Automatic Image Arranging System Based on the Semantic Analysis of the Surrounding Text Retrieved from the World Wide Web

Akash, Ashish Chaudhary, Namrita Pandita

Abstract—Image retrieval has become an important part of the day to day search for any internet user. Users generally type in certain keywords for image search, as the search engines are unaware of the context of the search they end up displaying some random images. This paper presents a system through which the user can specify the context of his search using an ontology file, thereby enabling the search engine to come up with much more relevant search results. Using this technique the search engine not only searches the images on web but also scans for keywords, provided in the ontology file, in the text surrounding the images. Thereafter it arranges the result according to relevance of images. This paper presents examples of the usage and performance of such a system, specifically focusing on image searches related to domain "football". The results of the idea proposed in this paper filters the image search so as to display the images which are more related to the domain "football" before the ones less related. In addition to analyzing the efficiency of this technique this paper also presents some limitations and proposes few future developments which could improve its efficiency furthermore.

Index Terms— Ontology, Text based image search, RDF, OWL, semantic analysis, python, image search engine

1 Introduction

When someone on the internet, search for the images on image search engines like Yahoo , Bing or Google, then these search engines don't know what exactly the user is looking for. For example, suppose someone is a fan of the Manchester United Football Club and likes to search for the images related to the club. So when he/she searches for the images related to that by entering the keyword Manchester, then yahoo will show results related to both the club and the Manchester city, the place. Thus, user will see some images randomly arranged for both the Manchester Football Club and the Manchester city. The paper presents a solution to remove this problem of random arrangement by performing semantic analysis on some information related to the image and then rearrange them for the user, with more relevant images coming first in the new arrangement.

In today's internet age an image without some relevant information about that image is of no use. So, the idea presented in this paper is to use this information associated with the image for the arrangement and determining that how relevant this image can be for the user.

- *Akash, B.Tech, Computer Science and Engineering Department, Motilal Nehru National Institute of Technology, Allahabad, India, PH-+91-9967494869. E-mail: mail2akash.lucky@gmail.com*
- *Ashish Chaudhary, B.Tech, Computer Science and Engineering Department, Motilal Nehru National Institute of Technology, Allahabad, India, PH-+91-9945170576. E-mail: ashishchaudhary9211@gmail.com*
- *Namrita Pandita, B.Tech, Electronics and Communication Engineering Department, Motilal Nehru National Institute of Technology, Allahabad, India, E-mail: namrita.pandita@gmail.com*

This paper use the surrounding text associated with the image and do semantic analysis on the text and come up with a number which determined the relevancy of the image and using this number to arrange the images. Thereby displaying images arranged as per the preference for the user.

Fig. 1 Image from yahoo image seach engine

Now there are various image analysis tools and methods available to do the semantic analysis on the images [1] like Content Based Image Retrieval (CBIR) [2], text content based image retrieval, meta-search engine implementation etc. These

International Journal of Scientific & Engineering Research Volume 4, Issue 11, November-2013
ISSN 2229-5518

methods are all very different in the approach they take to image ordering. This paper focuses on the method of text content based image retrieval.

Most images on the internet are not found alone, as an image without information is of very less use. They are embedded in an html page along with the text related to the image. The system, known as Semantic Order Image Arranging System (SOIAS) uses the surrounding text to understand the relevance of images. For example, the image in Fig 1 is taken from the yahoo image search engine.

As you can see that just by looking at the image in Fig 1 no one can tell much about the image, but when the text associated with the image is considered which is "Valley of flowers national Park is an Indian National Park, nestled in West Himalayas and is located in Uttarakhand state", anyone can tell accurately that the image is located where and what exactly the image is talking about. The idea is to use this information to accurately tell what the picture is about.

This paper describes a system which is very flexible and alows the users to change the ontology used for determining the semantic relevance for the images. Thus according to his or her need can change the relevancy for the images. Hence, the system propsed is a general system which can be tailored according to the need and use of the user by changing the ontology file accordingly. This paper uses the football ontology and thus the results presented in the paper are be related to the images associated with the game "football".

