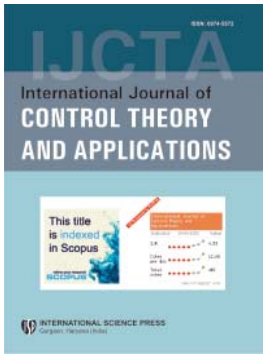

International Journal of Control Theory and Applications

ISSN : 0974–5572

Volume 9 • Number 40 • 2016

Adapting Images and URLs for Semantic Web Search

Borra Sai Sandeep[a] M. Sangeetha[b] Thallada Akhil[a] and Vudutha Akhil[a]

[a]*B.Tech. Dept of IT, SRM University, Chennai, Tamil Nadu, India*

[b]*Asst. Prof. Dept of IT, SRM University, Chennai, Tamil Nadu, India*

Abstract: The semantic images are those which are based on the presence of objects, their attributes and their relations to other objects within the image. But exactly characterizing this dependence requires extracting visual information from an image, which is generally a difficult problem. In this project, I propose studying semantic information within the images (images can be real and abstract images) created by us. We thoroughly examined the datasets to know semantically important features and the associations of words to visual features and methods for measuring semantic similarity. Finally, we study the relation between the notability and memorability of objects and their semantic importance.In this project, We have integrated wordnet for analysing all possible synonyms for the keywords given. Hence search efficiency, accuracy shall be improved.we propose a visual-attribute joint hyper graph learning approach to model the relationship of all images. Our aim of the project is to develop a meaning based search engine and increase the search accuracy and relevancy of search data for both images and web URL's.

Keywords: *Semantic,Hyper graph distance measure, Reranking, Keyword Aided.*

1. INTRODUCTION

Semantic mining has more scope and attraction in educational and industrial areas. Many existing search engines like Google, Yahoo and Bing executes the results based on the given keywords. Text based image retrieval is little challenging because the user should define the text associated to a specific images correctly. However if the attributes are not defined properly difficulty arises during text based image retrieval. Now visual based re-ranking is been defined for text based image search. In this visual information features are been extracted and compared during image search. Visual re-ranking methods are of 3 types, cluster based, classification based and graph based methods. The resulting images are re-ranked based on determining the relevance and number of times accessed by the users.

Identifying the fragments of attributes for image semantically is a challenge. Several works have been explored in figuring the semantic features for the images. Semantic meaning means understanding the meaning for the keywords and relevant keywords and relations between them.

In recent years, people are in eager of getting information from the internet. Search engines are used to fulfill the need of them. Even though the existing search engines are used to retrieve theinformation, the people will get irrelevant information for their search.Semantic search engine is used to retrieve the relevant information.

In order to provide the service to end user, a search engine extracts the Web pages from WWW repeatedly and store them using indexing scheme. There are broadly following five components of a search engine.

1. **Crawler:** The purpose of a Crawler is to crawl the Web pages from WWW and to create its repository.
2. **Indexer:** The purpose of Indexer is to index all the crawled Web pages into index repository.
3. **Searcher:** The function of Searcher is to search the relevant URLs based on user query from Indexed repository and put them in a list.
4. **Ranker:** The purpose of Ranker is to create ranked list of all retrieved URLs by searcher and handover it to user interface.
5. **User Interface:** User interface facilitates user interaction with the search engine.

