Intelligent Information Management, **2013, 5, 196-203**
Published Online November 2013 (http://www.scirp.org/journal/iim)
http://dx.doi.org/10.4236/iim.2013.56022

Towards More Efficient Image Web Search

Mohammed Abdel Razek[1,2]
[1]Deanship of E-Learning and Distance Education, King Abdul-Aziz University, Jeddah, KSA
[2]Mathematics 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

Copyright © 2013 Mohammed Abdel Razek. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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 F_1-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 context 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.

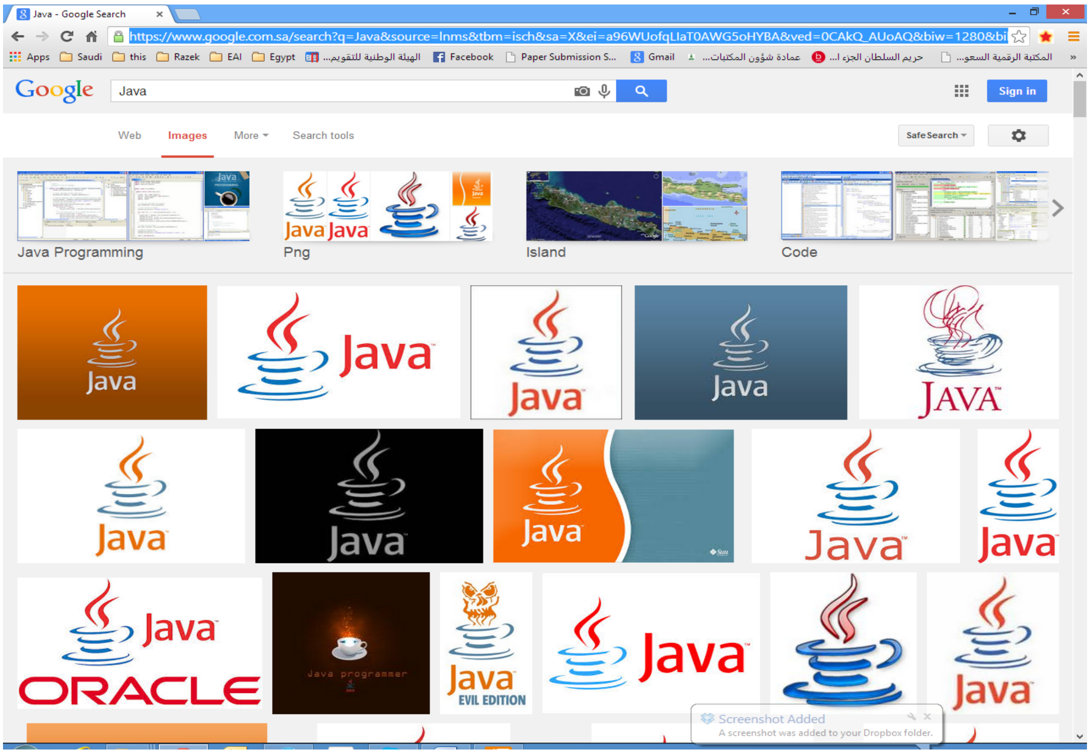

Java - Google Search
https://www.google.com.sa/search?q=Java&source=lnms&tbm=isch&sa=X&ei=a96WUofqLIaT0AWG5oHYBA&ved=0CAkQ_AUoAQ&biw=1280&bi
Apps Saudi this Razek EAI Egypt Facebook Paper Submission S... Gmail
Google
Java
Sign in
Web
Images
More
Search tools
SafeSearch
Java Programming
Png
Island
Code
Java
JAVA
ORACLE
Java programmer
Java EVIL EDITION
Screenshot Added
A screenshot was added to your Dropbox folder.

Figure 1. The results of Google images for "Java".

