2012

TsoKaDo: An Image Search Engine Performing Recursive Query Recommendation Based on Visual Information

Tsochatzidis, Lazaros T.

IARIA

TsoKaDo: An Image Search Engine Performing Recursive Query Recommendation Based on Visual Information

Lazaros T. Tsochatzidis Athanasios Ch. Kapoutsis Nikos I. Dourvas
Savvas A. Chatzichristofis Yiannis S. Boutalis Konstantinos Zagoris
Department of Electrical & Computer Engineering,
Democritus University of Thrace, Xanthi, Greece
{lazatsoc, athakapo, nikodour, schatzic, ybout, kzagoris}@ee.duth.gr

***Abstract*—This paper tackles the problem of the user's incapability to describe exactly the image that he seeks by introducing an innovative image search engine called TsoKaDo. Until now the traditional web image search was based only on the comparison between metadata of the webpage and the user's textual description. In the method proposed, images from various search engines are classified based on visual content and new tags are proposed to the user. Recursively, the results get closer to the user's desire. The aim of this paper is to present a new way of searching, especially in case with less query generality, giving greater weight in visual content rather than in metadata.**

Keywords-Image Retrieval; Metadata; Image Annotation; Query Recommendation Systems.

I. INTRODUCTION

In the last few years the idea of Computer-Human Interaction is evolving continuously with fast steps. It is believed that this is the most crucial and promising direction to research for the technological future of humanity. Nowadays, there are many applications belonging to Computer-Human Interaction systems. A great example of that is the web search engines on the Internet and more specific the image search engines. Images are a very important part of our daily life. Google, Bing, Ask or Flickr, which are visited by millions of people follow a web image searching procedure which is based on the keywords of the user and the metadata of the images. Metadata can be keywords, tags or any other information from the web page that the image belongs to. But, there are some problems with that technique. Sometimes, images are often available without any metadata. In other cases, the annotation of the images is not correct and does not correspond to the actual description of the image (noisy annotations). Furthermore, the disadvantage of using textual query (keyword) for the image search is that the user is not always fully capable of describing his exact wish. As it is commonly said, 'one image is consisting of a thousand words', so it is impossible for the seeker to describe the image exactly as it is. As a result, the images proposed to him by the web search engines are often not relevant and far from his desires.

On the other hand, there could be a method that is based only on low level features of the image without using any textual information. Content Based Image Retrieval (CBIR) is based on the visual content of the image, e.g., color, texture, shape or information from local patches. CBIR is defined as any technology that in principle helps to organize digital image archives by their visual content [1]. By this definition, anything ranging from an image similarity function to a robust image annotation engine falls under the purview of CBIR. But, the 'weak spot' of CBIR is that it seems to be notoriously noisy for image queries of low generality [2]. If the query image happens to have a low generality, early rank positions may be dominated by spurious results. As a result, the simultaneous employment of CBIR techniques and metadata were found to be significantly more effective and reduce the communication gap that exists between the Humans and the Computers more than the text-only and image-only baseline [3].

In this paper, we propose a new query expansion recommendation system which tries to reduce the semantic gap between '**What**' the user wants to find and '**How**' he describes it. Query expansion is the process of reformulating a seed query to improve retrieval performance in information retrieval operations. The proposed system initially uses several parsers to take images from Google, Bing and Ask image search engines for a given textual query (keyword). More details about parsers are given in Section 2. Then, utilizing Color and Edge Directivity Descriptor (CEDD) [4] the visual content of these images are described. More details about the description of the visual content are given in Section 3. Next, the well-known K-means classifier, separates the image descriptors in a preset or in a dynamically calculated number of clusters. In order to calculate the number of clusters needed for each image set, a Self Growing and Self Organized Neural Gas Network (SGONG) is employed. More details about SGONG are given in Section 4. The images, whose descriptors have the minimum distance from the center of each class, are considered to describe better the subject of the current class. Consequently, those descriptors are used to retrieve the top-K, visually similar, images from

 ISBN: 978-1-61208-177-9

a collection of ones parsed from Flickr, using the same keyword with the other three search engines. The manually annotated tags of these top-K images are retrieved and, subsequently, using Wu and Palmer [5] method, we are calculating the semantic similarity between these tags and classify them into C classes. This process is illustrated in Section 5.

Depending on the number that each tag appears, a tag cloud is constructed. Those tags will be used afterwards to improve the query and the search process. The entire procedure is described in details in Section 6 while early experimental results are drawn in Section 7. Finally, the conclusions are given in Section 8. A preliminary version of this paper has been presented in [6].

