Towards More Efficient Image Web Search

Mohammed Abdel Razek1,2
1Deanship of E-Learning and Distance Education, King Abdul-Aziz University, Jeddah, KSA
2Mathematics and Computer Science Department, Faculty of Science, Azhar University, Cairo, Egypt
Email: abdelram@azhar.edu.eg

Received October 30, 2013; revised November 15, 2013; accepted November 29, 2013

ABSTRACT

With the flood of information on the Web, it has become increasingly necessary for users to utilize automated tools in order to find, extract, filter, and evaluate the desired information and knowledge discovery. In this research, we will present a preliminary discussion about using the dominant meaning technique to improve Google Image Web search engine. Google search engine analyzes the text on the page adjacent to the image, the image caption and dozens of other factors to determine the image content. To improve the results, we looked for building a dominant meaning classification model. This paper investigated the influence of using this model to retrieve more efficient images, through sequential procedures to formulate a suitable query. In order to build this model, the specific dataset related to an application domain was collected; K-means algorithm was used to cluster the dataset into K-clusters, and the dominant meaning technique is used to construct a hierarchy model of these clusters. This hierarchy model is used to reformulate a new query. We perform some experiments on Google and validate the effectiveness of the proposed approach. The proposed approach is improved for in precision, recall and F1-measure by 57%, 70%, and 61% respectively.

Keywords: Web Mining; Image Retrieval; Dominant Meaning Technique; K-Means Algorithm, Web Search

1. Introduction

The continuous growth in the size and use of the Web information imposes new techniques to extract Web contents. The taxonomy of Web mining contains three categories: Web content mining, Web structure, and Web usage. The first category is Web content mining which presents the process of extracting information and knowledge from web WebPages. It may also deal with the content data of the Web pages which consist of text, images, audio, video, or structured records such as lists and tables. This research will focus only on the Web content mining which is the mining of pictures of a Web page to find out the weight of the content of the search query. The images on the web are considered as part of Web contents [1].

In a major part of this project, we will try to answer the following challenges: how to construct a query model based on the dominant meaning; how to improve the results of Web images search. To overcome, we use the following algorithm to improve the query results of image search:

- Collecting specific dataset related to some application domain;
- Using K-means algorithm to cluster the dataset into K-clusters;
- Using dominant meaning technique to construct the Hierarchy of meaning;
- Constructing a new query based on the dominant meaning algorithm;
- Using the new query to Google;
- Filter results based on the dominant meaning words.

This project uses a clustering method called K-means to classify dataset into k-clusters. Clustering is the process of partitioning or grouping a given set of patterns into disjoint clusters. This project will use one of the clustering methods called K-means. The k-means presented an effect in producing good clustering results for many practical applications [2]. However, a direct algorithm of k-means method requires time proportional to the product of a number of patterns and a number of clusters per iteration [3]. Following [4-6], this project briefly illustrates the direct K-means algorithm.

The idea behind this research is to improve the Image Web search using the dominant meaning technique [7].
We apply dominant meaning words, along with a machine learning method to classify WebPages. The dominant meaning definition is known as “the set of keywords that best fit an intended meaning of a target word” [7]. This technique sees a query as a target meaning plus some words that fall within the range of that meaning. It freezes up the target meaning, which is called a master word, and adds or removes some slave words, which clarify the target meaning.

2. Motivation

This research tackles to solve the Web mining content. For the semi-structured data, all the works utilize the HTML structures inside the WebPages and some utilized the hyperlink structure between the WebPages for WebPage representation. As for the database view, in order to have the better information management and querying on the web, the mining always tries to infer the structure of the web site to transform a web site to become a database.

For HTML web pages, there are many research and commercial systems available which use also image captions, e.g. Google image search: “Google analyzes the text on the page adjacent to the image, the image caption and dozens of other factors to determine the image content. Google also uses sophisticated algorithms to remove duplicates and ensure that the highest quality images are presented first in your results” [8], and [9]. In this sense, this project is using dominant meaning technique [7] and how it can be used to improve Web images searches. How does it influence search results?

The dominant meaning definition is known as “the set of keywords that best fit an intended meaning of a target word” [7]. This technique sees a query as a target meaning plus some words that fall within the range of that meaning. It freezes the target meaning, which is called a master word, and adds or removes some slave words, which clarify the target meaning.

For example, suppose that the query is “Java”. Figure 1 shows the results of the word “Java”. As shown, the most results are representing some images for the three well-known meanings of java: Java (computer program language), Java (coffee), and Java (Island).

The idea of this research is to clarify the target meaning with some slave words. Accordingly, if we need to look for java (computer program language), we need to add some slaves of java such as, computer, program, and language.

Figure 2 shows the results of Java with its slaves. This result, as we see, is more close to java language program.