2 The Working of the Semantic Order Image Arranging System (SOIAS)

2.1 Basic Flow of the System

Search term is given to the Module 1 of the SOIAS system which uses some internet image search engine to download the images and the surrounding text associated with the image. It then parses the text and stores in database.

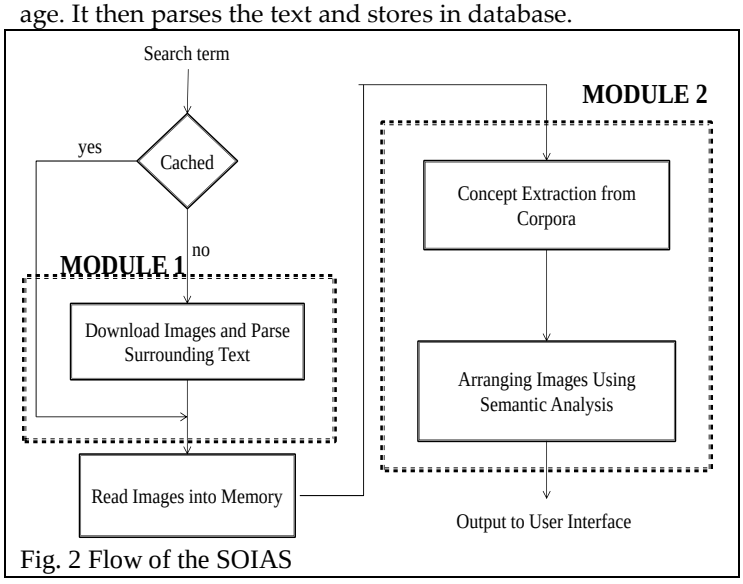

Fig. 2 Flow of the SOIAS

Module 2 then extracts the concepts from the corpora and then arranges images using the semantic analysis. Then the arranged images are shown to the user in a proper interface. The basic flow of the whole system is shown in Fig 2.

2.2 Module 1

The flow of the Module 1 is shown in the Fig 3. The various steps in the Module 1 are:

1. In the first step, the web search URL is made. There are some search constraints in the system which user can use to modify the search type the user wants to do. These are number of the images to be downloaded and image size

2. In next step, the system using the URL created in first step, does the web search. From the result page obtained, the image URL's are retrieved. Then the images and their HTML pages are downloaded from the URL retrieved using the python's mechanize library [3]. The downloading of the images in done in parallel threads to improve the downloading speed. The threading is implemented using the producer-consumer threads. The producer thread creates a new thread for every new image and put them in a queue. Each thread downloads the image and its corresponding HTML file. The consumer thread takes the image from the corresponding the producer threads and then create an entry in the database serially for maintaining the consistency.

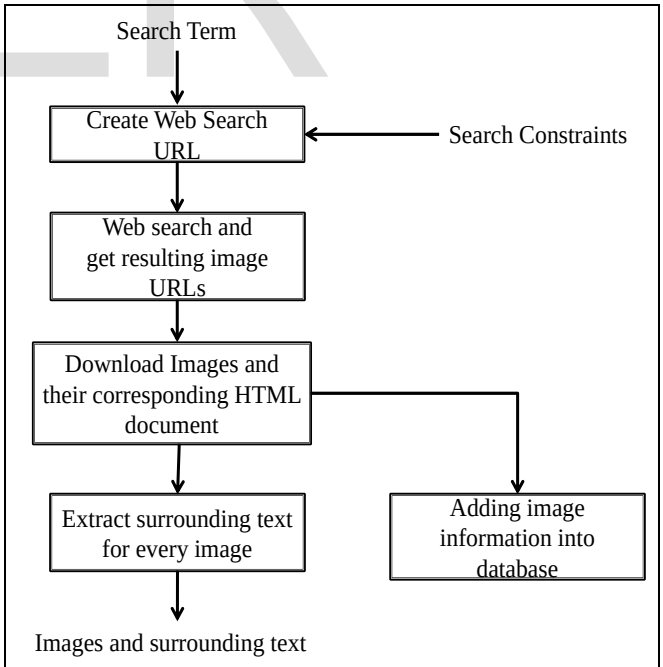

Fig. 3 Flow of Module 1

3. Surrounding text is then extracted from the image's HTML page using the python's BeautifulSoup [4] library and is saved as files with .corpora file extensions in the system along with the corresponding images.