The ideal scenario looks like this: A user inputs the keyword 'Paris'. The application searches through the internet using the three popular search engines with this keyword and generates a pool of results. The CEDD of each image is computed and the results are classified into two classes. The images, whose distance from the classes center is the minimum, are used for the retrieval of top-K relevant images, among the ones parsed from Flickr, under the same keyword. From those images and after the semantic classification of their tags, two classes are constructed: one with the tags 'Paris Eiffel Tower' and another one with 'Paris Hilton'. By clicking on a tag the application repeats the whole process using the improved query. So, the user of our web search engine can communicate in more specific way with his computer, and finally through this improved interaction to find exactly what he wishes.

Several re-ranking by visual content methods has been seen before, but mostly in different setups than the one we consider or for different purposes [7], [8]. Some of them used external information, e.g., an external set of diversified images [9] or training data [10]. The authors in [11] proposed a similar to our approach tag recommendation system based on visual similarity. According to their approach, the main problem of today's image search is the ambiguous definition of tags given from users. They are trying to bypass this phenomenon by recommending, the closest to image, tags. This is achieved by extracting the low-level visual features of the image. In our paper, we take advantage of this optimization of tags given to images, so we can propose more effective queries to users and help them get better results according to their desires.

II. PARSERS

The word 'parse' means to analyze an object specifically. Parsing refer to breaking up ordinary text. For example, search engines typically parse search phrases entered by users so that they can more accurately search for each word. Some programs can parse text documents and extract certain information like names or addresses.

In this occasion, parsers are used to search in the produced html (Hypertext Markup Language) page of each search engine (Google, Bing, Ask, flickr) to find the url of each image they return. Despite the fact that web search engines propose their Application Program Interfaces (APIs) for performance of text search, they do not have any APIs for image searching, except from Flickr. So, it was urgent need to generate parsers to bypass that serious obstacle.

III. COLOR AND EDGE DIRECTIVITY DESCRIPTOR (CEDD)

The recently proposed Color and Edge Directivity Descriptor belongs to the family of Compact Composite Descriptors (CCDs)[12]. An important thing about CEDD is that it uses only 54 bytes per image for indexing them, rendering this descriptor suitable for use in large image databases. Also the results of CEDD are very effective so the descriptor is very suitable for this web application.

In the technical part, CEDD initially separates images into a preset number of blocks. Each image block is classified in to one, or more than one of the $n = 6$ preset texture areas. Each texture area consists of $m = 24$ sub-regions. Using 2 fuzzy systems, CEDD classifies the colors of the image blocks in a 24-color custom palette. Then texture extraction is achieved by a fuzzy version of five digital filters proposed by MPEG-7 Edge Histogram Descriptor, forming 6 texture areas. When CEDD is used to describe an image block, each section of the image goes through 2 units: 1) the color unit and 2) the texture unit.

The color unit classifies the image block into one of the 24 shades used by the system in a color area, $m, m \in (0, 23)$. The texture unit classifies the image into a texture area, $n, n \in (0, 5)$. The image block is classified in the bin $n \times 24 + m$. This process is repeated for all the image blocks. At the end, the histogram produced, is normalized within the region [0, 1] and quantized for binary representation in a three bits per bin quantization.

IV. SELF-GROWING AND SELF-ORGANIZED NEURAL GAS NETWORK

The Self Growing and Self Organized Neural Gas Network (SGONG) [13] is an innovative neural classifier. This network was proposed in order to reduce the number of colors in a digital image. It collects the advantages of the Growing Neural Gas (GNG) and the Kohonen Self-Organized Feature Map (SOFM) neural classifiers. The main advantage of the SGONG network is that it controls the number of created neurons and their topology in an automatic way. This feature is very important for that web application because it provides a method to compute the number of the classes automatically.

There are also some other characteristics of this neural classifier:

 ISBN: 978-1-61208-177-9

- The dimensions of the input space and the output lattice are always identical.
- In order to determine the classes and to ensure fast convergence, it uses some criteria.
- Except for color components, the SGONG neural network can also be used if its entrance is other local spatial features.
- The color reduction results obtained are better than the other two techniques which it combines.
- For the training procedure, the Competitive Hebbian Rule (CHR) is used to dynamically create or remove the connections of neurons.

