Image Retrieval with A Multi-Modality Ontology

A Thesis Submitted to the
School of Computer Engineering
of the Nanyang Technological University

by

Wang Huan

in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy (Ph.D)

2009

Abstract

Ontology represents domain concepts and relations in a form of semantic network. Different from the traditional flat structured feature vectors, ontology provides more information through concept definition and relation inference. Many research works use ontologies in information matchmaking and retrieval. This trend is further accelerated by the convergence of various information sources supported by ontologies in the multimedia research field. Efforts have been made to shift traditional *content-based* retrieval approaches toward *concept-based* approaches. Both text and image features are extracted and integrated to improve the classification and retrieval performance. However, there are many open issues on the ontology understanding, definition, construction, utilization and implementation. The extra work required by ontology based approaches becomes one of the major difficulties that hedge against the development.

This thesis mainly focuses on finding an effective ontology model for multimedia information retrieval. More specifically, we look at the problem of image retrieval in the dynamic and noisy web environment, and propose a multi-modality ontology to better understand web image by incorporating both image and text features. Prototype systems are set up to prove the feasibility of the model. We then investigate the problem of scalability in ontology construction. Later, by taking advantage of both structural and content features of the online encyclopedia Wikipedia , real world objects are formalized in terms of concepts and relationships. Association rule mining algorithm is designed to improve the quality of the generated ontology. The retrieval performance by the automatically built ontology is comparable to the previous manually built one.

Besides the ontology construction issue, the ontology model is also investigated from the inference perspectives. Firstly, traditional description logic based reasoning is discussed and used in the semantic matchmaking process. Based on the output from the

reasoner, a ranking algorithm is presented to enhance the matching result. Well-defined weight can be utilized to emphasize different ontology modalities. Later, towards larger scale ontology inference, we provide a novel understanding of the ontology and consider an ontology as certain type of semantic network, which is similar to brain model in the cognitive research field. Spreading Activation Techniques, which have been proved to be effective information processing models in the semantic network, are consequently introduced for inference. Through comprehensive experiment, the inference method is shown to be a scalable architecture for larger ontology.

Moving from concept ontology which contains explicit semantic relations, we introduce the idea of building large scale concept ontology in the form of *concept thesaurus* from Wikipedia Corpus. This thesaurus aims to provide the maximum coverage of concepts, together with various relations among them. By effective utilization of the thesaurus information, concept detector space is constructed for concept similarity calculation. The concept detectors help to achieve better precision, automation and scalability in real world web image retrieval.

Acknowledgments

First of all, I would like to express my sincere gratitude to my supervisor, Dr. Liang-Tien Chia, who has given me invaluable advice on this research. I appreciate his time and effort spent in supervising me. Without his trust and support I would never have the confidence to start my research life. He has been a great supervisor, and I would always remember those precious moments, when he stayed up late to review our papers, when we celebrated together for our achievement, when he sent sincere bless to my wedding... There is more to remember and again, I am grateful to his guidance and understanding.

I would also like to express my thanks to Mr. Liu Song, Mr. Jiang Xing, Mr. Zhou Chen, Mr. Chu Yang, Ms. Wang Surong, Mr. Hu Yiqun, Mr. Cheng Xiangang, Mr. Gao Shenghua, and Mr. Wang Zhengxiang. Thanks for all the precious advices and thought-provoking discussions. I am so lucky to have their accompaniment during my candidature.

I own a lot to my family for their continuous understanding and encouragement. They are always there to listen to me and to give me advice. Special thanks must go to my hubby, He Jun, who is always there supporting me. Especially during the later stage of my Ph.D candidature, I have been in poor health. He took good care of me and supported me through all the pressure.

Finally, I would like to thank the Lab technicians, Chua Poo Hua, Oh Hwee May, Lim-Tan Lay Choo, Ng-Siom Siew Ling and all the other CeMNet members. Thanks for their support during all these years.