International Journal of Scientific & Engineering Research Volume 4, Issue 11, November-2013
ISSN 2229-5518

4. For every new image downloaded, a new entry is made in the database along with the information like the image URL, website URL, search term and the image's name in the hard disk.

2.3 Module 2

Module 2 consists of following steps as shown in the Fig 4 and described below:

1. Takes the corpora from the Module 1 and then removes the irrelevant stop words using a predefined dictionary of stop words. Stop-words are common words that carry less importance than keywords. Usually search engines remove these words from Keyword phrase.
2. In the next step the stemmer algorithm is used to obtain the morphological root of the words in the corpora. Stemmer algorithm [5] is applied using the python's whoosh library [6]. For Example, Chatter and Running words are converted to chat and run.
3. Concepts are extracted from the corpora by comparing key words against the given standard ontology concepts in the OWL format [7]. The number of the concepts extracted is stored in the database corresponding to the image.
4. Based on the number of the concepts extracted the images are arranged and are shown in an interface.The results with the highest corpora frequency are shown to the user first in the arrangement.

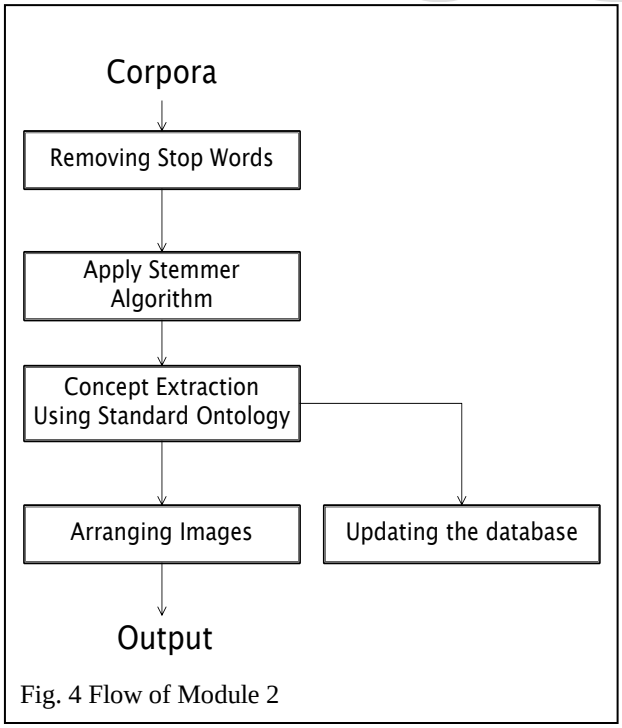

Fig. 4 Flow of Module 2

2.4 Database Representation of the SOIAS

A common RDF [8] file is created for storing the information. . A general entry of image is made into the database as shown below:

<rdf: Description rdf: about = "URI of IMAGE">
<image:image_name>… </image:image_name>
<image:URL_webpage>… </image:URL_webpage>
<image:URL_image>… </image:URL_image>
<image:Search_Term>… </image:Search_Term>
<image:keyword>… </image:keyword>
<image:concept_name>… </image:concept_name>
</rdf:Description>

2.5 Ontology Used in the System

Ontology [9] is a formal explicit description of concepts in a domain of discourse (classes (sometimes called concepts)), properties of the each concept and attributes of the concept and restrictions on the slots. Classes are the focus of most ontology. Classes describe concepts in the domain. For example, a class in the system's football ontology is Team. The ontology used in this paper is top-down ontology. For SOIAS system all the concepts are given equal importance. There is no such implementation in which the more importance is given to the concepts in subclasses.

