I.J. Information Technology and Computer Science, **2015, 01, 41-47**
Published Online December 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2015.01.05

SANB-SEB Clustering: A Hybrid Ontology Based Image and Webpage Retrieval for Knowledge Extraction

Dr. **Anna Saro Vijendran**
Director, Department of MCA, SNR Sons College, Coimbatore- 641 006, India
Email: saroviji@rediffmail.com

Deepa .C
Assistant Professor, Department of Information Technology, SNR Sons College, Coimbatore- 641 006, India
deepa_pkd@rediffmail.com

***Abstract*—** Data mining is a hype-word and its major goal is to extract the information from the dataset and convert it into readable format. Web mining is one of the applications of data mining which helps to extract the web page. Personalized image was retrieved in existing systems by using tag-annotation-demand ranking for image retrieval (TAD) where image uploading, query searching, and page refreshing steps were taken place. In the proposed work, both the image and web page are retrieved by several techniques. Two major steps are followed in this work, where the primary step is server database upload. Herein, database for both image and content are stored using block acquiring page segmentation (BAPS). The subsequent step is to extract the image and content from the respective server database. The subsequent database is further applied into semantic annotation based clustering (SANB) (for image) and semantic based clustering (SEB) (for content). The experimental results show that the proposed approach accurately retrieves both the images and relevant pages.

***Index Terms*—**Web Structure Mining, Ontology, Semantic Annotation, Block Acquiring Page Segmentation (BAPS), Semantic Annotation Based Clustering (SANB), Semantic Based Clustering (SEB).

I. INTRODUCTION

Web mining is the function of data mining methods to find patterns from the web. In accordance with determination goals, web mining can be split into three distinct types, and they are web usage mining, web content mining and web structure mining. Web usage mining is the extraction of valuable information from the data being produced through web page visits, transactions, and so on. Web content mining is the extraction of useful information from web pages.

Web structure mining is the advancement of useful information from the links integrated into web documents. According to the web structural data type, web structure mining can be split into two:

- Extracting patterns from hyperlinks on the web
- Mining the document structure

Knowledge extraction is the formation of knowledge from structured and unstructured sources. The resultant knowledge wants to be in a machine-understandable and machine-interpretable format and must signify knowledge in a manner that assists inference. The main criteria are that the extracted product goes ahead of the formation of structured information or the conversion into a relational schema.

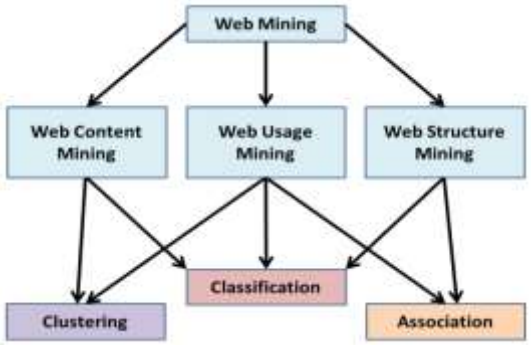

Fig. 1. Web mining model

Ontology is a prescribed design of a shared conceptualization. Ontologies are depicted as graphs, but not trees. Semantic annotation is the glue that binds ontologies into document spaces, via metadata. Instruction manual metadata manufacture cost is too high. Information extraction needs enlarging to target ontologies and a range of industrial document stores and the web.

The ontology-based information extraction (OBIE) is a subpart of information extraction, in which at least single ontology is used to show the process of information extraction from natural language text. At the semantic annotation of ordinary language text this one is improved with metadata, which should build the semantics of limited terms machine-readable. In this process, which is generally semi-automatic, knowledge is extracted, and then ontologies are established and thus knowledge is gained.

In the proposed work, the semantic annotation phase is answerable for inserting metadata to a collection of raw images and web pages. The work is divided into two major parts: uploading database on the server, and extracting images and web page based on semantic

Copyright © 2015 MECS

annotation. For extracting knowledge (i.e., image and web page), SANB clustering and SEB clustering are introduced. Thus, weight is evaluated based on semantic annotation, next it is ranked by weight and finally the image and web page are retrieved.