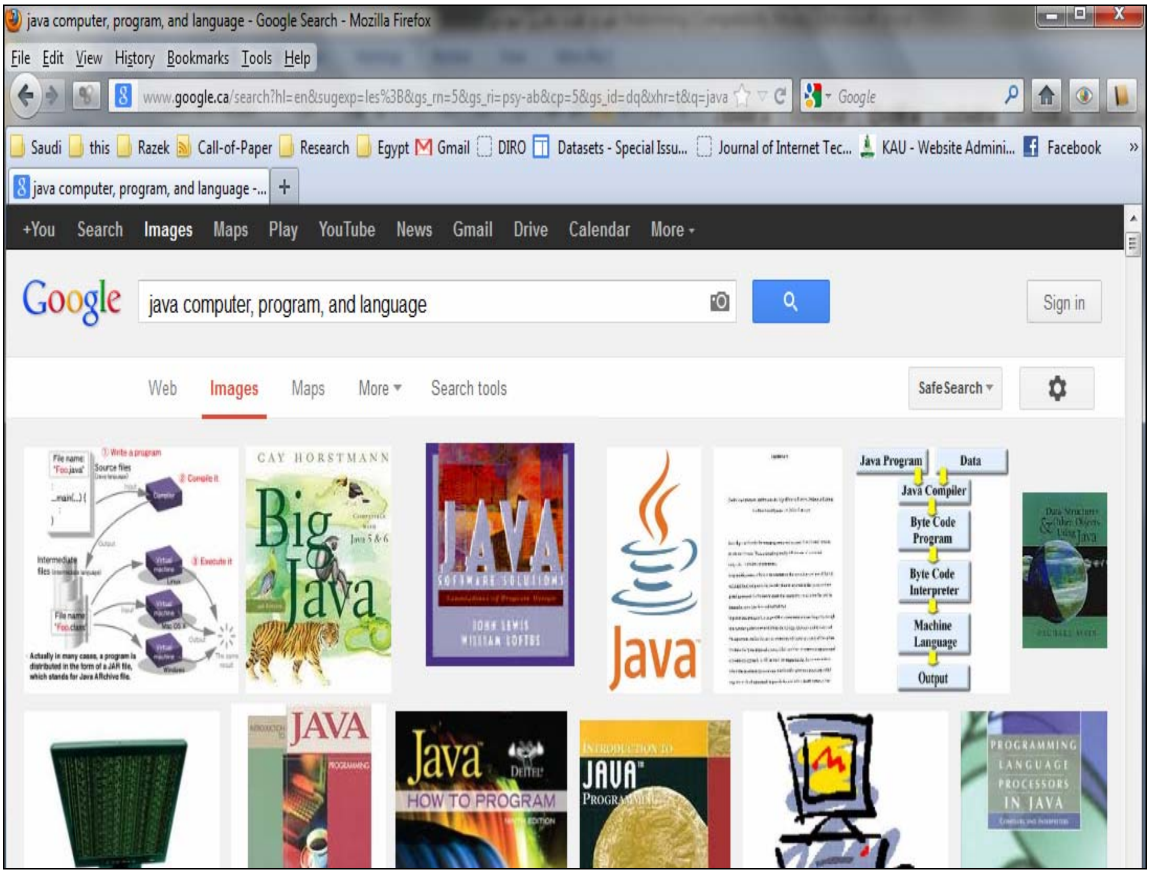
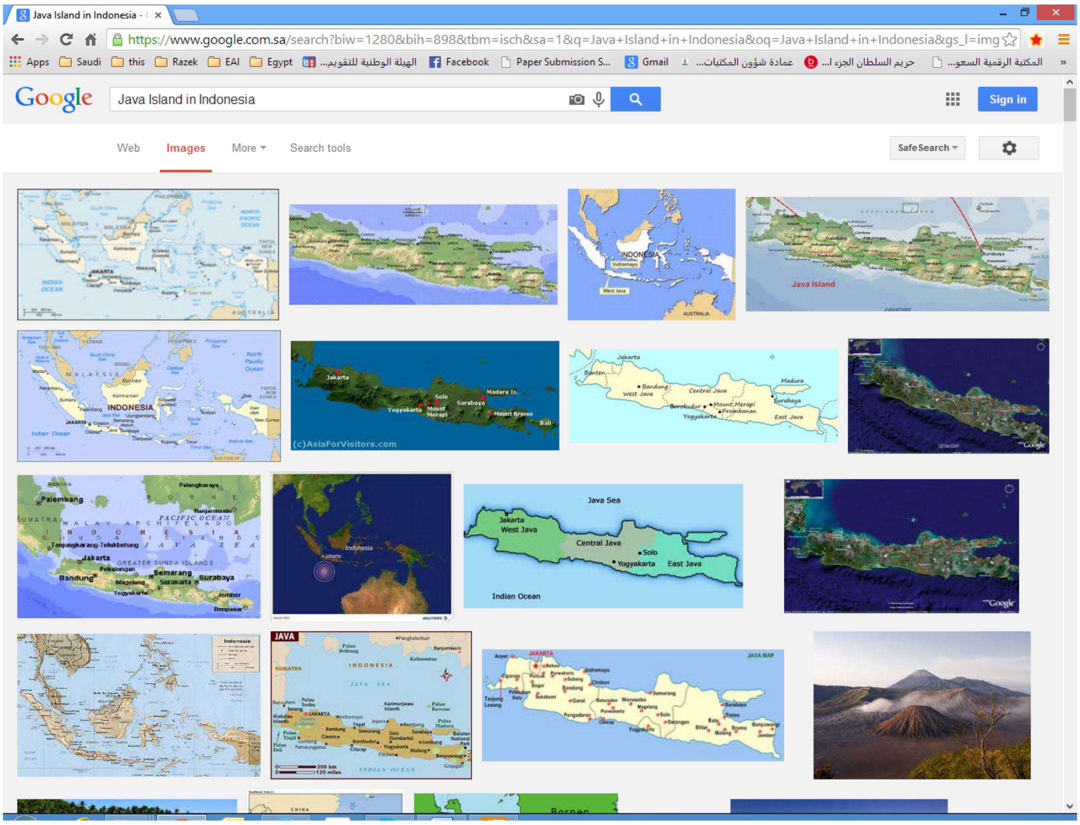