3 Testing and Results

The system used has the ontology associated with the football. So the results must show the images whose corpora are more related to the terms described in the ontology. The system puts a constraint on the images that there should be a minimum count of 10 concepts to be associated with an image. If no image is able to fulfill the above criteria it is discarded.

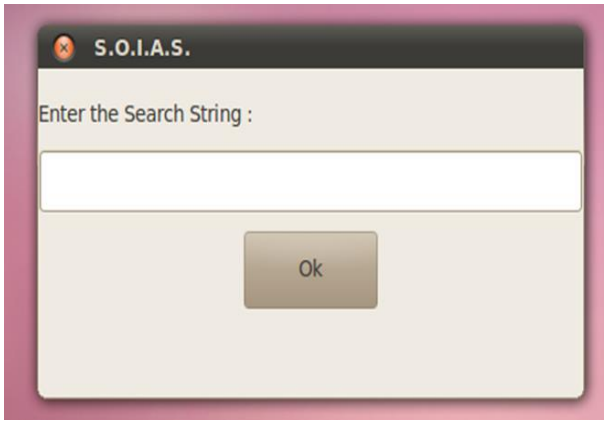

Fig. 5 GUI for entering search term

The SOIAS provides an interface shown in the Fig 5 to let the user enter terms to be used to search images.
The GUI is provided using the wxPython [10] library. Once the keyword is entered the images which are considered more relevant to the user are shown in the interface.

International Journal of Scientific & Engineering Research Volume 4, Issue 11, November-2013
ISSN 2229-5518