The rest of the paper is systematized as follows. Section II briefly overview the related happening in the semantic annotation and image retrieval techniques. Section III involves the detailed explanation about the proposed method. Section IV describes the implementation details. Section V summarizes with a brief conclusive remark and discussion on future works.

II. Related Work

This section deals with the works related to the semantic annotation based search and several image retrieval techniques.

Lux et al proposed semi-automatic tag suggestion which was used to maximize the quality and quantity of social annotation which further acts as a metadata. The overall process of the system was that the user assigned tag for the image, and then system found tag suggestions, next system re-ranked to suggest tags, and finally system presented top ranked tags to the user [1]. *Sieg et al* presented a structure for relative information access using ontology and established that the semantic knowledge entrenched in an ontology integrated with long-term user outlines. This can be used for personalized web search based on user's concern and preference [2]. *Lerman et al* presented two methods for personalizing results of image search on Flickr. Everyday behavior on Flickr had taken by users as a metadata for annotating their images. It can assist users by lowering the number of irrelevant results and it also significantly increase search precision [3].

La cascia et al proposed an integrated approach of latent semantic indexing for text and visual statistics for images. It was used to improve performance in conducting a content based search. It was unable to form a page zero query, because it was hard to depict the content of a particular test image with words [4]. *Zhang et al* suggested a novel personalized image retrieval scheme based on visual perception which discovers the region of interest whose attributes were extracted and analyzed. It was used to confine the semantic gap, also build user profiles [5]. *Liu et al* discussed a survey of content based image retrieval towards narrowing down the semantic gap. It needs the integration of relevant low-level attribute mining, effective knowledge of high-level semantics, friendly user interface, and efficient indexing tool [6].

Kodmelwar et al developed the social annotations and put forward a novel framework allowing for both the user and query relevance to discover to personalized image search. Ranking based multi-correlation tensor factorization model was proposed to accomplish annotation calculation and user-specific topic modeling was introduced to plan the query relevance and user preference [7]. *Bradshaw et al* proposed a probabilistic, multiple level approaches to the semantic labeling of images. Using multiple levels drastically increases the accuracy of the posterior probabilities. It can only distinguish between classes that are evenly separable [8].

Muller et al reviewed content based medical image retrieval systems. It helps to reduce the number of applications developed and also occupies more time on the essential tasks of unification and improvement of new methods and system development [9]. *Vogel et al* presented a computational image representation that diminishes the semantic gap between the image empathetic of the human and the computer. It was based on the categorization of local semantics thoughts. The major goal was to be ranked according to their similarity to the query [10]. *Klima et al* created the database of images: open source (DEIMOS), an open source for video image quality estimation and scientific intention. It also covers many distinct application fields [11].

Hsu et al discussed the integrated color-spatial technique which yields a better average accuracy for image recovery and was more permissive to noise in the image and was adaptable in dealing with the same objects of dissimilar colors [12]. *Dai et al* proposed a novel representation-visual group which was to develop the recovery precision by combining the analytically related attributes. It can assure the accuracy of image identical with high competency. Bag-of-visual words (BoW) had been used in large scale image recovery [13]. *Su et al* presented a novel image retrieval system named intelligent semantic image explorer (iSMIER) using image annotation, concept matching and fuzzy ranking techniques. It provides better performance by cross-media retrieval model for a visual domain to the semantic domain [14].

Hyvonen et al discussed that semantic web ontology was to retrieve images from a database and also help to the user in evaluating the information need, the query, and the answers. The images were annotated and expedite focused image retrieval [15]. *Singhal et al* proposed a migrating crawlers approach, in which migrants after moving to the web servers downloads the .TVI (table of variable information) file only for sustaining the freshness of search engine depository [16].

Rui et al presented a human-computer interaction approach to CBIR based on relevance feedback. This allows the user to present an improper primary query and constantly refine information need via relevance feedback. It reduces the user's effort [17]. *Tang et al* introduced an integrated graph-based semi-supervised learning framework to employ the multiple-instance or single-instance representation concurrently for image annotation. It provides better performance and reduce the computational complexity [18].