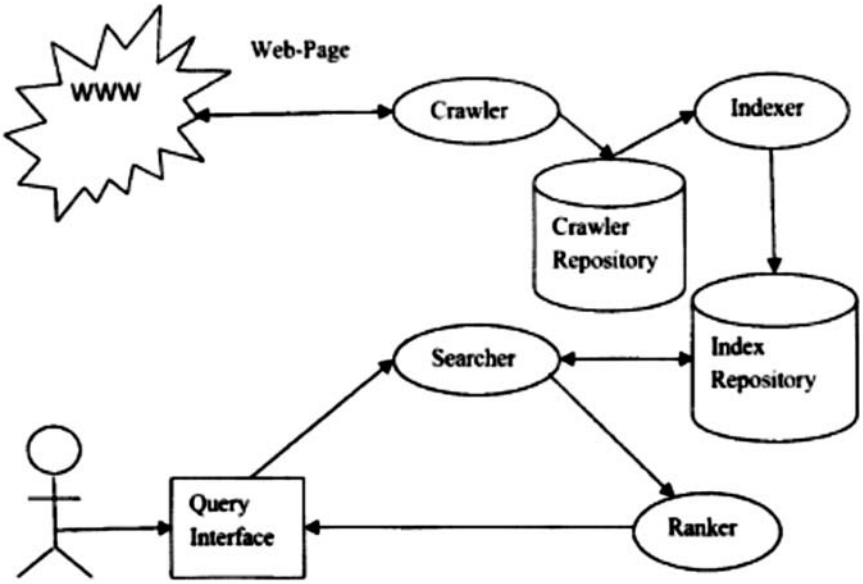

Figure 1: General Architecture of a Search Engine

1.1. Keyword-Based Search Engines

On the basis of Web for which a search engine works, they are divided into two types: the Keyword-Based Search Engine (KBSE) and the Semantic Search Engine (SSE). However in this, we will provide a brief description of the organization of the information and the problems with them.

KBSE's are based on plain web i.e. World Wide Web (WWW) , the meaning/context of keyword present in the document is not considered for indexing and retrieval but only the occurrence of keywords is considered for both the actions [2]. Performance of any search engine depends on how Web pages are organized or indexed. TF-IDF indexing scheme is a commonly used scheme for this type of search engines. This is a vector space modeling scheme [3, 4, 5] to index Web pages. In this scheme, importance of a term is decided based on the occurrence of term in the Web page. Terms having higher frequency is considered more important than a term having less frequency. Weight is calculated using the Inverse.

1.2. Problems with the Keyword Based Search Engines

Though these types of search engines are highly successful for both technical and commercial purposes, but still they are having some serious problems listed below.

1. **High recall, low precision:** Even if the main relevant pages are retrieved, there are other thousands of mildly relevant or irrelevant documents which areretrieved [1].
2. **Low or no recall:** It happens often that we don't get any relevant answer for our queries, or that important and relevant pages are not retrieved [1],
3. **High precision, low recall:** This situation occurs when all the retrieved documents are relevant but very few are fetched as compared to actually available in database [1],
4. **Results are highly sensitive to vocabulary:** Often our initial keywords in query do not get the results we want; in these cases the relevant documents use different terminology from the original query. This is imperfect because semantically similar search should return similar results [1].
5. **Results are single Web pages:** If we need information that is scattered over various documents, we must start several queries to collect the relevant documents, and then we must manually extract the partial information and put it together [1]
6. **Information Overkill:** It means the required information is lost (though available in responses) because of huge amount of information in the database [6].
7. **Problem of Synonymy:** Synonymy is the circumstance where different words explain the same theory. Thus a query in keyword based retrieval system may fail to retrieve a relevant document that does not contain the words which appear in the query [7,8]. For example, a search for "doctors" may not return a document containing the word "physician" and the word "cardiologist".
8. **Problem of Polysemy**: Polysemy is the circumstance where the same word has multiple meanings. So, a query in keyword based retrieval system may retrieve irrelevant documents containing the desired words in the wrong meaning [7, 8]. For example, a programmer and geographer looking for word "java" probably desire different sets of documents.
9. **Lack of Semantics:** In keyword based retrieval system, keyword list is used to explain the content of information object. Keyword list is a description that does not say anything about semantic relationships between keywords.

2. SEMANTIC WEB AND ITS ARCHITECTURE

Due to the above listed problems of keyword based search engines, researchers were forced to think about an alternate type of search engine. And finally they came up with a solution where the current Web page is supposed to be supplemented by additional information regarding the meaning of the page [5], this new type of web is termed as Semantic Web (SW).

SW proceeds in steps, each step building a layer on top of another so that additional information can be supplemented.

At the bottom we find *XML*, a markup language that enables creation of structured data documents. It gives a syntax for structured data documents, but urges no semantic limitations on the meaning of these documents.