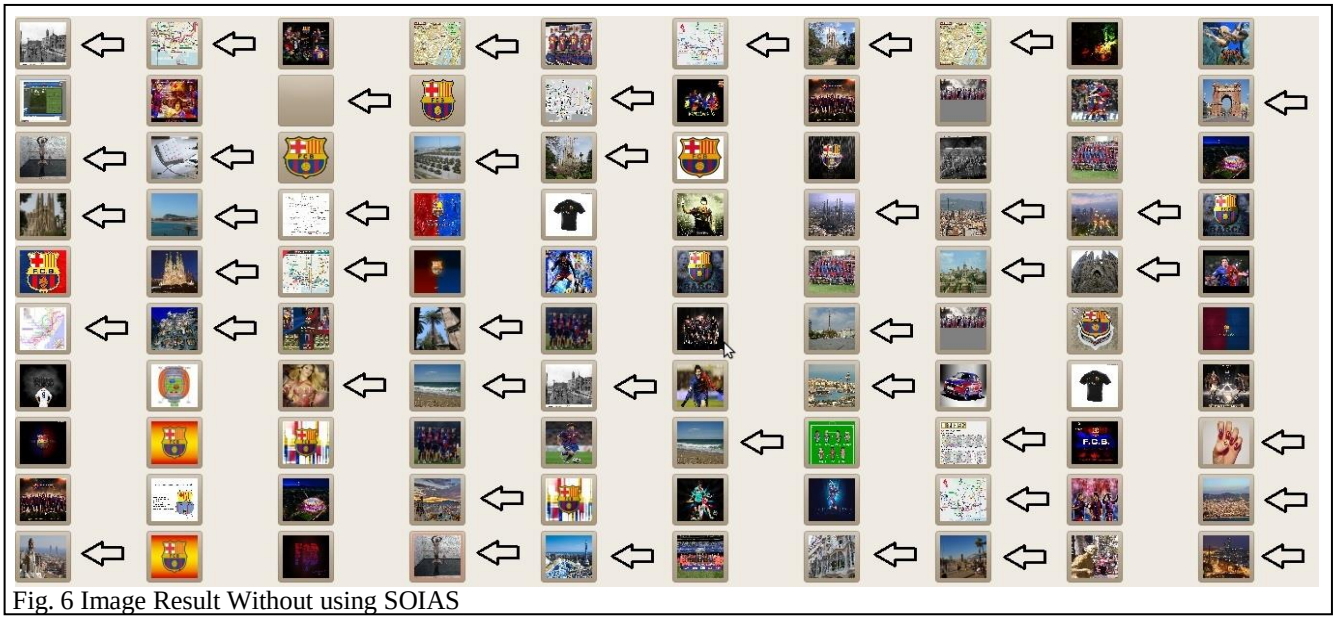

Fig. 6 Image Result Without using SOIAS

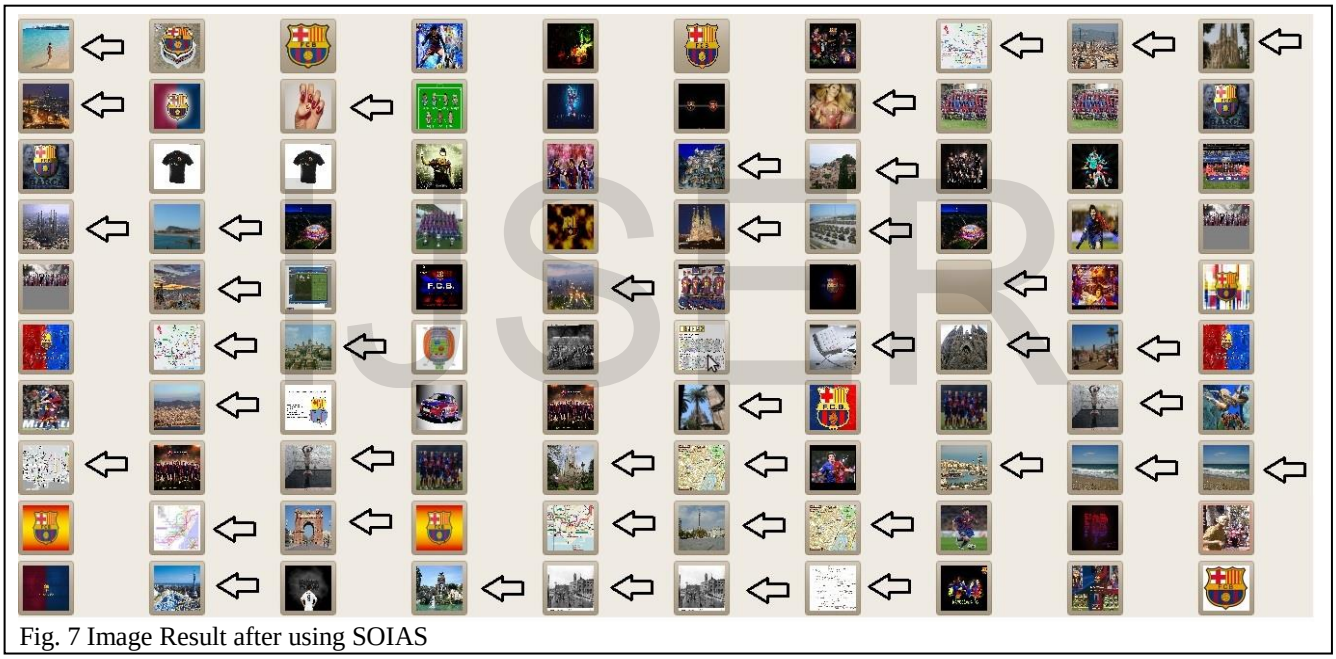

Fig. 7 Image Result after using SOIAS

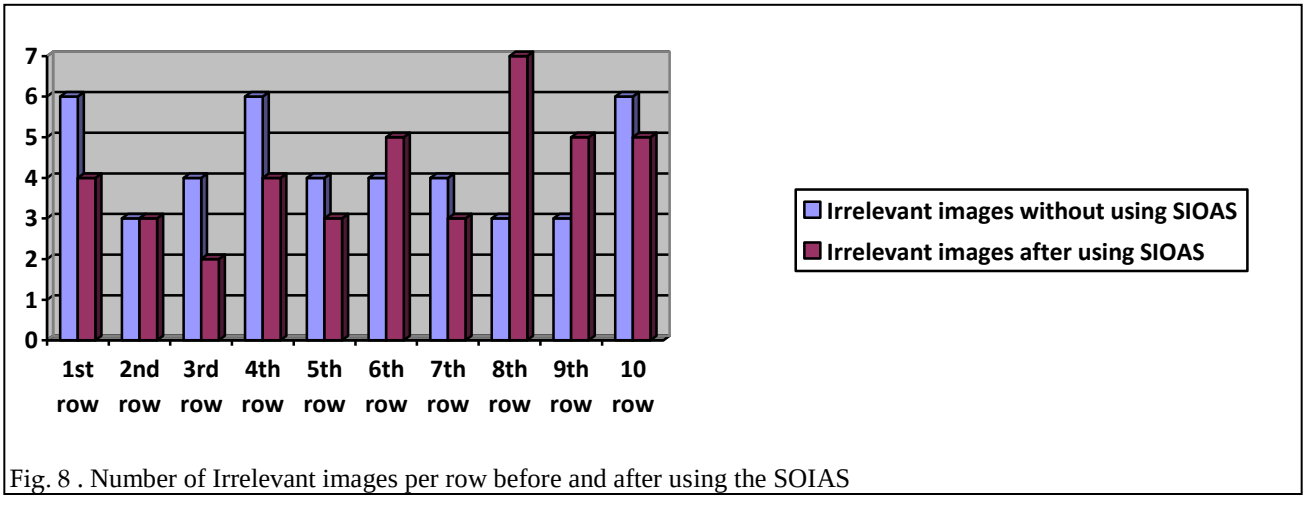

Fig. 8 . Number of Irrelevant images per row before and after using the SOIAS

International Journal of Scientific & Engineering Research Volume 4, Issue 11, November-2013
