $|R|$ is the total number of relevant images estimated by the regression method. These are plotted in Figure 14. It is interesting to see that the marginal recall also follows a linear curve, describable by the following equations respectively for the three search engines,

$$Z_k(recall - Google) = -0.0060 * k + 0.108. \qquad \textbf{(30)}$$

$$Z_k(recall - Yahoo) = -0.0057 * k + 0.105. \qquad \textbf{(31)}$$

$$Z_k(recall - MSN) = -0.0095 * k + 0.14. \qquad \textbf{(32)}$$

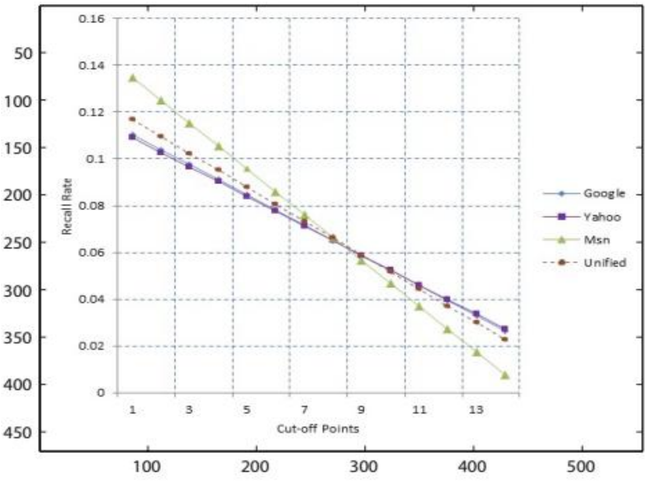


Fig. 14. Marginal Recall for ISEs with 400 returned images

As indicated earlier, most searches tend to only examine the first few pages, and we see that certainly for the first few pages of the results, MSN outperforms Yahoo, which in turn outperforms Google. It is noteworthy that the opposite ordering of the three search engines is true in the case of precision (Fig. 12). This seems to confirm the principle that simultaneously optimizing both the recall and precision is not possible. The unified recall equation is

$$Z_k(recall - Unified) = -0.0071 * k + 0.12. \qquad (33)$$

From these results, it is possible to conclude that for image discovery, where one is not specific about retrieving particular images and where precision is emphasized, Google provides the best performance. On the other hand, for image recovery, where one is specific about retrieving particular images and where recall is being emphasized, MSN provides the best performance.

4. Conclusion

Most image search engines tend to return an inordinate number of images which the engines present as relevant, but actually, quite a proportion of such images are not relevant. To predict the distribution of relevant images over the returned results for image search engines, we develop a regression model to allow their behavioral properties to be represented. We find that the regression model is able to furnish relatively accurate and useful prediction of search behavior, and regression equations are provided for describing both the precision and recall behavior. Consequently, the results of this research will have a direct bearing on future search engine design as well as providing informative guidance to users on the retrieval of relevant images, allowing them to optimize their strategy in the recovery and discovery of images.

References

[1] Yihun Alemu, Jong-Bin Koh, M. Ikram, and Dong-Kyoo Kim. 2009. Image Retrieval in Multimedia Databases: A Survey. In *Intelligent Information Hiding and Multimedia Signal Processing, 2009. IIH-MSP '09. Fifth International Conference on*. 681–689. https://doi.org/10.1109/IIH-MSP.2009.159

[2] Bernard J. Jansen Amanda Spink. 1977. *Searching multimedia federated content web collections*. Online Information Review, Vol. 30. Emerald Group Publishing Limited. https://doi.org/10.1108/14684520610706389

[3] Michael E. Hanna Barry Render, Ralph M. Stair Jr. 2006. Quantitative Analysis for Management-Chapter 4. (2006). http://books.google.com.hk/books?id=5I84AAAACAAJ

[4] Jr. Black, J.A., G. Fahmy, and S. Panchanathan. 2002. A method for evaluating the performance of content-based image retrieval systems. In *Image Analysis and Interpretation, 2002. Proceedings. Fifth IEEE Southwest Symposium on*. 96–100. https://doi.org/10.1109/IAI.2002.999897