Contents

List of Figures

List of Tables

List of Abbreviation

ACCR	Average Correct Classification Rate
AP	Average Precision
CBIR	Content-Based Image Retrieval
CLD	Color Layout Descriptor
CRT	Composite Region Templates
CSD	Color Structure Descriptor)
DCD	Dominant Color Descriptor
DL	Description Logics
EHD	Edge Histogram Descriptor
IR	Information Retrieval
LDA	Latent Dirichlet Allocation
MAP	Mean Average Precision
MPEG	Moving Pictures Experts Group
QBE	Query-By-Example
RDF	Resource Description Framework
RF	Relevance Feedback
ROI	Region of Interests
RQL	RDF Query Language
SAT	Spreading Activation Theory
SCD	Scalable Color Descriptor
SIFT	Scale Invariant Feature Transform
SOM	Self-Organizing Map
SVM	Support Vector Machine
WWW	World Wide Web

Chapter 1

Introduction

Nowadays there has been an emergence of a new multimedia information retrieval paradigm called concept-based retrieval. It goes beyond multimedia features representations and tries to build up dynamic concept models of human knowledge for machines to understand, reason, classify and retrieve information like human-beings. This process involves knowledge acquisition, representation and reasoning. Ontologies[1], which are domain specific and represent concepts and relations in a form of semantic network, provide such an effective semantic structure to describe visual content and improve the retrieval performance through concept matchmaking. They aim to overcome the semantic heterogeneity among domains and provides a shared and common understanding of a domain that can be communicated between people and across application systems. The main goal of this thesis is to build an effective ontology model to improve the current web image retrieval performance. The thesis also investigates various ways of inference to derive extra knowledge from the built ontology.

1.1 Background and Motivation

Today's web is flooded with explosive digital multimedia information made available by the fast developing media and network techniques. Multimedia information retrievals with both efficiency and efficacy become popular research issues, and web image retrieval is no exception. Most popular web image retrieval systems are based on keywords. Some famous commercial search engines include $Yahoo^{(TM)}$ and $Google^{(TM)}$ image searches.

However, this approach suffers from certain limitations. Firstly, text-based image searching needs adequate text information which provides highly related information about the images. Otherwise, the loosely-coupled relationship between Web images and Web textual contents may provide misleading information and affect the final retrieval result. Secondly, as human natural language is of high complexity, precise semantic interpretation for text is not available in the simple keyword matchmaking. The search result relies heavily on users' knowledge to narrow down the search target. Moreover, the useful image features are totally ignored in these attempts. There are also a substantial proliferation of works on content-based image retrieval (CBIR)[2], which uses image processing operations to extract the low-level features automatically from image content. The image features are converted to vectors and transposed into other spatial data array. These image processing operations include color processing, image texture processing and local shape processing. Image retrieval is performed by matching the features of query image with features inside database. However, due to the *semantic gap*[2], extracting semantically meaningful image content from low-level features is still an open issue.

As we can see, most of the aforementioned works only use single-modality information – either text or image features, that they suffer from certain limitations. In order to explore various available information resources and develop an image retrieval approach which utilizes multi-modality information, efforts have been made to introduce the concept of ontology and semantic matchmaking into image retrieval to bridge the *semantic gap*. Ontologies is defined as "a specification of a representational vocabulary for a shared domain of discourse — definitions of classes, relations, functions, and other objects."[1] The definition can be further understood as that ontologies are domain specific and represent concepts and relations in a form of semantic network. They aim to overcome the semantic heterogeneity among domains and provide a shared knowledge.

Since most information retrievals are at a higher human-concept level, rather than the low-level visual or textual features, this thesis proposes a different approach to improve web image retrievals. We believe ontology would help the retrieval system to understand both the user query and image content in a more effective way. We build up a novel ontology from the domain knowledge of *animal* and incorporate information from multiple modalities, which includes both text and image features. Such a semantically rich

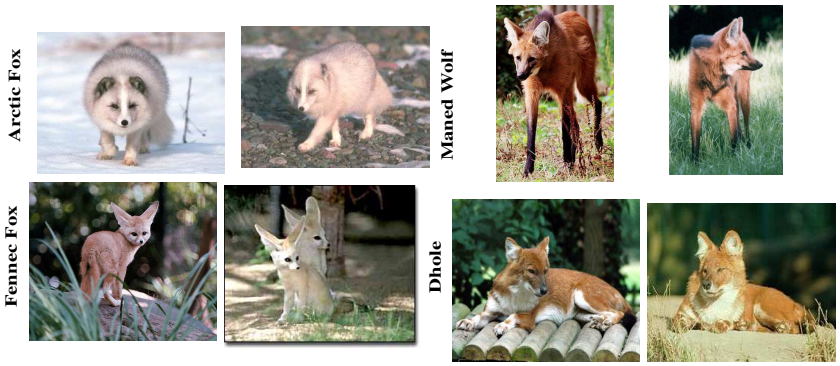


Figure 1.1: An example of web image classes in our data set.

ontology addresses the need for complete descriptions of image retrieval and improves the precision of retrieval compared to conventional methods.