Figure 3 shows the results of Java Island with its slaves. It’s clear that the results are more close to Java Island in Indonesia, and there is no images related to java language program.

On the other hand, Figure 4 presents the results of Java Coffee with its slaves. It’s clear that the results do not include neither images for Java language program or Java Island. Therefore, we use the learner’s context of interest and domain knowledge to individualize the content of this target word. We do that by looking for keywords in the user profile (the learner’s context of interest) to help in specifying the intending meaning. Because the target meaning is “computer program language”, we look for slave words in the user profile that best fit this specific meaning—words such as “computer”, “program”, “awt”, “application”, and “swing”.

The main question now is how to specify the core cluster of a query. To overcome this question, we must give answers for the following three questions: How can we construct a dominant meaning for image search? How can the system decide which intended meaning for the image requested? And how can it select words that must be added to the original query?

The following subsections give an answer for each of them in detail.

3. Methodology

This section presents the methodology to cluster the data collected from the Web, and also shows how to use this clusters for forming the model of the dominant meaning.

Figure 5 presents the architecture of our approach to improve the results of Google image search engine. This project follows some instructs to create and then improve the query results of image search.

- Firstly, we collect a specific datasets related to some application domain.
- Using K-means algorithm to cluster the dataset into K-clusters. Each collection is divided into K-classes. Each cluster is related to one meaning and contains some words to identify his meaning called slave words.
- Using dominant meaning algorithm is to classify slave words under its master words to identify the meaning coming from the cluster. This technique generates a hierarchy model for the dominant meaning of each cluster.
- The query is reconstructed based what is appropriate slave words to be added the query can be very important.
- Send the original and the new query independently, to search Google Image Search Engine.
- Choose the top-1000 items coming from the results for both queries.
- Compare the precision and recall of the results for both queries.
Figure 1. The results of Google images for “Java”.

Figure 2. Search results for java with its slaves.
Figure 3. The results of java island with its slaves.

Figure 4. The results of java coffee with its slaves.
3.1. K-Means Algorithm

The procedure of K-means algorithm attempts to find normal groups of data based on some similarity. It classifies a given data set through a certain number of clusters (assume $K$-clusters) fixed a priori. It assigns $K$-point (K-centroids) as one for each cluster. These points must be chosen in a good way because the place of the point impact on the accuracy of the results of clusters. The algorithm will assign each point in the data set to the nearest K-centroid which it divides a set of data points into non-overlapping groups. Therefore, points in a group are “more similar” to one another than points in other groups.

The first step is completed when no point is pending in the queue and an early group-age is done. The standard measure of the spread of a group of points about its centroid is the difference, or the sum of the squares of the distance between each point and the centroid. If the data points are close to the centroid, the difference will be small. The error measure is called the objective function $\Psi$ which is the sum of all the differences:

$$\Psi = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \delta(x_{ij}, z_i)$$

(1)

where the notation $\delta(x_{ij}, z_i)$ stands for the distance between $x_{ij}$ and $z_i$. The $x_{ij}$ is the $j^{th}$ point in the $i^{th}$ cluster, $z_j$ is the reference point of the $i^{th}$ cluster, and $n_i$ is the number of points in that cluster. Accordingly, to reach a delegate clustering $\Psi$ should be as small as possible.

The algorithm is composed of the following steps:

$k$-means algorithm
1) Select $K$ points for initial group centroids.
2) Assign each object to the group that has the closest distance to the centroid.
3) When all objects have been assigned, recalculate the positions of the K-centroids.
4) Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

The results of K-means algorithm contain $m$ clusters. These clusters are used to build the dominant meaning model. The subsection presents the methodology to build this model.

3.2. Construction of Dominant Meanings Tree

Following [7], suppose that the result of clusters $\Pi$ consists of $m$ classes, i.e. $\Pi = \{C_1\}^{m}_{i=1}$, and each cluster $C_i$ is represented by a finite set of WebPages $C_i = \{D_r \mid r = 1, \ldots, r_i\}$.

The question now is how can we use those WebPages to construct dominant meanings for the corresponding cluster?

To overcome this question, each Webpage is represented by a finite set of words $D_r = \{w_j \mid j = 1, \ldots, n_r\}$. A weight $f(w_j)$ is assigned to each term $w_j$ in a document for that term, which depends on the number of occurrences of the term in the document. This weight is a statistical measure used to evaluate how important a word is to a document in a collection of a data set.

The aim of this method is to find a top-$N$ words which represents cluster $C_i$. To complete the computations, suppose that a word $w^d$ represents the cluster $C_i$.