[5] E. Cakir, H. Bahceci, and Y. Bitirim. 2008. An Evaluation of Major Image Search Engines on Various Query Topics. In *Internet Monitoring and Protection, 2008. ICIMP '08. The Third International Conference on*. 161–165. https://doi.org/10.1109/ICIMP.2008.9

[6] Stoney G deGeyter. 2009. April 2009 Search Engine Market Share | Nielsen Online. (2009). http://www.polepositionmarketing.com/emp/april-2009-search-engine-2/

[7] R.G. Demirci, V. Kismir, and Y. Bitirim. 2007. An Evaluation of Popular Search Engines on Finding Turkish Documents. In *Internet and Web Applications and Services, 2007. ICIW '07. Second International Conference on*. 61–61. https://doi.org/10.1109/ICIW.2007.15

[8] M.T. Elagoz, M. Mendeli, R.Z. Manioglulari, and Y. Bitirim. 2008. An Empirical Evaluation on Meta-Image Search Engines. In *Digital Telecommunications, 2008. ICDT '08. The Third International Conference on*. 135–139. https://doi.org/10.1109/ICDT.2008.10

[9] Tom Fawcett. 2006. An Introduction to ROC Analysis. *Pattern Recogn. Lett.* 27, 8 (June 2006), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[10] Alexander G. Hauptmann and Michael G. Christel. 2004. Successful Approaches in the TREC Video Retrieval Evaluations. In *Proceedings of the 12th Annual ACM International Conference on Multimedia (MULTIMEDIA '04)*. ACM, New York, NY, USA, 668–675. https://doi.org/10.1145/1027527.1027681

[11] T. Ishioka. 2003. Evaluation of criteria for information retrieval. In *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*. 425–431. https://doi.org/10.1109/WI.2003.1241232

[12] Hua Jiang. 2009. Study on the Performance Measure of Information Retrieval Models. In *Intelligent Ubiquitous Computing and Education, 2009 International Symposium on*. 436–439. https://doi.org/10.1109/IUCE.2009.105

[13] Xiangyu Jin and James C. French. 2003. Improving Image Retrieval Effectiveness via Multiple Queries. In *Proceedings of the 1st ACM International Workshop on Multimedia Databases (MMDB '03)*. ACM, New York, NY, USA, 86–93. https://doi.org/10.1145/951676.951692

[14] M. L. Kherf, D. Ziou, and A. Bernardi. 2004. Image Retrieval from the World Wide Web: Issues, Techniques, and Systems. *ACM Comput. Surv.* 36, 1 (March 2004), 35–67. https://doi.org/10.1145/1013208.1013210

[15] Clement H. C. Leung and Horace Ho-Shing Ip. 2000. Benchmarking for Content Based Visual Information Search. In *Proceedings of the 4th International Conference on Advances in Visual Information Systems (VISUAL '00)*. Springer-Verlag, London, UK, UK, 442–456. http://dl.acm.org/citation.cfm?id=647061.714444

[16] Longzhuang Li and Yi Shang. 2000. A new statistical method for performance evaluation of search engines. In *Tools with Artificial Intelligence, 2000. ICTAI 2000. Proceedings. 12th IEEE International Conference on*. 208–215. https://doi.org/10.1109/TAI.2000.889872

[17] Wei-Hao Lin, Rong Jin, and A. Hauptmann. 2003. Web image retrieval re-ranking with relevance model. In *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*. 242–248. https://doi.org/10.1109/WI.2003.1241200

[18] S. Marchand-Maillet and Marcel Worring. 2006. Benchmarking Image and Video Retrieval: An Overview. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR '06)*. ACM, New York, NY, USA, 297–300. https://doi.org/10.1145/1178677.1178718

[19] T.W. Miller. 2005. *Data and Text Mining: A Business Applications Approach*. Pearson Prentice Hall. http://books.google.com.hk/books?id=ldyLQgAACAAJ

[20] Henning MÃijller and Antoine Geissbuhler. 2004. Benchmarking image retrieval applications. In *In Proceedings of the 7 th International Conference on Visual Information Systems*. Springer. XX.