XML Namespaces provides a way to use markups from different sources. They are used to refer to different sources in one document.

XML Schema is a kind of language which is used for controlling the structure of XML documents.

Resource Description Framework (RDF) is a type of language for generating a data model for objects (or resources) and link among them. It enables to presenting information in the form of graph. Though, the RDF data model does not depend on XML, but RDF is realized through XML-based syntax.

Resource Description Framework Schema (RDFS) gives basic vocabulary for explaining properties and classes of RDF resources. Using RDFS it is possible to create ranking of classes and their properties.

Web Ontology Language (OWL) extends RDFS by adding more advanced constructs to describe the semantics of RDF statements. It allows additional limitations, such as for example cardinality, restrictions of values, or characteristics of properties like transitivity. It appears as a way to capture more semantics and formally describe the meaning of terminology used in Web documents. It is based on descriptive logic and so bring reasoning power to the knowledge representation.

The Logic layer is used to increase the ontology language in addition and to permit the writing of application-specific declarative knowledge.

The Proof layer involves the actual deductive process based on the basis of complex properties as well as the representation of proofs in Web languages (from lower levels) and proof validation.

3. SEMANTIC WEB TECHNOLOGIES

In the SW, information is shown as an important asset of assertions called statements made up of three parts: subject, predicate, and object. Because of these three parts, statements inferred to as triples. The subject of a statement is the thing that statement describes, and the predicate states a connection between the subject and the object.

4. RELATED WORKS

KuhanandhaMahalingam et al (1997) reviewed that ontologies area in dynamic way to organize query formulation and semantic pacification inhuge and scattered information environments. Ontologies capture thesesemantic relationships, whether they exist among keywords or among thetables and fields in a database. Ontology is a network structure that givesusers with an abstract view of a domain specific information space. Ontologies are said to be well suited for knowledge sharing in a distributedenvironment (Iqbal et al 2009).

Ontologies have abenefit overunstructured text-based information spaces for linking values to differentunits or formats, since query outcomes do not typically contain informationabout the units of returned values. For example, if a query asks foremployee salaries, the outcomes do not indicate whether the salaries are indollars or pounds or both. Ontologies are best for resolving such problems. Butthey are difficult to build. So, the authors implemented the Java OntologyEditor (JOE) to benefit users build and browse ontologies. It also enables queryformulation at several levels of abstraction, including a very abstract levelcomfortable for novice users.

Amit Sheth (2002) proposed that search engines do not consider theuser query's context. In some of the cases, the search could be restricted to the technologycategory, but this alone still does not show the difference between the operating system andthe product. A more difficult query is to find movies that Robert Redforddirected, but not those in which he acted with a different director.

Non-semantic search engines cannot correctly answer such queries becausethe keywords "director" and "Robert Redford" could occur in documents notsatisfying these criteria. They designed Semantic Content Organization andRetrieval Engine (SCORE) for building a comprehensive solution for researchand analysis-oriented semantic applications that deal with a broad variety ofcontent sources.

MudassarIlyas et al (2004) proposed a conceptual architecture for asemantic search engine. This architecture has ontology editor,ontology mapper, ontology translator, Web page annotator, ontology crawler,Web crawler, query builder, knowledge base and inference engine. Also, theymainly focused on the reference engine and uttered that the relational databaseis used to store the knowledge base, and the ontologies are used to removethe flaws in the current reference engines.

Liang Bangyong et al (2004) proposed a novel method to improvethe semantic search engine using ontology language like OntologyInterference Language (OIL). The authors proposed that current web searchmainlydepends on the keywords in the web pages. This method is lacking semanticsin many forms. For example, a search for a person by the person's name meansto find the web pages that contain the text of the name. On the contrary,semantic search is to find the information about the person in the real world.It is hard to attain this mark in current content based web search enginesbecause text is not useful during reference.

Fang Yuan et al (2004) suggested that search engine is the mostimportant information extraction tool. However, most popular search enginesare based on HTML documents which lack semantic retrieval and personalizedservice. The authors presented the concept of extendedMarkupLanguage (XML) search engine and evolved a framework of an intelligentXML (John Miller and SonaliSheth 2000) search engine and also debatedthe important techniques in intelligent XML search engine.