java computer, program, and language - Google Search - Mozilla Firefox
File Edit View History Bookmarks Tools Help
www.google.ca/search?hl=en&sugexp=les%3B&gs_rn=5&gs_ri=psy-ab&cp=5&gs_id=dq&xhr=t&q=java
Google
Saudi this Razek Call-of-Paper Research Egypt Gmail DIRO Datasets - Special Issu... Journal of Internet Tec... KAU - Website Admini... Facebook
java computer, program, and language -...
+You Search Images Maps Play YouTube News Gmail Drive Calendar More
Google
java computer, program, and language
Sign in
Web Images Maps More Search tools
SafeSearch
Big Java
CAY HORSTMANN
JAVA
Java
Java Program
Data
Java Compiler
Byte Code Program
Byte Code Interpreter
Machine Language
Output
JAVA
Java HOW TO PROGRAM
INTRODUCTION TO JAVA PROGRAMMING
PROGRAMMING LANGUAGE PROCESSORS IN JAVA

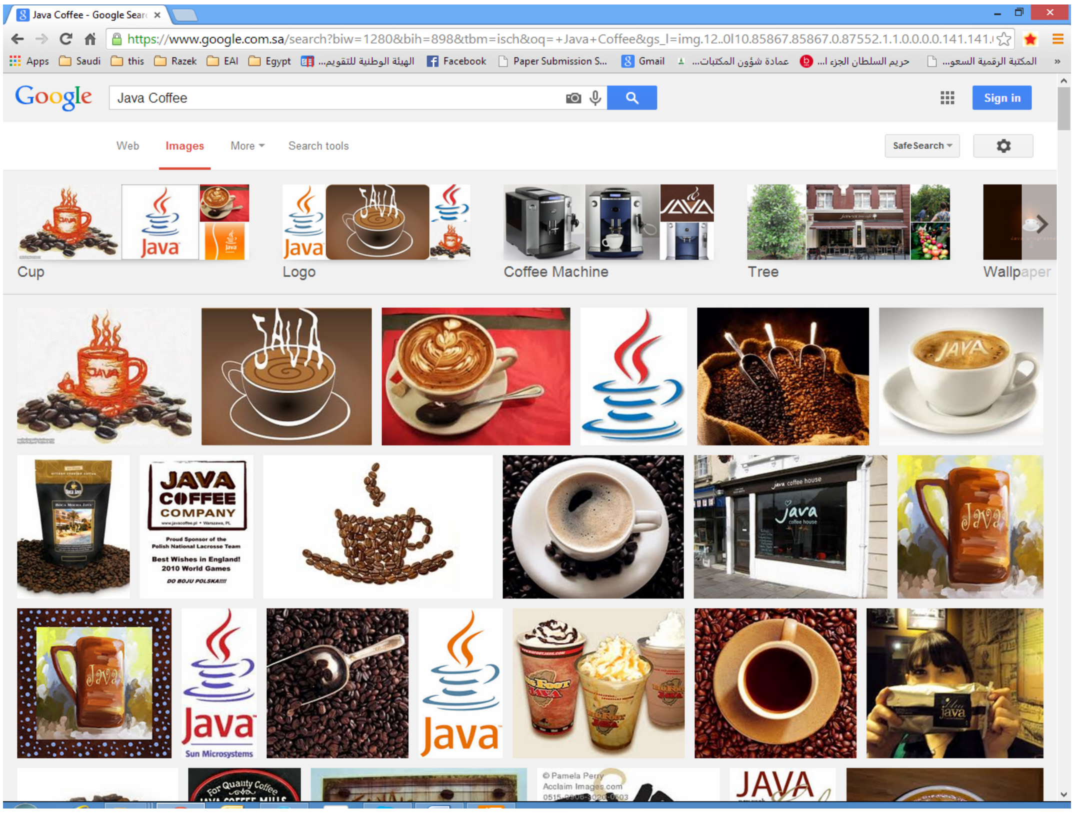
Figure 2. Search results for java with its slaves.



https://www.google.com.sa/search?biw=1280&bih=898&tbm=isch&sa=1&q=Java+Island+in+Indonesia&oq=Java+Island+in+Indonesia&gs_l=img

Java Island in Indonesia

Figure 3. The results of java island with its slaves.



https://www.google.com.sa/search?biw=1280&bih=898&tbm=isch&oq=+Java+Coffee&gs_l=img.12..0l10.85867.85867.0.87552.1.1.0.0.0.0.141.141.

Java Coffee

Figure 4. The results of java coffee with its slaves.

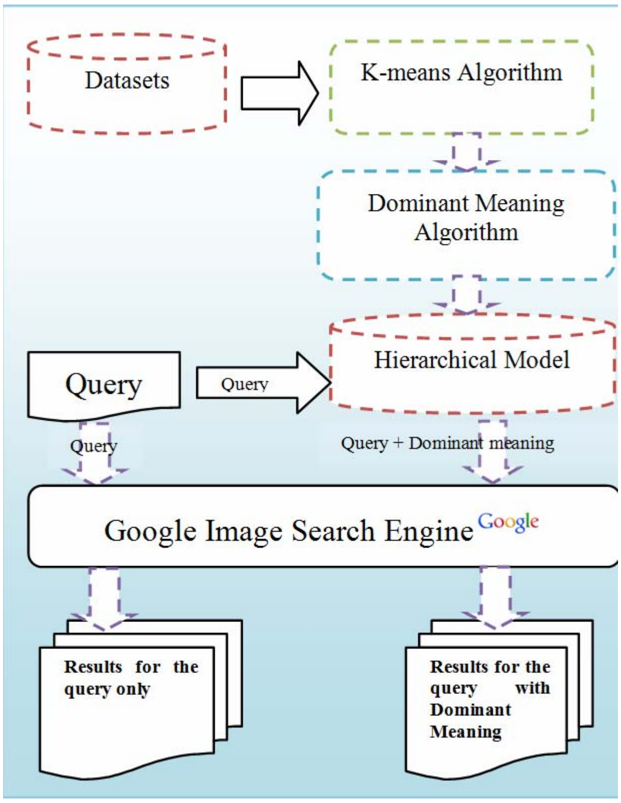

Datasets
K-means Algorithm
Dominant Meaning Algorithm
Hierarchical Model
Query
Query
Query
Query + Dominant meaning
Google Image Search Engine Google
Results for the query only
Results for the query with Dominant Meaning

Figure 5. Architecture of the methodology.