As it is mentioned above, the SGONG uses some criteria in order to determine the number of created neurons. At the end of each epoch, three criteria that modify the number of the output neurons and make the proposed neural network to become self-growing are applied. These criteria are applied in the following order:

- remove the inactive neurons,
- add new neurons,
- and finally, remove the non important neurons.

V. WORDNET-BASED SEMANTIC SIMILARITY MEASUREMENT

To define sematic similarity we have to consider the calculation of the conceptual similarity between words that are not lexicography similar such as the word 'car' and the word 'automobile'. These procedures is accomplished by comparing the results of their relationship with a third ontology (like 'wheeled vehicle' in our example). In general sematic similarity helps to detect duplicate (high scores) or complementary (medium scores) content. In bibliography there are several methods for this particular job. The four most important are [14]:

- the edge counting method
- the information content method
- the feature based method
- the hybrid method.

Wordnet version 3.0 (2006) [15] is a lexical database which is available online and provides a large repository of English lexical terms. Wordnet was designed to establish the connections between four type parts of Speech (POS) such as noun, verb, adjective and adverb. Those POS's are grouped into synonyms sets called synsets which are the smallest units in Wordnet and represent terms or concepts. The synsets are also organized into 'senses' which basically represent different meanings of the same term.

To calculate the semantic similarity between two synsets, the path length measurement was used. In [16], the authors are using a similar to our approach for enchanting search privacy on the Internet, focusing on plausible deniability against search engine query-log. Path length uses hyponyms and hypernyms. The hypernym represents a certain set of discrete objects and the hyponym represents a smaller part of the hypernym. The taxonomy of the two synsets is treated as an undirected graph and measures the distance between them in Wordnet. In this graph, a shared parent of two synsets is known as a sub-sumer. The Least Common Sub-sumer (LCS) of two synsets is the sumer that does not have any children that are also the sub-sumer of two synsets. In this paper the method followed for this process is the Wu and Palmer similarity metric. This method measures the depth of the two synsets in the Wordnet taxonomy, and the depth of the least common sub-sumer (LCS) and combines these figures into a similarity score given below:

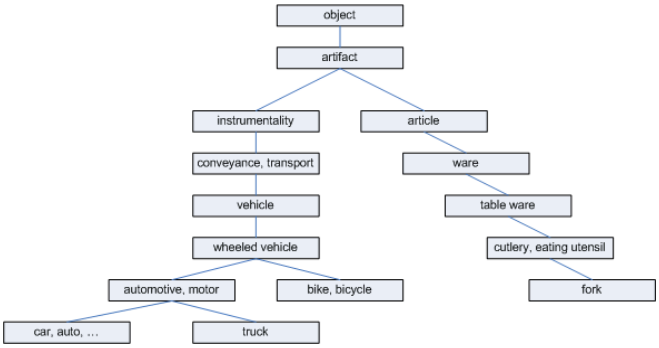

Figure 1. Example of the Hyponym Taxonomy in Wordnet Used For Path Length Similarity Measurement

$$\text{Sim} = \frac{2 \times \text{depth(LCS)}}{\text{depth(Synset1)} + \text{depth(Synset2)}} \quad (1)$$

In contrary to the other dictionaries, Wordnet does not have any information about etymology, pronunciation and the forms of irregular verbs but only bounded information about the usage.

The actual lexicographical and semantic information is maintained in lexicographer files, which are then processed by a tool called grind to produce the distributed database. Both grind and the lexicographer files are freely available in a separate distribution, but modifying and maintaining the database requires expertise.

VI. IMPLEMENTATION - METHOD OVERVIEW

All the technical characteristics described before are combined into an image web search engine called TsoKaDo.

At first, the user is asked to provide the search engine with an initial query, the number of images to be fetched from each search engine (M) and the number of classes to be created (K). Using, the user's query, TsoKaDo fetches the top-M results from the well-known search engines Google [17], Bing [18], Ask [19] and Flickr [20]. The difference between those search engines is that the Flickr store human imported tags for each image and this is TsoKaDo's source of tags. A second diference between those is in their parsers as it mentioned in Section II.

 ISBN: 978-1-61208-177-9

In order the image's visual information to be used, the Color and Edge Directivity Descriptor (CEDD) is extracted. The CEDD will produce a vector of 144 numbers that describes the color and the texture areas of the image. These vectors will be used from now on, and each consequent process will be taking place in the R^{144} space.