1.2 Problem Definition

Among the numerous possibilities enabled by ontology, we are interested in exploiting ontology for a better understanding of web image on the World Wide Web. This presents us with issues of image processing, scene understanding, text analysis, ontology construction, and semantic matchmaking. We are especially interested to investigate the improvement on the retrieval performance brought by ontology in *animal* domain where various object shapes and diverse surrounding environments are involved. Taking the images of four different animal classes in Figure1.1 as an example, it is easy for human-beings to identify the classes: *Arctic Fox* has light-colored fur; *Fennec fox* has a pair of grotesque ears and ET-style face; *Maned Wolf* is featured with its black long legs; and *Dhole* has white fur spreading from its jaw to abdomen. However, it is not easy for even the most advanced image processing techniques to tell the classes apart due to lack of discriminating low-level features.

The goal of using a multi-modality ontology,which uses features from both text and image modalities, is to create a machine-processable representation for web image retrieval. This image representation overcomes the semantic heterogeneity among various images. As a semantic network consisting of nodes and arcs, ontology provides a useful

structure to help computer learn and understand human knowledge. The concepts encoded in the ontology takes the place of the human pre-knowledge. And the semantic matchmaking process mimics the human inference model and derives additional facts from the available information.

To achieve our goal, the following open issues are inevitably involved:

- How to design a proper structure of an ontology for the purpose of multimedia information retrieval? Since we are looking beyond text modality and aim to extend to image modality, more relations should be captured in our ontology structure. After the structure is defined, how can we find a proper reusable external resource to extract necessary knowledge, and proceed to use the knowledge to build up the ontology?

- Given an effective ontology structure, how to implement ontology construction more efficiently and avoid the tedious manual work? Some methods tried to automate the construction process by extracting partial ontology with hierarchical structure from WordNet[3–6]. However, the challenge is that most ontology construction are still based on large collaborative projects which involve substantial manpower. We want to investigate more efficient ways to construct an ontology with rich semantic information.

- What are the effective inference models when using an ontology, especially when a large ontology domain is involved? For image retrieval purpose, we need to compare the concepts derived from web images with the concepts in the knowledge base. The problem is how to implement the comparison process through knowledge inference. Scalability is another challenge, as when ontology grows bigger, the efficiency of inference process can become a bottleneck. To answer this question we should first look into traditional description logic based matchmaking approaches with semantic reasoners. Later we also consider ontology as a kind of semantic network and study neural network based inference models, which are more adaptive to larger concepts domains.

- How to extend the ontology to a large scale from both ontology construction and inference perspectives? Researchers have been persistently pursuing for a more "complete" ontology in the sense that it could cover most of the concepts in general knowledge. This is a big challenge as it is difficult to define the mass concepts. There are many efforts and discussions about how to define such a thesaurus-like ontology and most importantly, how to utilize such a huge knowledge base. In this thesis, we discuss the possibility of building ontology for extensive domains which cover millions of concepts, and explore ways of using such a thesaurus in web image retrieval experiments.

Since this thesis works on web image retrieval problem, the above problems can be concluded into two main issues: the multi-modality ontology construction and the semantic matchmaking for image concepts. To answer these questions, we need to decide our experimental domain, collect web images with web pages and set up database. As a preliminary attempt, our experiment is conducted on web images of the animal domain of *animal*. As shown in Figure1.1, it is a challenging domain due to animals' varied shapes and complex living environments. The taxonomy structure of animal classes is good for concept hierarchical relation inference test. And the rich information used to describe animal species allow us to mine useful concepts and relations for concept identification. Ontology is effective on capturing and integrating various aspects of information. To prepare our data set, we crawl animal images of different classes with their related web pages from the Internet to set up an image database. The experimental scenario is defined as: Given one input query, we want to have a high accuracy of retrieval result, especially for top retrievals, as these results are most concerned by web users.

1.3 Research Contributions

In this thesis, we mainly focus on the solution of the aforementioned two open issues in ontology-based web image retrieval field, which are the ontology construction and ontology inference problems. In a typical experiment scenario, after a query is submitted to the image retrieval system, we want to find the most relevant images to the query. Based

on the constructed ontology which provides us knowledge base, each web image is analyzed and depicted by an unknown/anonymous concept. It is from the inference process between the unknown concept and the knowledge base, that we can locate the positions of the related concepts in the ontology structure and further detect the concept(s) associated with the web image. Therefore, we aim to tackle both ontology construction and inference problems, and provide our original answers and discussions.