**Dominant Meaning Algorithm (K-Clusters)**

1) Calculate the values of $f(w^d), f(w^d), \forall j, r$
2) Calculate $F^d = \max \left\{ \max_{j=1}^{n_r} f(w^d) \right\}$
3) Define a set $\Gamma_r$ that contains the top-N maximum value of $f_j = f(w^d)$ for a document $W_r$, $\Gamma_r = \{f_j \mid j = 1, \ldots, N\}$, where $0 < f_j < F^d$.
4) For each cluster $C_i$, we rank the terms of collection $\Gamma_r$ in decreasing order. As a result, the dominant meanings of the cluster $C_i$ can be represented by the set of words that is corresponds to the set $f^d$. Return $C_i = \{w^d, w^d, \ldots, w^d\}$.
4. Experimental Results

To ensure that our algorithm works in practice, we conducted experiments with images collected directly from the Web.

4.1. Data Set

The data set consists of 314 web pages from various web sites at the University of Waterloo, and some Canadian websites [10]. The data is categorized into 10 categories as shown in Table 1 and Figure 6.

4.2. Dominant Meaning Model and Formulate Query

Based on K-means and dominant meaning algorithms, Figure 7 shows the hierarchical model the categories of the proposed dataset shown in Table 1. Many research used ontology and meaning to reformulate query [11], and [12]. For example, if we used this model to reformulate a query of a word “Query = \{ w_1, w_2 \}”, we would get the set of corresponding clusters as \{ C_2 \}. We observe that cluster \( C_2 \) contain two words as, \{ w_2, w_2 \}. Consequently, the new query will contains New-Query = \{ w_2, w_2 \}.

4.3. Recall and Precision

Recall is the ability of a retrieval system to obtain all or most of the relevant documents in the collection [13], [14]. The relative recall can be calculated using following the formula: Relative recall = Total number of sites retrieved by a search engine/ Sum of sites retrieved.

To compare two experiments, we use \( F_1 \) performance measure [15] to determine the performance of both of them. It is given by:

\[
\text{precision} = \frac{\# \text{ of correct classes proposed}}{\# \text{ of classes in test data}}
\]

and

\[
\text{recall} = \frac{\# \text{ of correct classes proposed}}{\# \text{ of classes proposed}}
\]

\[
F_1 = \frac{2 \times \text{precision} \times \text{recall}}{(\text{precision} + \text{recall})}
\]

As shown in Table 3 and Figure 6, in case of the Query using Dominant meaning, searching data campus-network (D2) had the highest relative pr value (0.71) followed by the data set snowboarding-skiing (0.69) with the least relative recall for the data set river-rafting (D8) (0.41).

As shown in Figure 6, in case of bag-of-words query, searching data campus-network (D2) had the highest relative recall value (0.57) followed by the data set snowboarding-skiing (0.56) with the least relative recall for the data set river-rafting (D8) (0.23).

As shown in Figure 8, in case of the Query using dominant meaning, searching data black-bear-attack (D1) and winter-Canada (D10) had the highest relative precision value (0.57) followed by the data set snowboarding-skiing (0.69) with the least relative precision for the data set campus-network (D2) (0.39).
Figure 9 shows a comparison for average precision for query using dominant meaning vs. query using bag-of-words. In case of bag-of-words query, searching data campus-network (D9) had the highest relative precision value (0.43) followed by the data set winter-Canada (D10) with (0.56) with the least relative precision for the data set Canada-transportation-roads (D3) with (0.27).

Figure 10 shows the $F_1$-measures for each application domain in the cluster for both the original query and the reformulated query using the proposed technique. The highest values are for D1 and D10 with $F_1$-measures 0.61, and 0.61 respectively.

We also notice that our approach can achieve better performance in terms of $F_1$ for categories D2, D3 with the same value (0.5). It is clear that the query which is reformulated with the dominant meaning approach has a great improving for the results than the original query. For the improving in $F_1$ values of the best four categories of the testing dataset (D1, D2, D5, and D4), we can see that, compared with the original query, improve the $F_1$ measure by 17.9%, 15.4%, 13.3%, and 12.4%, respectively.

5. Conclusion

In this article, we studied the effectiveness of reformulating the query using a dominant meaning technique on the results of Google image search engine. To apply the technique, we used the dataset which consists of 314 web pages classified into 10 categories. K-means algorithm is used to cluster each category in the dataset into K-clusters. We applied the dominant meaning algorithm on each cluster to extract some meaning to build a hierarchy model. We used this model to reconstruct a new query. We investigated into the influence of the results coming from Google search engine to the performance of the original query and the restructured query. As experimental results shown, the proposed technique in this paper had a considerable performance for precision, recall and $F_1$-measure.

6. Acknowledgements

This project was funded by Deanship of Scientific Research, King Abdulazize University, Saudi Arabia, under award number (521/214/1432). And, the authors would like to thank the Deanship of Scientific Research for their supports and encouragements.

REFERENCES