Muller et al showed the various relevance feedback strategies on the query result. Negative feedback can be prevented by using a Rocchio's technique of individualized weighting positive and negative attributes. It provides better outcomes for negative relevance feedback [19]. *Chen et al* proposed a novel web media semantic concept retrieval framework which includes both tag removal and model fusion for content-based and

Copyright © 2015 MECS

tag-based model which helps to remove the irrelevant tag for the query [20].

III. Proposed Work

This section presents a detailed description about the extraction of both images and the contents of the web page (simply, Knowledge extraction) by using clustering algorithms. To overcome the existing technique (TAD), the two major components are introduced in the proposed work and they are explained as follows:

- Server Database (SDB) Upload
- Semantic based knowledge extraction

A. Server Database (SDB) Upload

For uploading a server database, this proposed system follows two datasets for images and contents which are taken from the web pages. Fig.3 presents the flow of uploading a server database. Thus, it developed SBD for both the images and the contents.

1) Image SDB

Image are uploaded by collecting several images from the user and stored in a database with appropriate tags and annotation. The data will be considered as images with the basic set of formats including JPEG, GIF, BMP, and PNG. Fig.2 depicts a sample set of input images. These images are further provided with a user preferred tags and annotations. Annotations are given descriptively to each of those images and then it is stored in an SDB.

Fig. 2. Sample sets of input images

2) Web Page SDB

In order to examine a web page for content extraction, efficient page is first decided through HTML parsers that adapt the HTML and generates a DOM (Document Object Model) tree representation of the web page. In structure analysis, the tags presents in a web page are in HTML. It is parsed and is requires recognizing tags accessible in page blocks, child tags and in addition tags over inner block.

In tag tree parsing, a DOM tree for getting analysis structure is constructed. The input of DOM tree is xml files have to perform the file conversion, and the HTML file is converted into XML file. After attaining the structure of the web page, the unwanted tags are detached. For content extraction, block acquiring page segmentation (BAPS) is introduced and is expressed as follows.

Block Acquiring Page Segmentation (BAPS) Algorithm

1. *Algorithm segmentation (node)*
2. *while (node!=null)*
3. *if (node is acquiring node & no child)*
    a. *remove node*
4. *else if (node is acquiring node & one child)*
    a. *if (node is not gathering node)*
    b. *node to archive*
    c. *set ranking as b*
5. *else if (node is acquiring node & more child)*
    i. *if (node is gathering node)*
    ii. *not to archive*
    iii. *set ranking as a*
6. *else*
    a. *remove node*

Initially, unnecessary tags in the web page are detached and then the subsequent information is detached. In this proposal, tag such as HTML, HEAD, and BODY are used to extract the content.