ISSN 2229-5518

Fig. 9 Image Result Without using SOIAS

Fig. 10 Image Result using SOIAS

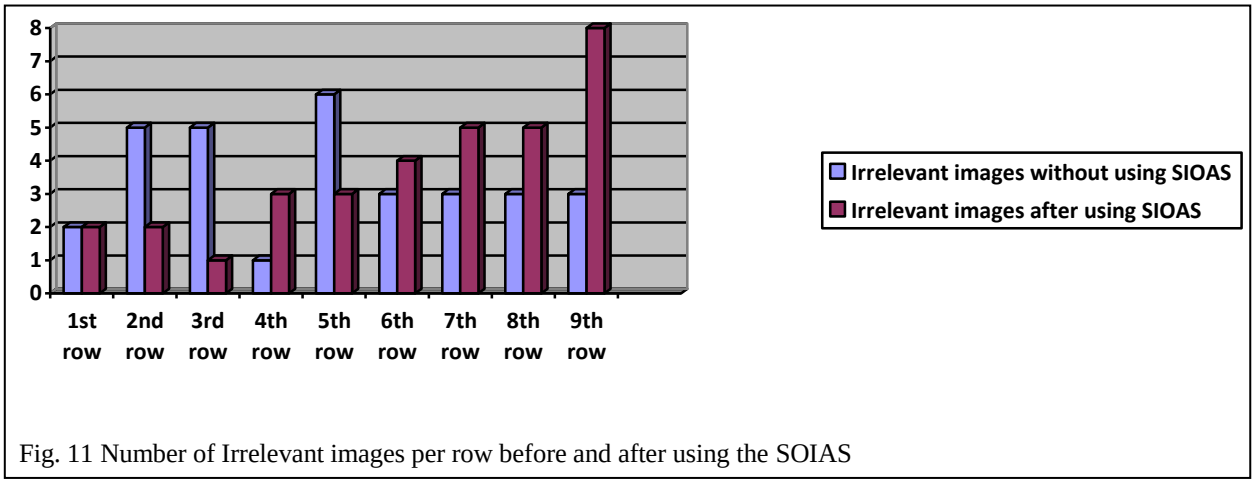

Fig. 11 Number of Irrelevant images per row before and after using the SOIAS

International Journal of Scientific & Engineering Research Volume 4, Issue 11, November-2013
ISSN 2229-5518