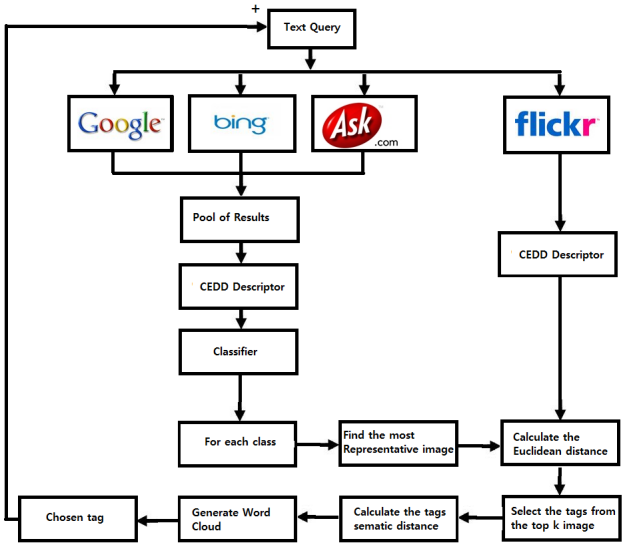

Figure 2. Steps Followed by Tsokado Image Web Search Engine.

At the next step, the pool of results of the first three search engines (Google, Ask, Bing), which is defined as:

$$P = M_{\text{Google}} \cup M_{\text{Bing}} \cup M_{\text{Ask}} \quad (2)$$

are classified using K-means algorithm or the Self-Growing and Self-Organized Neural Gas network (SGONG), depending on the users choice. If the user has selected an explicit number of classes ($K = 2, 3, 4, 5$) only the K-means algorithm is used. Otherwise, if the option 'Auto' is selected, SGONG will propose the number of classes needed for reliable classification and the K-means will classify the images.

For each class, having its center available, TsoKaDo determines the image that has the least Euclidian distance from the class center. This image is considered to be the most representative image of the whole class and it is compared with each image returned from Flickr in the first step. The comparison is being performed by calculating the Euclidian distance and the tags of the image with the least one are fetched to describe the class.

Finally, in order to filter the tags collected, TsoKaDo uses Wordnet to calculate the semantic distance between them and stores them in a $Y \times Y$ array, where Y the number of tags for the class. These are sorted according to their appearance frequency and then are proposed to the user by a tag cloud.

VII. Experimental Results

In order to show the functionality of TsoKaDo, some representative examples will be presented in this section.

For the first example, the keyword 'Greece' is chosen and various images about Greece are being retrieved from the search engines such as landscapes, islands and maps. These images are being classified into classes as it is shown in Figure 3.

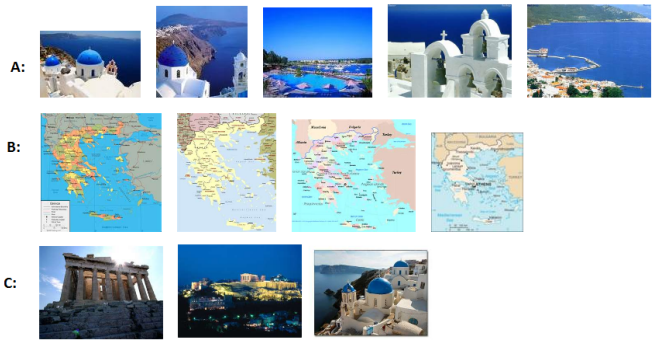

Figure 3. Classification of Images Under the Keyword 'Greece'.

Finally, new keywords are being proposed to the user for each class (see Figure 4). The tag cloud of class C, illustrated in Figure 4, contains the keywords 'Parthenon' and 'Thessalonica' and Class A tag cloud contain the terms 'Athens', 'Cyclades' and 'Santorini', which are all famous and popular destinations for vacations in Greece, also shown in the images of the class. In addition, class B contains many maps available on the web about Greece.

Figure 4. Tag clouds for Some Classes of Figure 3.

The user proceeds to the next step choosing a keyword to expand his search. In this example we choose to search more about 'Santorini and the search engines response is shown in Figure 5 and the tag clouds in Figure 6, respectively.

As it is clearly shown, the new results of TsoKaDo are about the island Santorini only. In the class A there are some maps of the island and in the other classes some photos and landscapes. If the user wants to expand further his search, there is the keyword 'stairs' available and the final result is shown in Figure 7. To conclude with, this example demonstrates how effective the search can be extended with new keywords based on visual information.