Firstly, we design an effective ontology model for knowledge representation, which integrates information from multiple modalities and facilitate the reasoning over such information. The main purpose of designing such an ontology model is that it could help machines to discover useful semantic knowledge from the noisy web text information, such as image captions, file names, surrounding texts, and image itself. Therefore, the concept structure in the proposed ontology should be defined according to the general human knowledge. The major contributions in this part can be categorized as follows: We differentiate our ontology model from other ontology models in [4–6] in the sense that the proposed model preserves more concepts and relations according to general knowledge. Therefore the final concept set preserves more relationships than commonly-used hierarchical class relationships, and is able to extract additional information from the web pages to increase the concept detection accuracy. A method to semi-automatically generate the ontology model is firstly proposed. Class and relationship definitions in the ontology model are extracted from the BBC Science &Nature Animal category[1]. The proposed multi-modality ontology is proved to work well in our web image retrieval task. Since the semi-automatic method still involves tedious and time consuming manual work, the question is further extended to how to efficiently build the ontology model. We consider both the ontology completeness and scalability problems, and provide our original solution which is to build ontology automatically from Wikipedia[2] in a cost effective way. As we get expert knowledge from the bigger knowledge base, the final concept set preserves more relationships than manually built ontology. Combined with the middle concepts from image features, we arrive at an ontology model which combines multi-modality information for web image retrieval. More specifically, the text cues in

[1]http://www.bbc.co.uk/nature/wildfacts/animals_a_z.shtml
[2]http://en.wikipedia.org/wiki/Main_Page

the ontology is extracted from natural language in the web page text, while the image cues are formulated as middle-level concept through a feature training process. Both cues are combined seamlessly in the final ontology.

Given such an ontology, the next issue is how to apply it to image retrieval. What is an effective method to inference over the knowledge base and the generated concepts? For this purpose, we first try the traditional reasoners (semantic matchmakers), whose inference results are somehow roughly categorized according to human cognition (e.g.:exact match, subsume match). To provide further measurement, we also propose an enhanced ranking algorithm to quantify the initial inference result. Later, to cope with extended concept structure, we consider our multi-modality ontology as a type of semantic network, since they are of similar structures. Inspired by the spreading activation process in the semantic network of the cognitive field, an inference model based on spreading activation theory is proposed for ontology inference.

As a final part of our work, we discuss the possibility of using the structure and content features of Wikipedia to build a concept thesaurus. The motivation is to consistently improves performance in web image understanding and retrieval with the incorporation of large-scale concepts. By leveraging the prior knowledge from the thesaurus, we are able to mine semantic salient concepts from the sparse natural language descriptions. We move a step forward to combine the similarity distance from the image feature space to boost the web image retrieval performance.

1.4 Dissertation Organization

The rest of this thesis is organized as follows: Chapter 2 gives a literature survey, including an introduction of the background knowledge. Some existing works are also discussed. Chapter 3 focuses on the construction of the multi-modality ontology model. This model is composed of 3 sub-ontologies, which involves analysis for both text information and image content. Chapter 4 discusses the semantic matchmaking part in the image retrieval. We first introduce the background knowledge of description logic behind the semantic matchmaking. A survey is also done on the existing semantic reasoners. After we elaborate on the matching algorithm, we explain the work flow model and the

overall structure of our image retrieval system. Experiment results are also shown in Chapter 4, where we compare the result of Google with both single-modality ontology and multi-modality ontology, and show that the multi-modality ontology outperforms the other approaches. Chapter 5 discusses our attempts to build a large-scale concept thesaurus from Wikipedia and use the thesaurus to detect salient concepts in the noisy web environment. A conclusion is given in Chapter 6, where we iterate our main contribution and propose future research direction.

Chapter 2

Literature Review

Web images are digital images placed on the web. They are usually attached with captions, file names, and surrounding text descriptions. Due to the exponentially increasing number of web images, how to effectively retrieve images with high precision has become a challenging topic. This chapter starts with a discussion of traditional image retrieval approaches, which are widely adopted by most web image search engines. It is followed by an introduction of recent research of semantic related image retrieval and a review to some existing works in web image retrieval. By discussing the pros and cons of these works, the motivation of importing ontology and semantic matchmaking to image retrieval is shown.