Tikk et al (2006) discussed that users often face the problem offinding the relevant outputs on the result pages for their search. Theauthors grouped the answers for the query into topics of a fixed subjecttaxonomy. In this manner, the original problem can be changed to the indexing of queries and the results with the topic names using organized learning algorithm. The authors introduced ferrety algorithm that performs topic assignment, which also works when there is no directly available training data that describes the semantics of the subject taxonomy.

Amasyah (2006) has used similarity measure for the word classification. The author has taken Turkish words. Semantic similarity is needed to solve several natural language applications. For developing schema, they examined XML node path index, semantic keyword index and element tag index. The semantic link coefficient in between the two nodes and between a keyword and a node and by referring from ontology base is formulated in the search algorithm. Capasso (2006) proposed that effective search and retrieval is required for realizing the full potentiality of the Web. Although now a days search engines execute better than the search engines a few years ago, big betterments are still needed with respect to the relevance of the retrieved documents to the user's query and the presentation of the results. In his work, a prototype retrieval system is developed using WordNet for identifying the related documents and ranking them according to their relevance to the query.

Du Zhi-Qiang et al (2007) implemented a framework of semanticsearch engine derived from on ontology to solve the flaws of the low queryaccuracy and the limitedness in understanding the user's query intention thatoccur in a traditional search engine. They proposed the information extractionalgorithm based ontology. They developed a prototype of search engine by using oflucene, and the search result is better than that of the common search engine.HaiyanChe et al (2007) implemented a blueprint of semantic-based searchengine framework. This framework can retrieve factual knowledge fromChinese natural language documents automatically by combiningtechnologies of semantic Web, information retrieval, natural languageprocessing and a novel theme-based method. Instead of listing of documentlinks, results of user's query request returned from framework aresemantically coherent reports, which can satisfy users completely.

Li Yong (2007) proposed a personalized search engine tocontrol the lack of personalization, poor recall and precision (NasraouiandZhuhadar 2010) ratios and lower efficiency of search. The core of the sharedknowledge base and personalized query is ontology technology, and this can obviously enhance the recall and precision ratios. Moreover, the application of the users' interest base can improve the qualityand efficiency of customer service.

Iosif Elias et al (2007) presented two novel web-based metrics forsemantic likeness computation between words in a web search engine. Thefirst metric considers only the page counts resulted by a search engine. Thesecond metric downloads a frequency of the top ranked documents. Theproposed metrics work automatically, without seeking any humanattributedknowledge resource. The metrics' performance is assessed in termsof connection with respect to the pairs of the commonly used Charles – Millerdataset. The proposed "wide-context" metric gains 71% correlation, whichis the highest score gained among the fully unsupervised metrics in theliterature till date.

Existing system: The existing system is a keyword based search engine where precision and relevancy is lacking. No semantic based similarities is implemented in the existing system. In google, which is an attribute based search engine that have relied on corresponding textual information of the images against queries given by users. Existing re-ranking approaches are low-level visual features.Visual reranking method divided into Clustering based,Classification based and Graph based method.The cluster based re ranking method stem from the key observation that a wealth of visual characteristic.Purely based on low level visual feature while generally do not take any semantic assosiations among initial ranked list in to the consideration.

Proposed System: Since there is lack of accessebility of abstract images, We are using real images for processing. To extend the research we propose to implement semantic based search in web URL's also by various datasets in various categories. The proposed method uses images for three purposes-To make set of similar semantic words, With use of saliency, we can relate many words. The Co-saliency is replicated as a linear combination of the single image saliency map and multi-image saliency map. The memorability of the picture has strong influence in our mind.Proposed to purify text-based search results by utilize the visual information contained within the images. After a query "boy" is submitted, an initial output is acquired via a text-based search engine. It is noticed that text-based search often provides "inconsistent" results. A fast and accurate scheme is proposed for grouping Web image search results into semantic clusters. It is obvious that the clustering based reranking methods can work well when the initial search results contain many near identical media documents. Hypergraph Distance Measure Algorithm is been proposed for search optimization.