[21] Henrik Nottelmann and Norbert Fuhr. 2003. Evaluating Different Methods of Estimating Retrieval Quality for Resource Selection. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '03)*. ACM, New York, NY, USA, 290–297. https://doi.org/10.1145/860435.860489

[22] P. Over, C. Leung, H. Ip, and M. Grubinger. 2004. Multimedia retrieval benchmarks. *MultiMedia, IEEE* 11, 2 (Apr 2004), 80–84. https://doi.org/10.1109/MMUL.2004.1289045

[23] Jaroslav Pokorny. 2004. Web Searching and Information Retrieval. *Computing in Science and Engg.* 6, 4 (July 2004), 43–48. https://doi.org/10.1109/MCSE.2004.24

[24] N.V. Shirahatti and K. Barnard. 2005. Evaluating image retrieval. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1. 955–961 vol. 1. https://doi.org/10.1109/CVPR.2005.147

[25] Luo Si and Jamie Callan. 2003. Relevant Document Distribution Estimation Method for Resource Selection. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in In-*

formation Retrieval (SIGIR '03). ACM, New York, NY, USA, 298–305. https://doi.org/10.1145/860435.860490

[26] J.R. Smith. 1998. Image retrieval evaluation. In *Content-Based Access of Image and Video Libraries, 1998. Proceedings. IEEE Workshop on*. 112–113. https://doi.org/10.1109/IVL.1998.694520

[27] Raquel Kolitski Stasiu, Carlos A. Heuser, and Roberto Silva. 2005. Estimating Recall and Precision for Vague Queries in Databases. In *Advanced Information Systems Engineering*, Oscar Pastor and JoÃčo FalcÃčo e Cunha (Eds.). Lecture Notes in Computer Science, Vol. 3520. Springer Berlin Heidelberg, 187–200. https://doi.org/10.1007/11431855_14

[28] K. Stevenson and C. Leung. 2005. Comparative evaluation of Web image search engines for multimedia applications. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. 4 pp.–. https://doi.org/10.1109/ICME.2005.1521641

[29] Duygu Tumer, Mohammad Ahmed Shah, and Yiltan Bitirim. 2009. An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia. In *Proceedings of the 2009 Fourth International Conference on Internet Monitoring and Protection (ICIMP '09)*. IEEE Computer Society, Washington, DC, USA, 51–55. https://doi.org/10.1109/ICIMP.2009.16

[30] F. Uluc, E. Emirzade, and Y. Bitirim. 2007. The Impact of Number of Query Words on Image Search Engines. In *Internet and Web Applications and Services, 2007. ICIW '07. Second International Conference on*. 50–50. https://doi.org/10.1109/ICIW.2007.61