This second example intends to demonstrate the advantages of using multiple search engines. In Figure 8 presents

Figure 5. Classification of images with query extended to 'Greece Santorini.

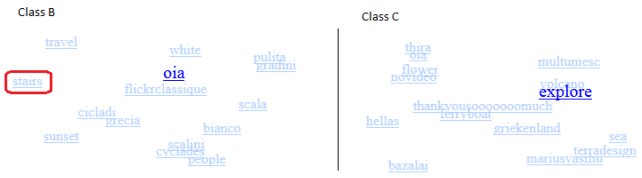

Figure 6. Tag clouds for Some Classes of Figure 5 - Tag: 'Santorini'.

the top results of the three such engines that TsoKaDo uses, given the keyword 'food'.

From the examples above, we conclude to the fact that TsoKaDo has two major advantages against conventional search engines:

To begin with, it combines the results of three search engines which very often return completely different results. It is known that every search engine stores a rank of web pages according to their importance based on some criteria.

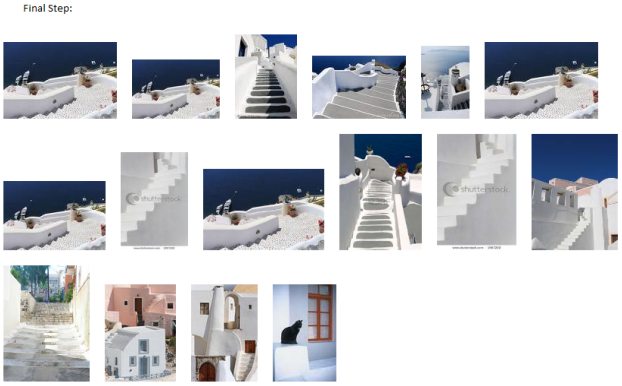

Figure 7. Filtering 'Santorini' Results Using 'Stairs' Tag.

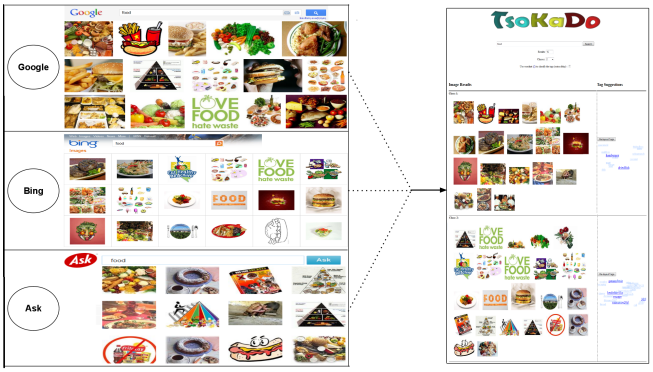

Figure 8. Comparison's results between TsoKaDo and the other search engines (for the same query 'food')

Combining three search engines gives the advantage of using three different ranks so it is even better than fetching more results from one engine regarding the diversity and quality of the images.

Furthermore, the most important advantage is that TsoKaDo manages to propose useful tags to the user. As it is mentioned earlier, the source of tags of this search engine is the Flickr which contains a huge database of images and photographs tagged with human imported keywords. This offers satisfying semantic recognition of objects, persons and locations that are depicted. The downside of using human imported tags from Flickr is that sometimes the keywords fetched are 'noisy'. As it is shown in Figure 9, besides the useful keywords there are many irrelevant terms. This can be improved at some point using Wordnet but this process may lose some useful data too. In Figure 9, it is shown how a tag cloud of a class with the query 'food is filtered using WordNet.

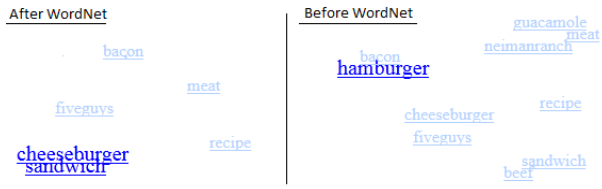

Figure 9. Filtering of tags using WordNet.

VIII. Conclusions and Future Work

In this paper, a new method in query expansion was proposed. With this method, users can expand their query with new keywords that are proposed to them relying on the extraction of low level visual features. This new image web search engine called TsoKaDo has a result of images, more corresponding to the users requirements.

Recently, Google introduces a new feature named 'sort by subject' [21], which at first may seem particularly close

to TsoKaDo's function. However, this is not true because Google does not use visual information of the image. Given that Google has access to almost every site in the web, it can determine keywords that in many important sites appear together. So it assumes they are relevant and proposes them to the user.