5. ARCHITECTURE DIAGRAM

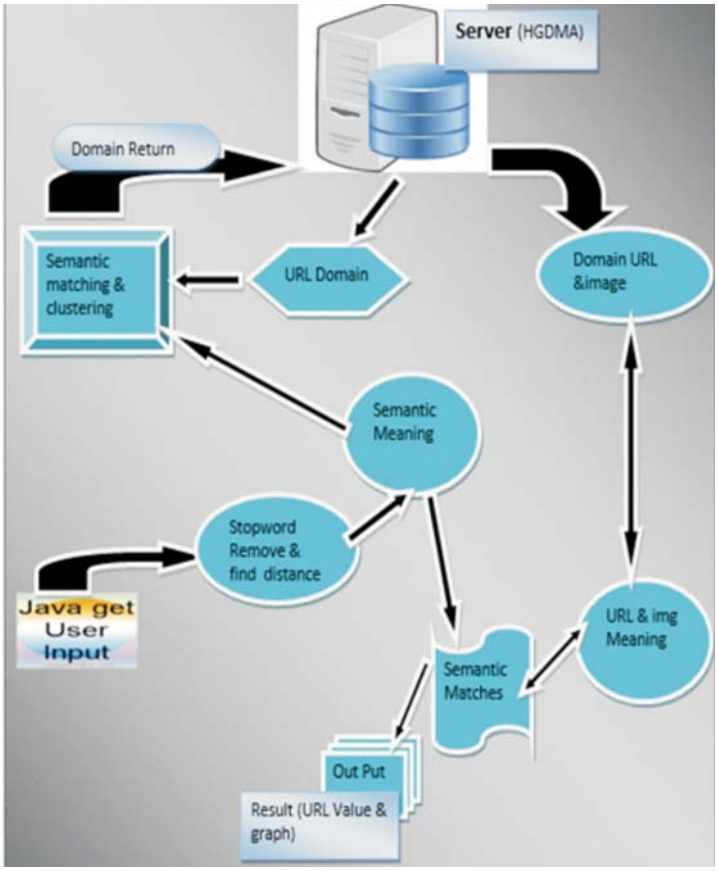

Figure 1

6. WORKING METHODOLOGY

Web image based re-ranking is one of the encouraging method for image retrieval. In this semantic based search engine is proposed to identify the features from the visual image and provide the results. The relevant results are been shown at the top and less relevant results are shown in the lower. For each image, attributes are been assigned during image upload. Example: Our hyper graph distance measure algorithm identifies the animal and car image by identifying the attributes. Also relevant or relative keywords are identified using WordNet. Thus it increases the image results and accuracy.

We propose hyper graph learning theory for finding the semantic attributes. Thus this algorithm finds the relevant and highest matching keyword first and other highest matching keywords are listed below. In the text based search visual information contained within the images are been obtained. Example: When a query named "baby" is been submitted, the relevant keywords are identified by integrating WordNet. Attribute based re-ranking is used for listing the most relevant search at the top and less relevant search at the bottom.Eventually, the most relevant results are moved to the top of the result list while the less relevant ones are reordered to the lower ranks