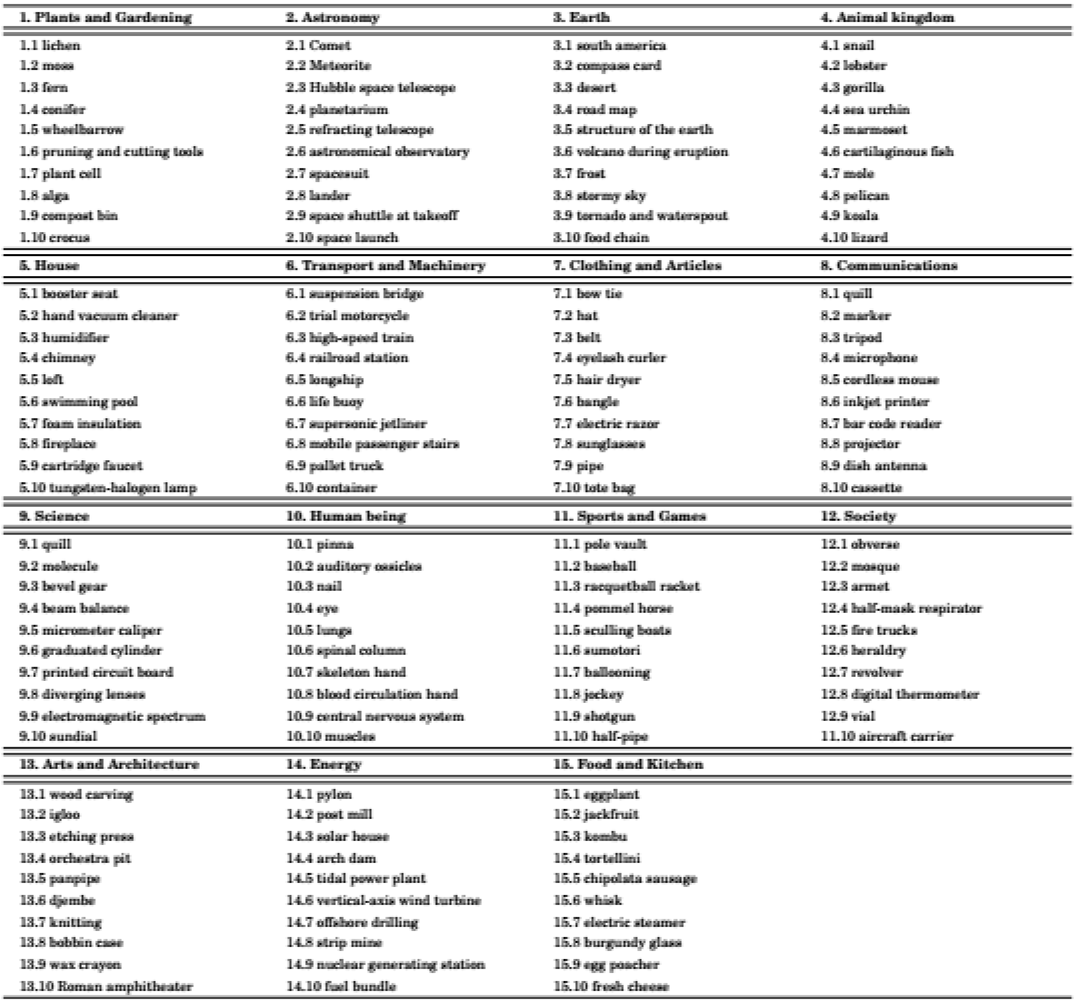
[31] Online VisualDictionary. 2013. Visual Dictionary Online. (2013). http://visual.merriam-webster.com/index.php

Biographies

Prof. Clement H.C. Leung is a Professor in Computer Science and Technology, United International College. He obtained his BSc (Hons) in Mathematics from McGill University, Canada, his MSc in Mathematics from Oxford University, and his PhD in Computer Science from University College London. His principal research interest is in the area of Multimedia and Visual Information Systems. He holds two US patents, and his publications include four books and well over one hundred research articles. Among the publications that his articles have appeared include: IEEE Transactions on Pattern Analysis and Machine Intelligence, ACM Journal of Multimedia, IEEE Transactions on Communications, ACM Transactions on Intelligent Systems Technology, ACM Computer Graphics, The Computer Journal, IEEE Transactions on Software Engineering, and IEEE Multimedia. He is acknowledged by the international research community to be a pioneer in the field of Visual Information Systems, and has published the first research volume on the subject. His services to the research community include serving as Program Chair, Program Co-Chair, Keynote Speaker, Panel Expert, and on the Program Committee and Steering Committee of major International Conferences. In addition to contributing to the Editorship of ten international journals, he has served as Chairman of the International Association for Pattern Recognition Technical Committee on Multimedia and Visual Information Systems, as well as well as on the International Standards (ISO) MPEG-7 Committee responsible for generating standards for digital multimedia, where he played an active role in shaping the influential MPEG-7 International Standard. He is listed in Who's Who in the World, Great Minds of the 21st Century, Dictionary of International Biography, Who's Who in Australia, and Who's Who in Australasia & Pacific Nations. He is a Fellow of the British Computer Society, and a Fellow of the Royal Society of Arts, Manufactures and Commerce.