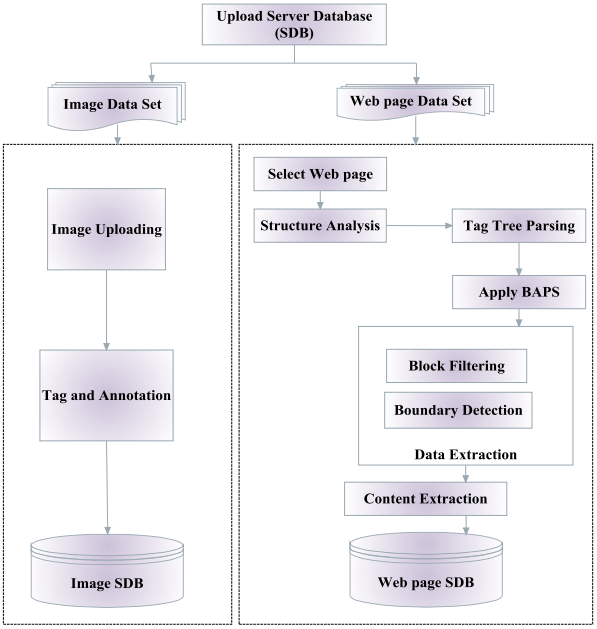

Fig. 3. Flow of server database upload

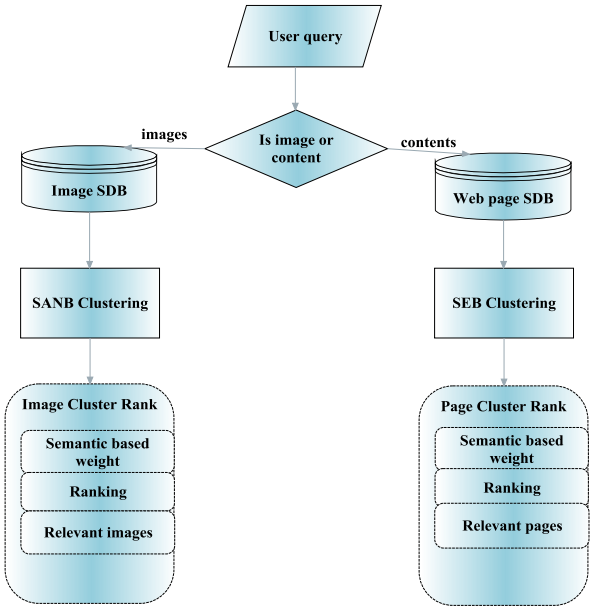

Fig. 4. Flow of semantic based knowledge extraction

Data extraction process is used to eliminate the boundary for each and every noisy block. Finally, SDB for both images and the contents are uploaded.

B. Semantic based Knowledge Extraction

The extraction of both images and contents are collectively termed as knowledge extraction. Fig.4 presents the flow of this process.

1) Semantic Annotation based (SANB) Clustering

The user quotes query, if that corresponding query is an image then it searches in the image server database. Further, it follows semantic annotation based clustering algorithm:

Semantic Annotation based Clustering (SANB) Algorithm

Input : Image description (di), annotation (ai)
Output : Cluster
1) //Similarity
2) for each i=1 to n // Number of images
a) get di,ai from i^{th} image
b) for each j=i+1 to n
i) get dj,aj from j^{th} image
(a) s1=semsim(di,dj) // using wordnet
(b) s2=semsim(ai,aj)
(c) sim=(s1+s2)/2
(d) sim[i][j]=sim
c) end
d) end
3) //Clustering
4) c:=1
5) for each i=1 to n
i) add img_i into cluster c
ii) for each j=j+1 to n
(1) if sim[i][j]>0.5
(2) add img_j into cluster c
(3) end if
iii) end
b) c++
6) end
7) return c

In SANB clustering algorithm, image description and annotation are taken as an input from *n* number of images. The selected image is compared with other images for similarity using wordnet. Wordnet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are collected into sets of synonyms and are termed as synsets, each expressing a different concept. It can be freely used and the main intention is twofold: to produce an association of dictionary and thesaurus that is more instinctively usable, and to sustain an automatic text analysis and artificial intelligence applications.

Grammatical similarity is checked for image description and is saved and another set of group consists of annotation. Similarity is taken by averaging the description and annotation and is further equivalent to array set. Each cluster has only one image. Therefore, for clustering the similarity is checked with the threshold value of 0.5. Finally, the image is placed in the cluster which acts as an output for above algorithm.

2) Semantic based (SEB) Clustering

The user quotes query, if that corresponding query is the content from web page then it searches in the web server database. Further, it follows a semantic based clustering algorithm:

Semantic based Clustering (SEB) Algorithm

Input : Web Page Contents
Output : Cluster
1. //Similarity
2. for each i=1 to n
3. get content ct_i
4. for each j=j+1
a. get content ct_j
b. s1=semsim (ct_i,ct_j)
5. end
6. end
7. //Clustering
8. c:=1
9. for each i=1 to n
10. add di to c
11. for each j=j+1
12. if(sim[i][j]>0.5) then
a. add dj to c
13. end if
14. end
15. c++
16. end

In SEB clustering algorithm, web page contents are taken as an input from *n* number of images. Selected web page content is compared with other contents for similarity. Subsequent contents are checked by excluding the clustered content.

Linguistic similarity is checked for web page contents and is saved as an array element. Therefore, for clustering the similarity is checked with the threshold value of 0.5. Finally, the web page content is placed in clusters and that content won't be repeated for further cluster.