2.1 An Overview of Traditional Image Retrieval

The concept of *image retrieval* is introduced after the rapid development and adoption of digital image techniques. Some basic operations are feature extraction, image representation and query processing. Feature extraction is usually done on two modalities: high-level text information and low-level image features. Based on these two approaches, existing image retrievals can be classified as text-based image retrieval and content-based image retrieval. In the thesis, traditional image retrieval is defined as those approaches exclusively based on single-modality information, e.g., text-based image retrieval and content-based image retrieval. These two fields have been discussed for a few decades and the techniques are mature.

2.1.1 Text-Based Image Retrieval

Ever since 1970s, text-based image retrieval has been studied intensively. Images are first indexed with text annotations and stored into a database. Some representative surveys are shown in [7, 8]. The most conventional approach would be matching the input query directly with text features of images, such as image captions, file names, web page titles, page contents and so on. Most text-based image search engines on the WWW, such as $Yahoo^{(TM)}$ and $Google^{(TM)}$ adopt such technique. Users of these search engines input keywords or select filtering values to find their targets. More complex queries are also available by a combination of the simple Boolean logic like *AND* and *OR*.

The text on the web page also provides useful cues for web image retrieval. Due to the nature of the web, it is important to preprocess the text and provide a clean text corpus prior to further analysis. Some basic preprocessing steps include stop words removal and word stemming. In the former step, trivial words are pruned according to a customized list of words, while the latter step is a process to transform the words back to their stem. After the preprocess, techniques from natural language processing fields are employed to analyze the surrounding text information. These techniques usually rely on statistics like word frequency, and construct a text feature space. Surrounding text for each web image is treated like a document. Text vectors are used to represent the document in the feature space and similarity is calculated based on the distance between vectors. A representative vector model is the tf-idf scheme[9], which is a statistical model for the distribution of words in a text corpus. A dictionary is first generated according to both the statistics of the global and local term frequency in one text corpus. Based on the frequency information, the terms in the final dictionary have higher discriminative power. For each web page in the corpus, the importance of the term is measured by the term frequency in the document, weighted by its global frequency in the text corpus. According to the generated dictionary, the text information for each web page is transformed into a vector, with each entry representing the weighted value of the term frequency. Based on the vectorial text features, similarity is calculated for image retrieval.

Text based approaches are efficient and straightforward. However, they suffers from some limitations. First, in a noisy web environment, many of the web images are not well annotated. Some of the annotations could be uninformative or misleading. Second,

synonymy and polysemy introduce another open issue of disambiguation. Third, due to the subjectivity of human judgment, web users could have different understandings for the same annotation. It is always the semantic meaning, rather than the standard keyword that really decides the matching accuracy. For example, if a user wants to find images of certain *fox* living in South Africa, he will not be able to get correct result by typing "fox with habitat in South Africa". Unless he can specify the exact keyword, such as *cape fox*, the search engine will not provide highly related results.

2.1.2 Content-Based Image Retrieval

Due to the limitation of text-based image retrievals, more research works focus on image content information. Content-based image retrieval (CBIR)[2] uses image processing operations to extract the low-level visual features automatically from image content. As a result, the tedious manual annotation process is bypassed. The basic visual features can be classified into color, shape and texture features. The image features are converted to vectors and transposed into spatial data array. The Information Retrieval(IR) process is based on these image features. A number of web image retrieval works[10–16], which have different emphasis on the image features, have been proposed.

Color feature has always been considered as a robust visual feature. Both global color information and local color information are useful in identifying the background and foreground objects. Representative color features include dominant color, color histogram, color moment, color layout, and color correlogram[17, 18]. MPEG-7[1] has defined a series of color descriptors, including dominant color descriptor(DCD), color layout descriptor(CLD) and color structure descriptor(CSD). These descriptors represent both global and local color information. In [10], Laaksonen et al. organize MPEG-7 descriptors into a self-organizing map(SOM) for image retrieval. The proposed framework PicSOM uses parallel SOMs to train separate descriptors, which are mainly color descriptors, and combines the responses together to form the final decision. However, in order to find a proper weight, users need to input a large number of descriptors to train the SOM at the beginning. This process does not set up a link between the descriptors and human concept. Color histogram is another commonly used feature, which depicts distribution