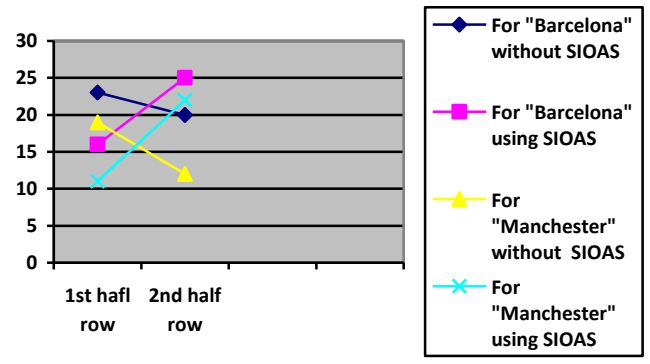

Fig. 12 Number of Irrelevant images in 1st half and 2nd half of the rows

If the user enters the term "Barcelona" as the search term, then the images as shown by the yahoo without using the semantic analysis tool is given in Fig 6 and arrangement of the images after applying the semantic analysis on the same set of images is shown in the Fig 7. The number of irrelevant images per row before and after applying semantic analysis is shown in Fig 8.

The symbol "←" in these images shows the image which is irrelevant to the user. The system tries to rearrange the images with the images related to the Barcelona Football Club coming first and rest of the images coming in the end.

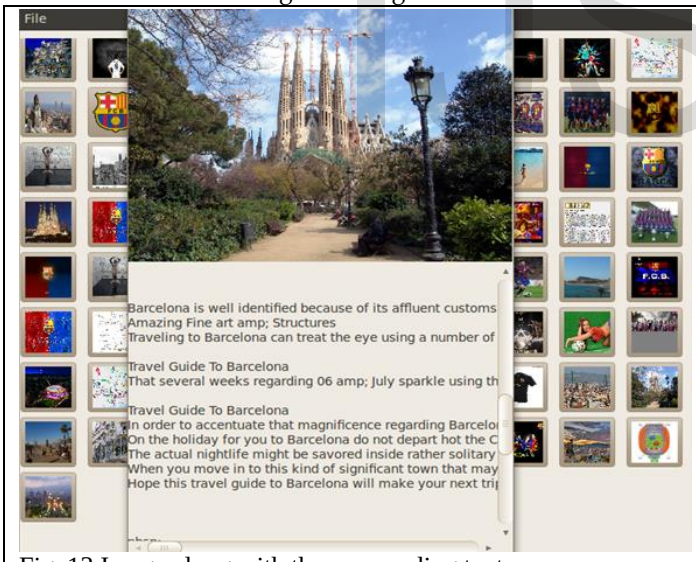

Fig. 13 Image along with the surrounding text

For the first search term "Barcelona", as shown by Fig 12, the number of irrelevant images in the first five rows, which is 23, is very high as compared to the irrelevant images in the later 5 rows which is 20. This number increases from last 5 rows to first 5 rows. So, if the user searchs for an image on any web browser, then the number of irrelevant images shown is very high in the initial rows, which is not what the user wants. This is where the SOIAS comes in the picture. If the user has specified its intent using the football ontology, then the results are quite different. The number of irrelevant images in the first 5 rows decreases to 16 and in the last 5 rows increases to 25. Hence, the system is able to move the more relevant images, which in this case is associated with the football, to the first 5 rows and less relevant to later of the rows.

As second search term "Manchester", the number of irrelevant images before using the SOIAS is shown in Fig 9 and after using SIAOS is shown in Fig 10. The number of irrelevant images per row before and after using the SIOAS is shown in Fig 11. The number of irrelevant images in the first 5 rows is 19 and 12 in the rest 4 rows, before using the SOIAS as shown in the Fig 12. But, this changes when SOIAS is used to arrange the images. The number of irrelevant images decreases to 11 in the top row and increases to 22 in last 4 rows.