3) Cluster Ranking

Clustered image or web page content is currently ranked by weight and it will provide the relevant images and web page contents as an output.

(a) Image Cluster Rank

The clustered image is ranked by its weight and the top number of results is returned as an optimized result. Thus, the relevant images are retrieved. The following algorithm described as follows:

Image Cluster Rank

Input : User query image, des, cluster, k (number of results)
Output : Relevant images
1. //Cluster weight
a. get query description q_d
b. for each i= 1 to c // each cluster
c. for each j= 1 to n // number of image in cluster
i. get des di from image j

ii. si= si+semsim (q_d,di)

d. end

e. end

2. //Ranking

a) Rank the cluster using cluster weight

b) Return the top k results

(b) Web Page Cluster Rank

The clustered web page content is ranked by its weight and the top number of results is returned as an optimized result. Thus, the relevant web page contents are retrieved. The following algorithm described as follows:

Web page Cluster Rank

Input : User query (q), cluster, k (number of results)

Output : Relevant web pages

1. //Web page weight

a. for each i= 1 to c // each cluster

b. for each j= 1 to n // number of web page content in cluster

i. si= si+semsim (q,dj)

c. end

d. end

2. //Ranking

a) Rank the cluster using weight

b) Return the top k results

Finally, the knowledge is extracted and its performance is analyzed in the next section with the existing work.

IV. Performance Analysis

This section presents the performance analysis of the existing work (TAD) and proposed work (SANB).

A. Computational Time

Computation time is the quantity of time for which a central processing unit (CPU) was used for developing instructions of a computer program or operating system.

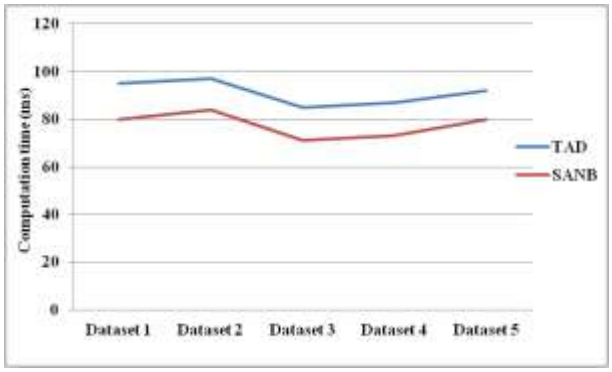

Fig. 5. Dataset vs. Computational time

Fig.5 shows the comparison graph between the existing works (TAD) with the proposed work (SANB). Here, for each dataset computational time is slightly increased and decreased for the proposed system which has lesser time than existing systems.

B. Precision

Precision is termed as the total number of correctly identified data records to the total number of data records identified.

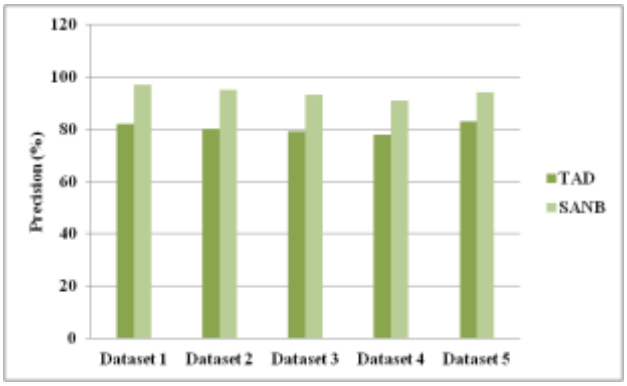

Fig.6. Dataset vs. Precision

Fig.6 presents the comparison between the TAD and SANB. The precision is higher for SANB comparing to TAD (existing system).

C. Recall

Recall is termed as an algorithm revisited most of the relevant results. Herein, the proposed system returned the relevant images and web page contents.