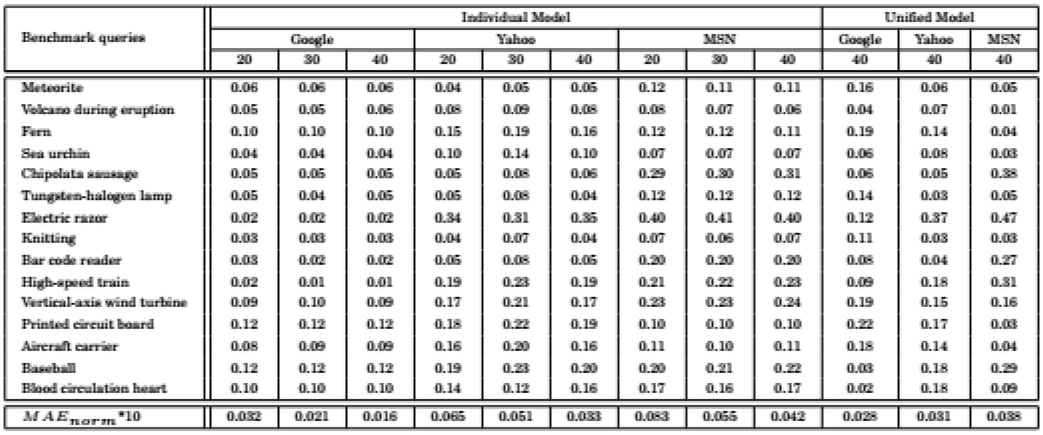
Dr. Yuanxi Li is currently a lecturer at Department of Computer Science, Hong Kong Baptist University. She received her Ph.D. degree in Computer Science and M.Sc. in IT Management from Hong Kong Baptist University. Her research interests include Semantic Multimedia Indexing and Retrieval, Adaptive Searching Engine and Semantic Ontology, etc. She has been serving as committee in several International Conferences and Workshops, including organizing committee RTCSA2015, co-chair of UCC-IDP2016, etc. Besides, she has been also reviewers for several Journal and conference, including the International Journal of Web Information Systems, the International Conference on Utility and Cloud Computing, etc.

Table. 1. Benchmark Query List



1. Plants and Gardening	2. Astronomy	3. Earth	4. Animal kingdom
1.1 lichen	2.1 Comet	3.1 south america	4.1 snail
1.2 moss	2.2 Meteorite	3.2 compass card	4.2 lobster
1.3 fern	2.3 Hubble space telescope	3.3 desert	4.3 gorilla
1.4 conifer	2.4 planetarium	3.4 road map	4.4 sea urchin
1.5 wheelbarrow	2.5 refracting telescope	3.5 structure of the earth	4.5 marmoset
1.6 pruning and cutting tools	2.6 astronomical observatory	3.6 volcano during eruption	4.6 cartilaginous fish
1.7 plant cell	2.7 spacesuit	3.7 frost	4.7 mole
1.8 alga	2.8 lander	3.8 stormy sky	4.8 pelican
1.9 compost bin	2.9 space shuttle at takeoff	3.9 tornado and waterspout	4.9 koala
1.10 crocus	2.10 space launch	3.10 food chain	4.10 lizard
5. House	**6. Transport and Machinery**	**7. Clothing and Articles**	**8. Communications**
5.1 booster seat	6.1 suspension bridge	7.1 bow tie	8.1 quill
5.2 hand vacuum cleaner	6.2 trial motorcycle	7.2 hat	8.2 marker
5.3 humidifier	6.3 high-speed train	7.3 belt	8.3 tripod
5.4 chimney	6.4 railroad station	7.4 eyelash curler	8.4 microphone
5.5 loft	6.5 longship	7.5 hair dryer	8.5 cordless mouse
5.6 swimming pool	6.6 life buoy	7.6 bangle	8.6 inkjet printer
5.7 foam insulation	6.7 supersonic jetliner	7.7 electric razor	8.7 bar code reader
5.8 fireplace	6.8 mobile passenger stairs	7.8 sunglasses	8.8 projector
5.9 cartridge faucet	6.9 pallet truck	7.9 pipe	8.9 dish antenna
5.10 tungsten-halogen lamp	6.10 container	7.10 tote bag	8.10 cassette
9. Science	**10. Human being**	**11. Sports and Games**	**12. Society**
9.1 quill	10.1 pinna	11.1 pole vault	12.1 obverse
9.2 molecule	10.2 auditory ossicles	11.2 baseball	12.2 mosque
9.3 bevel gear	10.3 nail	11.3 racquetball racket	12.3 armet
9.4 beam balance	10.4 eye	11.4 pommel horse	12.4 half-mask respirator
9.5 micrometer caliper	10.5 lungs	11.5 sculling boats	12.5 fire trucks
9.6 graduated cylinder	10.6 spinal column	11.6 sumotori	12.6 heraldry
9.7 printed circuit board	10.7 skeleton hand	11.7 ballooning	12.7 revolver
9.8 diverging lenses	10.8 blood circulation hand	11.8 jockey	12.8 digital thermometer
9.9 electromagnetic spectrum	10.9 central nervous system	11.9 shotgun	12.9 vial
9.10 sundial	10.10 muscles	11.10 half-pipe	11.10 aircraft carrier
13. Arts and Architecture	**14. Energy**	**15. Food and Kitchen**	
13.1 wood carving	14.1 pylon	15.1 eggplant	
13.2 igloo	14.2 post mill	15.2 jackfruit	
13.3 etching press	14.3 solar house	15.3 kombu	
13.4 orchestra pit	14.4 arch dam	15.4 tortellini	
13.5 panpipe	14.5 tidal power plant	15.5 chipolata sausage	
13.6 djembe	14.6 vertical-axis wind turbine	15.6 whisk	
13.7 knitting	14.7 offshore drilling	15.7 electric steamer	
13.8 bobbin case	14.8 strip mine	15.8 burgundy glass	
13.9 wax crayon	14.9 nuclear generating station	15.9 egg poacher	
13.10 Roman amphitheater	14.10 fuel bundle	15.10 fresh cheese	