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 centroids 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 Ψ which is the sum of all the differences:

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

where the notation $\delta\left(x_{ij}, z_i\right)$ stands for the distance between x_{ij}, and z_i. The x_{ij} is the j^{th} point in the i^{th} cluster, z_i 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 Ψ 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 Π consists of m classes, *i.e.*

$\Pi = \{C_k\}_{k=1}^{m}$, and each cluster C_k is represented by a finite set of WebPages

$$C_k = \left\{D_r \mid r = 1, \ldots, r_k\right\}.$$

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 $D_r = \{w_{rj} \mid j = 1, \ldots, n_r\}$. A weight $f\left(w_{rj}^k\right)$ 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_k. To complete the computations, suppose that a word w^k represents the cluster C_k.

```
Dominant Meaning Algorithm (K-Clusters)
1) Calculate the values of $f\left(w^k\right), f\left(w_{jr}^k\right), \forall j, r$
2) Calculate $F^k = \underset{j=1,\ldots,r_k}{Max}\left\{\underset{v=1,\ldots,n_r}{Max}\left\{f\left(w_{vj}^k\right)\right\}\right\}$
3) Define a set $\Gamma_r$ that contains the top-N maximum value of $f_j^r = f\left(w_{jr}^k\right)$ for a document $W_r$ $\Gamma_r = \left\{f_j^r \mid j = 1, \cdots, N\right\}$, where $0 < f_j^r < F^k$.
4) For each cluster $C_k$, we rank the terms of collection $\Gamma_r$ in decreasing order. As a result, the dominant meanings of the cluster $C_k$ can be represented by the set of words that is corresponds to the set $f_j^r$. Return $C_k = \left\{w_1^k, w_2^k, \cdots, w_T^k\right\}$.
```

4. Experimetal 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 Quarry

Based on K-means and dominant meaning algorithms, **Figure 7** shows the hieratical 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 "$\text{Query} = \left\{ w_2^1 \right\}$", we would get the set of corresponding clusters as $\left\{ C_2 \right\}$. We observe that cluster C_2 contain two words as, $\left\{ w_2^1, w_2^2 \right\}$ Consequently, the new query will contains

$\text{New-Query} = \left\{ w_2^1, w_2^1 \right\}$.

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).

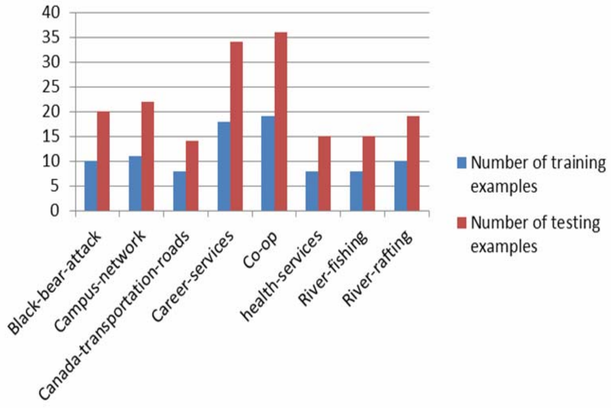


Figure 6. Number of training and testing examples.

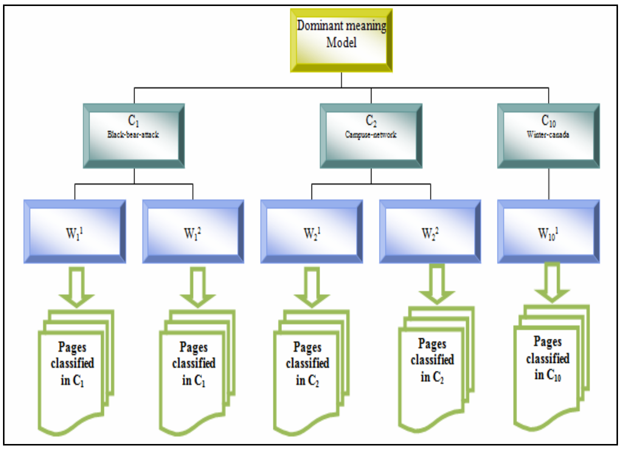


Figure 7. Dominant meaning model of the dataset.