7. EXPERIMENTAL RESULTS

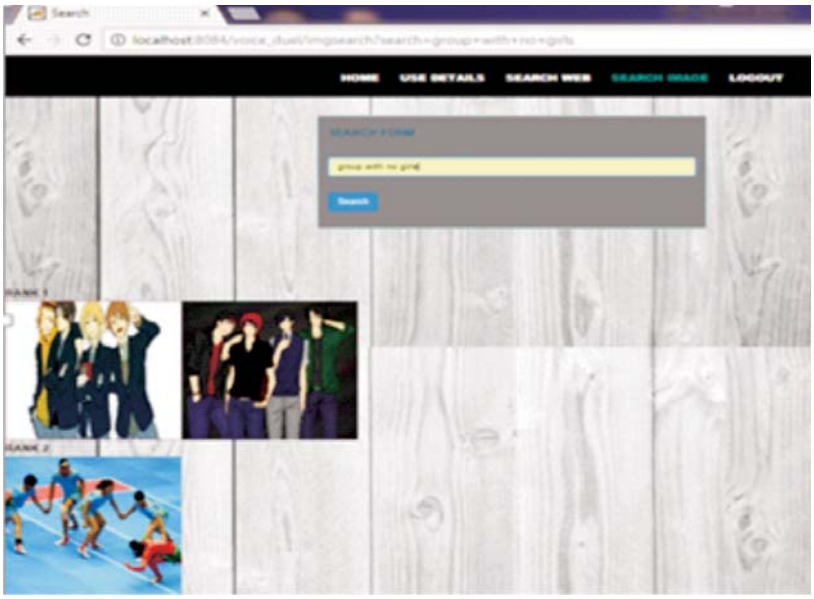

Figure 2

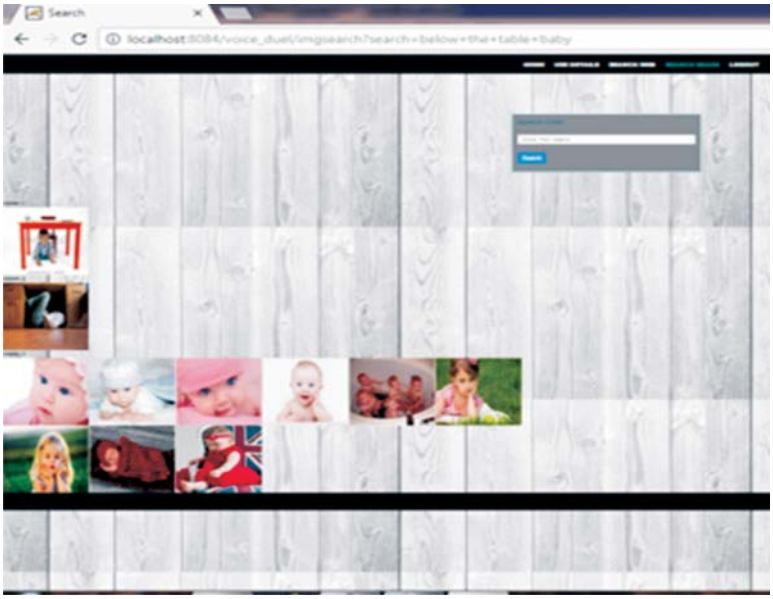

Figure 3

8. CONCLUSION

Image based re-ranking search combining semantics would enhance the results of keyword based search engine. WordNet is integrated to identify the relevant keywords. A hyper graph is used to identify the association between the visual features and attribute features. The basic idea is to improve the efficiency of the image results using hyper graph distance measure algorithm. Also for result optimisation attribute based image re-ranking is used.

REFERENCE

[1] R. P. Adams, Z. Ghahramani and M. I. Jordan. Tree-structured stick breaking for hierarchical data in 2010

[2] Xueqing Liu, Yangqiu Song, Shixia Liu, Haixun .Automatic Taxonomy Construction from Keywords.In ACM 978-1-4503-1462, in 2012.

[3] K. A. Heller and Z. Ghahramani.Bayesian Hierarchical Clustering. In ICML, volume 21,in 2005.

[4] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen. Short text conceptualization using a probabilistic knowledgebase. In IJCAI, pages 2330–2336, In 2011.

[5] T. Lee, Z. Wang, H. Wang, and S. Hwang. Web scale taxonomy cleansing. PVLDB, 4(12):1295–1306, in 2011.

[6] W. Wu, H. Li, H. Wang, and K. Q. Zhu.Probase: A probabilistic taxonomy for text understanding. In SIGMOD, 2011.

[7] J. Wang, H. Wang, Z. Wang, and K. Q. Zhu. Understanding tables on the web. In ER, in 2012.

[8] Y. Wang, H. Li, H. Wang, and K. Q. Zhu. Toward topic search on the web. In ER, in 2012.

[9] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen. Short text conceptualization using a probabilistic knowledgebase. In IJCAI, pages 2330–2336, in 2011.