Table. 2. MAE for Different Image Search Engines



Benchmark queries	Individual Model									Unified Model		
	Google			Yahoo			MSN			Google	Yahoo	MSN
	20	30	40	20	30	40	20	30	40	40	40	40
Meteorite	0.06	0.06	0.06	0.04	0.05	0.05	0.12	0.11	0.11	0.16	0.06	0.05
Volcano during eruption	0.05	0.05	0.06	0.08	0.09	0.08	0.08	0.07	0.06	0.04	0.07	0.01
Fern	0.10	0.10	0.10	0.15	0.19	0.16	0.12	0.12	0.11	0.19	0.14	0.04
Sea urchin	0.04	0.04	0.04	0.10	0.14	0.10	0.07	0.07	0.07	0.06	0.08	0.03
Chipolata sausage	0.05	0.05	0.05	0.05	0.08	0.06	0.29	0.30	0.31	0.06	0.05	0.38
Tungsten-halogen lamp	0.05	0.04	0.05	0.05	0.08	0.04	0.12	0.12	0.12	0.14	0.03	0.05
Electric razor	0.02	0.02	0.02	0.34	0.31	0.35	0.40	0.41	0.40	0.12	0.37	0.47
Knitting	0.03	0.03	0.03	0.04	0.07	0.04	0.07	0.06	0.07	0.11	0.03	0.03
Bar code reader	0.03	0.02	0.02	0.05	0.08	0.05	0.20	0.20	0.20	0.08	0.04	0.27
High-speed train	0.02	0.01	0.01	0.19	0.23	0.19	0.21	0.22	0.23	0.09	0.18	0.31
Vertical-axis wind turbine	0.09	0.10	0.09	0.17	0.21	0.17	0.23	0.23	0.24	0.19	0.15	0.16
Printed circuit board	0.12	0.12	0.12	0.18	0.22	0.19	0.10	0.10	0.10	0.22	0.17	0.03
Aircraft carrier	0.08	0.09	0.09	0.16	0.20	0.16	0.11	0.10	0.11	0.18	0.14	0.04
Baseball	0.12	0.12	0.12	0.19	0.23	0.20	0.20	0.21	0.22	0.03	0.18	0.29
Blood circulation heart	0.10	0.10	0.10	0.14	0.12	0.16	0.17	0.16	0.17	0.02	0.18	0.09
MAE_{norm}*10	0.032	0.021	0.016	0.065	0.051	0.033	0.083	0.055	0.042	0.028	0.031	0.038