[1] http://www.chiariglione.org/mpeg

of pixels according to different color ranges. Other than conventional color histogram approaches, variations on the color histogram which use different metrics for similarity measures have also been proposed in [11, 13, 19]. Color set[20], which is an approximation of color histogram has been proposed for fast retrieval. Even though the efficiency is improved, the result accuracy depends on the correct segmentation of the image. Color moment is also discussed as an improved feature of color histogram in [19], which uses the low-order moments of color distribution for similarity calculation. In [16], color correlogram[18] is extracted as low-level features to describe the web image content. This feature contains spatial correlation of colors and provides global distribution information of local spatial correlation of colors. This feature is robust when dealing with images of same scene but different viewing positions or partial occlusions.

Texture is another effective visual cue, as it provides surface pattern information of image independent from single color or intensity[12]. Traditional texture features are modelled as random field modelling[21], co-occurrence matrices[22] and Gabor filters[23]. In [21], images are considered as 2D homogeneous random fields and decomposed into mutually orthogonal components. This decomposition is an approximation of human perceptual properties of "periodicity", "direction", "randomness". This texture model is proved to be effective in retrieving perceptually similar natural textures. Co-occurrence matrix[24] has been proposed to include spatial information of texture. The texture representation is extracted from the co-occurrence matrix, which is built from orientation and distance of image pixels. In reference [23], a bank of Gabor filters is used to filter images, and the filtered result is subjected to nonlinear transformation for feature extraction. Spatial segmentation is produced by unsupervised clustering algorithms. The proposed texture segmentation achieves good retrieval result in both nature and artificial textures based images. Besides the aforementioned traditional models, in early 1990s, the wavelet transform becomes popular in texture representations[12, 14]. Texture information is extracted from the mean and variance of the wavelet subbands. Wavelet based transformation can be combined with co-occurrence matrix to improve texture analysis[25].

Most shape features are transformation invariant, which means that the features are independent from the translation, rotation and scaling. The representative shape features

include Fourier descriptor[15] and moment invariants[26]. Fourier descriptor is commonly used to represent the boundary-based information, while moment invariants represent the whole shape region. In [15], Fourier descriptors are used to analyze the closed curves in the plane. The function used to represent the curve is expanded in Fourier series. The coefficients obtained through the expansion, which are proved to be insensitive to curve's starting point, are used as descriptor for further shape similarity measure. The use of moments for shape description is first introduced by Hu in [26]. Seven transformation invariant moments are proposed for region-based shape representation. Variations of moment invariants are discussed in[27, 28]. In [27], Zernike moments project the image onto a spanned space and use the amplitude as the shape feature. This feature works well for character recognition of binary solid symbols. Chen[28] proposes improved moment invariants which are invariant to scaling, translation and rotation. Research[29] has shown that a combination of the Fourier descriptor and moment invariants outperforms either single feature. Curvature Scale Space(CSS) technique[30] provides another effective contour shape descriptor for planar curves. The curve is turned to path-based parametric representation, and convolved with a Gaussian function. With a varied standard deviation of the Gaussian, the curvature zero-crossing points of the resulting curves are extracted. Since this representation is robust to non-rigid deformations and perspective transformations, it has been chosen into the MPEG-7 standard for contour-based shape representation.

Recent researches[31–35] use Harris-Laplace detector[32] which is scale invariant and detects corner-like regions in images as interest points. Scale Invariant Feature Transform(SIFT) descriptor[36] is then used to represent the information around the interest point. SIFT feature extracts histograms of gradient orientations from the image region. It has been proved to be the most robust local feature descriptor. Because of its popularity, there have been many variations and extensions[37–39] of SIFT feature. In [37], Principle Component Analysis is applied to the gradient patch. To improve the efficiency of SIFT descriptor, fast approximated SIFT[38] is proposed by using integral image and integral orientation histogram for interest region detection and description. To include color information into the original SIFT feature, [39] proposes Colored SIFT to build the SIFT descriptors in a color invariant space.