Hence, the SOIAS has again proved that it is able to arrange images such that more relevant images are shown to the user in the initial rows of arrangement if the user is able to express his intent using the ontology.

Additional features which are also provided by the system are:

1. Keywords suggestion to the user: The keywords are suggested by comparing the frequencey of the words in the surrounding text and presenting words with the highest frequency.
2. Equality of the images: Equality of the images based on the number of concepts which are same in both the images. It is done by caculating the similar number of the concepts and the total number of the concepts and calculating the ratio
   RelativeEquality = commonConcepts / totalConcepts
3. Ontology Modification: In the system, user can specify the system to use his own ontology. Thus, the system will give relevance to the concepts specified in the user ontology. Thus, the system can easily be chaged to change the domain of the system.
4. Image with surrounding text: The system provides the user with the ability to see the enlarged version of the image, which is then fetched from the database, along with the surrounding text associated with the image. Fig 13 shows the image along with the surrounding text associated.

4 Conclusion

SOIAS system is able to arrange the images with images whose surrounding text is related to the football to come first and the other irrelevant images come later in the arrangement. System discards the images which have no corpora or surrounding texts associated with them or have concepts less than 10 associated with them. As shown in the results provided previously in which more relevant images move up in the arrangement and less relevant move later in the arrangement. There are some limitations observed in the system on which the future work is going to be focused on following points:

1. Firstly, if the corpora associated with the image is not relevant enough then the images even though they are

International Journal of Scientific & Engineering Research Volume 4, Issue 11, November-2013
ISSN 2229-5518

relevant to the user may end up coming later in the arrangement.

2. System is not able to determine and eliminate the images which are same.
3. The ontology used is not properly utilized with all the concepts been given the same importance. Thus, more generic concepts of football are given same importance as the specific concepts which need to be changed.
4. Some relevant images are discarded because of having less than the minimum number of concepts needed to be considered as relevant by the SOIAS.

Thus, there are some limitations on which the system needs to be improved upon. With more concentrated efforts in this direction, the system can be made to work in more efficient manner.

ACKNOWLEDGMENT

We would like to express our gratitude towards Professor B.D.Chaudhary of Computer Science and Engineering Department, MNNIT Allahabad, under whose esteemed guidance we were able to successfully implement this project. He continually and persuasively pushed us towards research. We would also like to thank Computer Science and Engineering Department of MNNIT, Allahabad, for providing us the resources and access to the computer labs required for the project.

REFERENCES

[1] Bo Luo, Xiaogang Wang, and Xiaoou Tang, "A World Wide Web Based Image Search Engine Using Text and Image Content Features", *SPIE* Vol. 5018 (2003)
[2] Content-based image retrieval (CBIR), http://en.wikipedia.org/wiki/Content-based_image_retrieval
[3] Mechanize library, https://pypi.python.org/pypi/mechanize
[4] Beautiful Soup Library, http://www.crummy.com/software/BeautifulSoup/
[5] Julie Beth Lovins, "Development of a Stemming Algorithm", *Mechanical Translation and Computational Linguistics*, vol.11, nos.1 and 2, March and June 1968
[6] Whoosh library, https://pypi.python.org/pypi/Whoosh/
[7] W3C Recommendation OWL Web Ontology Language Overview: http://www.w3.org/TR/owl-features/
[8] W3C Recommendation Resource Description Frame-work(RDF) Schema Specification: http://www.w3.org/TR/1998/WD-rdf-schema/
[9] Natalya F. Noy and Deborah L. McGuinness," Ontology Development 101: A Guide to Creating Your First Ontology", http://liris.cnrs.fr/alain.mille/enseignements/Ecole_Centrale/What%20is%20an%20ontology%20and%20why%20we%20need%20it.htm
[10] wxPython Library, http://www.wxpython.org/