Although TsoKaDo offers something new in the field of image searching, it still has problems to deal with. At first TsoKaDo creates these classes because CEDD is based only in the color and the texture of the image; so, it cannot make an absolute semantic grouping of the images. In addition, the tag clouds may not be so relevant because the tags taken by Flickr, which are attached to every image by its users, are very noisy and don't always correspond to the content of the image. For example, in many cases, uploaders are using as tag for the images that the upload, details about the digital camera, that they use. This type of information does not describe the content of the image.

To bypass the current TsoKaDo problems, a wide range of future work can be expected. At first visual words might offer better semantic recognition of objects inside the images. Furthermore, the obstacle of 'noisy' tags can be overran by extending the search of relative images and keywords to more reliable sources such as Wikipedia. Wordnet can be, also, replaced by EuroWordnet, the Multilanguage version of Wordnet with various European languages for a better sorting of the tags. In addition, in order to measure the effectiveness of the proposed scheme a detailed case study will be performed. An on-line early version of TsoKaDo is available at [22].

REFERENCES

[1] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, p. 5, 2008.

[2] A. Arampatzis, K. Zagoris, and S. Chatzichristofis, "Dynamic two-stage image retrieval from large multimodal databases," *Advances in Information Retrieval*, pp. 326–337, 2011.

[3] Arampatzis, A. and Zagoris, K. and Chatzichristofis, S., "Fusion vs. two-stage for multimodal retrieval," *Advances in Information Retrieval*, pp. 759–762, 2011.

[4] S. Chatzichristofis and Y. Boutalis, "Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval," in *Proceedings of the 6th international conference on Computer vision systems*. Springer-Verlag, 2008, pp. 312–322.

[5] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994, pp. 133–138.

[6] Lazaros T. Tsochatzidis, Athanasios C. Kapoutsis, Nikolaos I. Dourvas, S.A. Chatzichristofis, Yiannis S. Boutalis, "Query expansion based on visual image content," in *5th Panhellenic Scientific Conference for Undergraduate and Postgraduate Students in Computer Engineering, Informatics, related Technologies and Applications*, 2011.

[7] K. Barthel, "Improved image retrieval using automatic image sorting and semi-automatic generation of image semantics," in *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08. Ninth International Workshop on*. IEEE, 2008, pp. 227–230.

[8] R. van Leuken, L. Garcia, X. Olivares, and R. van Zwol, "Visual diversification of image search results," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 341–350.

[9] A. Popescu, P. Moellic, I. Kanellos, and R. Landais, "Lightweight web image reranking," in *Proceedings of the seventeen ACM international conference on Multimedia*. ACM, 2009, pp. 657–660.

[10] T. Berber and A. Alpkocak, "Deu at imageclefmed 2009: Evaluating re-ranking and integrated retrieval systems," *CLEFWorking Notes*, 2009.

[11] M. Lux, A. Pitman, and O. Marques, "Can global visual features improve tag recommendation for image annotation?" *Future Internet*, vol. 2, no. 3, pp. 341–362, 2010.

[12] Savvas A. Chatzichristofis, Yiannis S. Boutalis, *Compact Composite Descriptors for Content Based Image Retrieval: Basics, Concepts, Tools*. VDM Verlag Dr. Muller, 2011.

[13] A. Atsalakis and N. Papamarkos, "Color reduction and estimation of the number of dominant colors by using a self-growing and self-organized neural gas," *Engineering Applications of Artificial Intelligence*, vol. 19, no. 7, pp. 769–786, 2006.

[14] G. Varelas, E. Voutsakis, P. Raftopoulou, E. Petrakis, and E. Milios, "Semantic similarity methods in wordnet and their application to information retrieval on the web," in *Proceedings of the 7th annual ACM international workshop on Web information and data management*. ACM, 2005, pp. 10–16.

[15] G. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[16] A. Arampatzis, P. Efraimidis, and G. Drosatos, "Enhancing deniability against query-logs," *Advances in Information Retrieval*, pp. 117–128, 2011.

[17] [Online]. Available: http://www.google.com

[18] [Online]. Available: http://www.bing.com

[19] [Online]. Available: http://www.ask.com

[20] [Online]. Available: http://www.flickr.com

[21] [Online]. Available: http://www.google.com/landing/imagesorting/

[22] [Online]. Available: http://tsokado.nonrelevant.net/

 ISBN: 978-1-61208-177-9