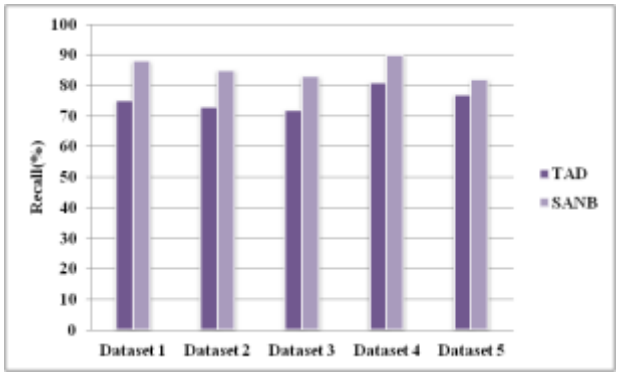

Fig. 7. Dataset vs. Recall

Fig.7 presents the comparison of TAD and SANB. In which the proposed system provides the best result with a higher percentage of recall.

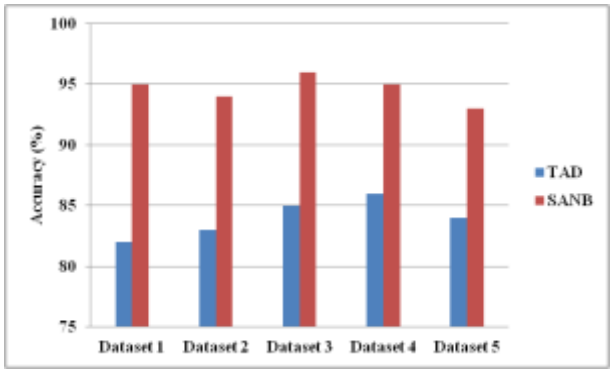

Fig. 8. Dataset vs. Accuracy

Copyright © 2015 MECS

D. Accuracy

Accuracy is the sum of correctly identified and incorrectly identified. Herein, the proposed system presents higher accuracy than existing systems.

Fig.8 shows the comparison of TAD and SANB where SANB presents the higher accuracy for all the datasets in the system.

E. Memory Usage

Memory usage is the memory which is used by the program when it is implemented.

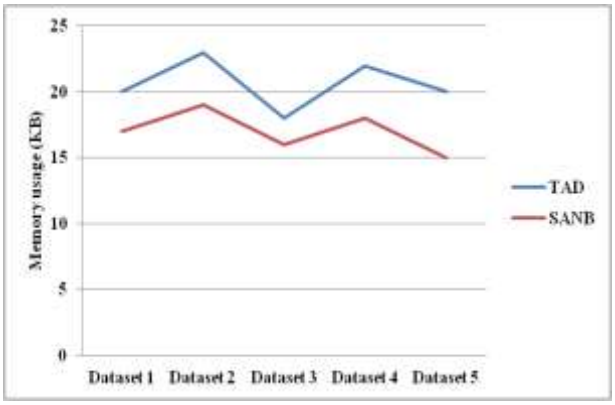

Fig. 9. Dataset vs. Memory usage

Fig.9 shows the comparison of memory usage for TAD and SANB. Herein, the proposed system presents the lower usage of memory for each dataset.

V. Conclusion

An efficient way for retrieving the relevant images and web page contents (simply, knowledge extraction), semantic annotation based clustering is proposed. Two major steps are followed in this proposed work. The former step is to upload the server database and the latter step is to extract the knowledge using semantic annotation based (SANB) clustering for image data set and semantic based (SEB) clustering for web page content. The experimental results present the retrieval results for relevant images and web page contents.

References

[1] M. Lux, *et al.*, "Using visual features to improve tag suggestions in image sharing sites," *Proceedings of knowledge acquisition from the social web, Graz, Austria*, 2008.

[2] A. Sieg, *et al.*, "Learning Ontology-Based User Profiles: A Semantic Approach to Personalized Web Search," *IEEE Intelligent Informatics Bulletin*, vol. 8, no. 1, pp. 7-18, 2007.

[3] K. Lerman, *et al.*, "Personalizing image search results on flickr," *Intelligent Information Personalization*, 2007.

[4] M. La Cascia, *et al.*, "Combining textual and visual cues for content-based image retrieval on the world wide web," in *Content-Based Access of Image and Video Libraries, 1998. Proceedings. IEEE Workshop on*, 1998, pp. 24-28.