Table 1. Number of WebPages in dataset.

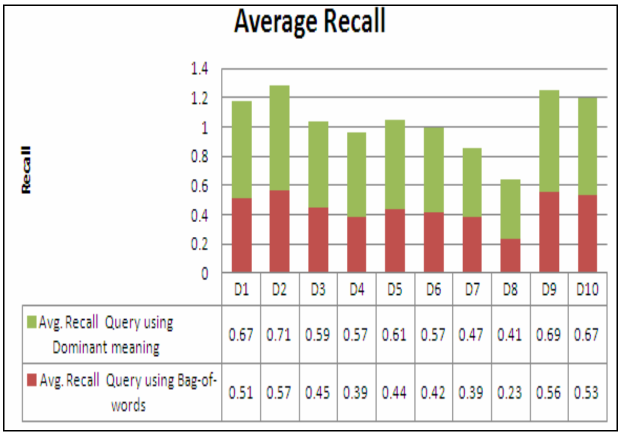
	Subject	Number of WebPages
D1	Black-bear-attack	30
D2	Campus-network	33
D3	Canada-transportation-roads	22
D4	Career-services	52
D5	Co-op	55
D6	Health-services	23
D7	River-fishing	23
D8	River-rafting	29
D9	Snowboarding-skiing	24
D10	Winter-Canada	23
	Total	314

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).

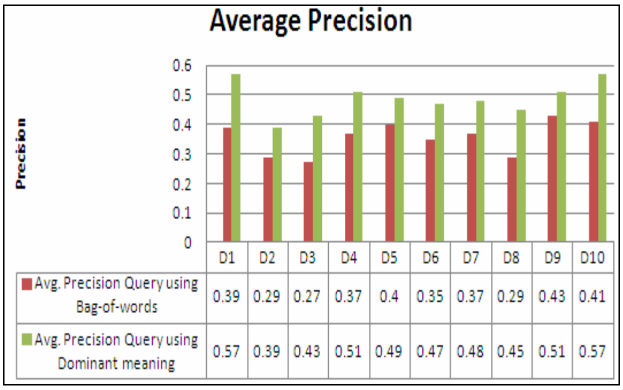
Figure 10 shows the F_1-measures for each application domain in the cluster for both the original query and the reformulated query using the proposed technique. The highest values are for D1 and D10 with F_1-measures 0.61, and 0.61 respectively.

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.



	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
🟩 Avg. Recall Query using Dominant meaning	0.67	0.71	0.59	0.57	0.61	0.57	0.47	0.41	0.69	0.67
🟥 Avg. Recall Query using Bag-of-words	0.51	0.57	0.45	0.39	0.44	0.42	0.39	0.23	0.56	0.53

Figure 8. Average recall for query using dominant meaning vs. query using bag-of-words.



	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
🟥 Avg. Precision Query using Bag-of-words	0.39	0.29	0.27	0.37	0.4	0.35	0.37	0.29	0.43	0.41
🟩 Avg. Precision Query using Dominant meaning	0.57	0.39	0.43	0.51	0.49	0.47	0.48	0.45	0.51	0.57

Figure 9. Average precision for query using dominant meaning vs. query using bag-of-words.

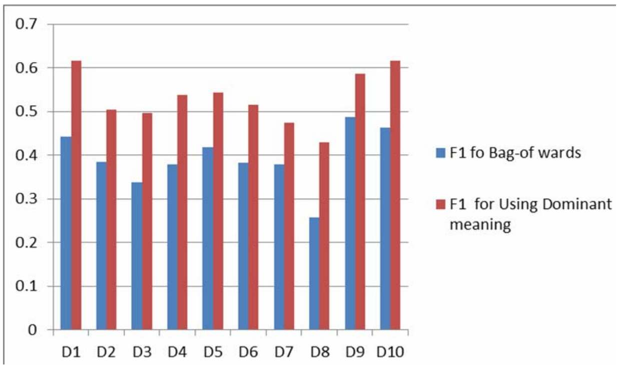


Figure 10. F_1-measures for the original query and for the query with dominant meaning.