[5] J. Zhang, *et al.*, "A personalized image retrieval based on visual perception," *Journal of Electronics (China)*, vol. 25, no. 1, pp. 129-133, 2008.

[6] Y. Liu, *et al.*, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262-282, 2007.

[7] M. Kodmelwar and P. Futane, "An Optimization Technique for Image Search in Social Sharing Websites."

[8] B. Bradshaw, "Semantic based image retrieval: a probabilistic approach," in *Proceedings of the eighth ACM international conference on Multimedia*, 2000, pp. 167-176.

[9] H. Müller, *et al.*, "A review of content-based image retrieval systems in medical applications—clinical benefits and future directions," *International journal of medical informatics*, vol. 73, no. 1, pp. 1-23, 2004.

[10] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 133-157, 2007.

[11] M. Klíma, *et al.*, "DEIMOS–an open source image database," *Radioengineering*, vol. 20, no. 4, pp. 1016-1023, 2011.

[12] W. Hsu, *et al.*, "An integrated color-spatial approach to content-based image retrieval," in *Proceedings of the third ACM international conference on Multimedia*, 1995, pp. 305-313.

[13] L. Dai, *et al.*, "Large scale image retrieval with visual groups," in *Proc. IEEE ICIP*, 2013.

[14] J.-H. Su, *et al.*, "Multi-modal image retrieval by integrating web image annotation, concept matching and fuzzy ranking techniques," *International Journal of Fuzzy Systems*, vol. 12, no. 2, pp. 136-149, 2010.

[15] E. Hyvönen, *et al.*, "Ontology-Based Image Retrieval," in *WWW (Posters)*, 2003.

[16] N. Singhal, *et al.*, "Reducing Network Traffic and Managing Volatile Web Contents Using Migrating Crawlers with Table of Variable Information," *World Applied Sciences Journal*, vol. 19, no. 5, pp. 666-673, 2012.

[17] Y. Rui, *et al.*, "Relevance feedback: a power tool for interactive content-based image retrieval," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 8, no. 5, pp. 644-655, 1998.

[18] J. Tang, *et al.*, "Image annotation by graph-based inference with integrated multiple/single instance representations," *Multimedia, IEEE Transactions on*, vol. 12, no. 2, pp. 131-141, 2010.

[19] H. Muller, *et al.*, "Strategies for positive and negative relevance feedback in image retrieval," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, 2000, pp. 1043-1046.

[20] C. Chen, *et al.*, "Web media semantic concept retrieval via tag removal and model fusion," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 4, no. 4, p. 61, 2013.

Authors' Profiles

Dr. Anna SaroVijendran received the Ph.D. degree in Computer Science from Mother Teresa Women's University, Tamilnadu, India, in 2009. She has 24 years of experience in teaching. She is currently working as the Director, Dept of MCA in SNR Sons College, Coimbatore, Tamilnadu, India. She has presented and published many papers in International and National conferences. She has authored and co-authored more than 50 refereed papers. She has

also acted as chair person in many National and International Conferences. Her professional interests are Image Processing, Image Fusion, Data Mining and Artificial Neural Networks.

Ms. C Deepa received her Masters Degree in Computer Applications from Bharathiar University, Coimbatore, Tamil Nadu, India in the year 2000 and received M.Phil degree in Computer Science from Bharathiar University, Coimbatore, Tamil Nadu, India in the year 2004. Currently she is working as Assistant Professor, Department of Information Technology, SNR SONS College, Coimbatore, TamilNadu . She has more than 12 years of experience in teaching. She is doing Ph.D in Computer science at SNR Sons College, Coimbatore, under the supervision of Dr. Anna Saro Vijendran, Director & Head, Department of Computer Applications, S.N.R Sons College. Her research interests include Data Mining ,Web Mining and OOPS

How to cite this paper: Anna Saro Vijendran, Deepa .C,"SANB-SEB Clustering: A Hybrid Ontology Based Image and Webpage Retrieval for Knowledge Extraction", International Journal of Information Technology and Computer Science(IJITCS), vol.7, no.1, pp.41-47, 2015. DOI: 10.5815/ijitcs.2015.01.05

Copyright © 2015 MECS