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

[1] X. Wang, S. Qiu, K. Liu and X. Tang, “Web Image Re-Ranking Using Query-Specific Semantic Signatures,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, No. 99, 2013, pp. 1-14.

[2] S. Sujatha and A. S. Sona, “New Fast K-Means Clustering Algorithm Using Modified Centroid Selection Method,” *International Journal of Engineering Research & Technology* (*IJERT*), Vol. 2, No. 2, 2013, pp. 1-9.

[3] C. Zhang and S. X. Xia, “K-Means Clustering Algorithm with Improved Initial Center,” 2*nd International Workshop on Knowledge Discovery and Data Mining* (*WKDD*), Moscow, 23-25 January 2009, pp. 790-792.

[4] M. Gautam and A. Xavier, “Speed Improvements to Information Retrieval-Based Dynamic Time Warping Using

Hierarchical K-Means Clustering," 2013 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Vancouver, 26-31 May 2013, pp. 8515-8519.

[5] D. Mavroeidis and P. Magdalinos, "A Sequential Sampling Framework for Spectral k-Means Based on Efficient Bootstrap Accuracy Estimations: Application to Distributed Clustering," *ACM Transactions on Knowledge Discovery from Data*, Vol. 7, No. 2, 2012, pp. 2-7.

[6] J. Wu, H. Xiong and J. Chen, "Adapting the Right Measures for k-Means Clustering," *Proceedings of the* 15*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, 28 June-1 July 2009,, pp. 877-886.

[7] M. A. Razek, C. Frasson and M. Kaltenbach, "Dominant Meanings towards Individualized Web Search for Learning Environments," In: G. D. Magoulas and S. Y. Chen, Eds., *Advances in Web-Based Education*: *Personalized Learning Environments*, IDEA Group Publishing, Hershey, 2006.

[8] Y. Jing, M. Covell, D. Tsai and J. M. Rehg, "Learning Query-Specific Distance Functions for Large-Scale Web Image Search," *IEEE Transactions on Multimedia*, Vol. 15, No. 8, 2013, pp. 2022-2034.

[9] G. Maderlechner, J. Panyr and P. Suda, "Finding Captions in PDF-Webpages for Semantic Annotations of Images," In: D.-Y. Yeung, *et al*., Eds., *Structural*, *Syntactic*, *and Statistical Pattern Recognition*, Lecture Notes in Computer Science Volume, Springer-Verlag Berlin Heidelberg, 2006, pp. 422-430.

[10] K. Hammouda, "Web Ming Dataset," 2013. http://pami.uwaterloo.ca/~hammouda/webdata

[11] P. Singh, R. H. Goudar, R. Rathore, A. Srivastav and S. Rao, "Domain Ontology Based Efficient Image Retrieval," 7*th International Conference on Intelligent Systems and Control* (*ISCO*), Coimbatore, 4-5 January 2013, pp. 445-452. http://dx.doi.org/10.1109/ISCO.2013.6481196

[12] D. Gowsikhaa, S. Abirami and R. Baskaran, "Construction of Image Ontology Using Low-Level Features for Image Retrieval," *International Conference on Computer Communication and Informatics* (*ICCCI*), Coimbatore, 10-12 January 2012, pp. 1-7. http://dx.doi.org/10.1109/ICCCI.2012.6158922

[13] B. T. Sampath Kumar and J. N. Prakash, "Precision and Relative Recall of Search Engines: A Comparative Study of Google and Yahoo," *Singapore Journal of Library & Information Management*, Vol. 38, No. 1, 2009, pp. 124-137.

[14] S. M. Shafi and R. A. Rather, "Precision and Recall of Five Search Engines for Retrieval of Scholarly Information in the Field of Biotechnology," *Webology*, Vol. 2, No. 2, 2005, pp. 42-47. http://www.webology.ir/2005/v2n2/a12.html

[15] M. Gagnon, A. Zouaq and L. Jean-Louis, "Can We Use Linked Data Semantic Annotators for the Extraction of Domain-Relevant Expressions?" *The International World Wide Web Conference Committee* (*IW*3*C*2), *WWW* 2013 *Companion*, Rio de Janeiro, 13-17 May 2013, pp. 1239-1246.