Researches on separate image features have already proved their success for image retrieval. In most cases, these features are combined together to boost the retrieval performance. Especially in web image retrieval domain, the heterogeneity of images has made it more difficult for visual feature selection. Many content-based image retrieval works[40–49] use relevance feedback(RF) mechanism to learn proper combination and weight of image features. Systems employing RF mechanisms will first require users to input an sample image as query, and then let users give feedback on the relevancy of the returned results. Given the users' feedback, the systems retrieve images again through training according to the new information. The feedback-relearning iteration may repeats several times to ensure an optimal result. A representative work MARS[40] aims to organize different visual features into a meaningful retrieval architecture, instead of finding a best feature representation. It integrates computers and human expertise for image retrieval. A relevance feedback architecture is proposed in MARS to understand human perceptual models. The top retrievals are examined by users and the next iteration shows returned results based on the users' feedback. PicToSeek[41] is another similar work, which retrieve web images by image content features and relevance feedback. Users input image as query, and assign different weight to the visual features, such as color and texture. The adjusted weight return more relevant images according to users' perception and subjectivity. Another system Web-WISE[42] also requires users to give either image samples or estimations of feature values as input query, and returns the user a set of web image thumbnails with similar color and texture feature. Real applications based on such content-based web image retrieval systems include RIME[43], which aims to find unauthorized image copies on the WWW. After the user gives a copy detection request, the system compares the image feature vectors with those stored in the feature repository, and fetches similar images as positive detection. In Cortina[45], high-level semantics from the text are first used to retrieve images. Later visual features are used for finer granularity of result. RF is also considered as a good practice to incorporate perception subjectivity to bridge the *semantic gap*. A medical image retrieval framework is proposed in[46] where both global image feature similarity calculation and RF are combined to better associate image features with the visual categories in the database. In [47], both real-world and artificially generated user interaction data is used for semantic

image clustering. Recently there are also studies[48, 49] on using different information, such as image excerpt and click-through data to help improve the feedback relevancy judgements.

It can be seen from the previous examples that most of the content-based web image retrieval systems require users to have at least a basic understanding of image features. Its Query-By-Example(QBE) strategy involves tedious interaction with users to specify and train the correct retrievals. The users must specify the important image features with their expected values. This requirement reflects the fact that in these systems, there is still a *semantic gap* between the image features and human perceptual concepts. Therefore, the system need users' expert knowledge to bridge the gap. The question of how to extract semantically meaningful image content from low-level features is still an open issue.

2.1.3 A Combination of Text and Content Approach for Image Retrieval

Researchers have tried to look beyond these single-modality approaches and combine text information and image features together to bridge the *semantic gap*. Direct integration of keywords and image features are proposed. Photobook[50] provides a set of tools for image searching and retrieval. It uses both text annotations and low-level features of image content, which reduces images to a small set of perceptually significant coefficients. VisualSEEk[51] is a text-and-image feature based search engine. The spatial correlations of image regions and visual features are extracted from images and further utilized in image retrieval. Users can specify the color, texture and the spatial relationship of image regions. Then the system tries to match the query with images from the database. Some earlier systems proposed in [52–54] try to combine the high-level text information with low-level image features. WebMars[55] aims to retrieval HTML documents using both text and image information. Its users can submit either text or image query, which is then refined to search for media information. The novelty of this work comes from a tree structure which combines different media types together, and helps narrow the gap between the medias. Benitez et al.[56] use WordNet for word disambiguation and

combine the text analysis result with image features to get a 3-15% improvement on image retrieval result.

It has been proved by literature[52–55, 57, 58] that web image retrieval benefit most from a utilization of all available information(both text and image). We argue that even though a combination of text and image has established certain connection between human concepts and image features, none of these works consider prior knowledge, for example, the domain knowledge of the target object, as an important cue to solve the problem of semantic interpretation in image retrieval. A little flavor of "mixture of text and image" does not significantly alter the basic structure of those earlier single-modality approaches.

2.2 Semantic Image Retrieval

In the context of semantic image retrieval, concept-based image retrieval[59] has shown some promising results, where semantic concepts are derived directly from low-level features through various means. The interesting issue is to bridge the *semantic gap* and help machine understand the media content like human-beings do at a higher abstraction level. The new intelligent interface should be supported with general knowledge and basic inference mechanism. Towards a better multi-modality based semantic way for image retrieval, recent works are trying to derive semantic content from both image features and surrounding text. Such efforts can be classified into four basic categories: (1) Mining implicit semantics through machine learning; (2) Image retrieval with customized semantic template; (3) Image retrieval with concept ontology; (4) Web image retrieval by fusion of text and image information. We put web image retrieval as a separate category, as it is different from the others due to the heterogeneity of images and additional text information. According to this categorization, the following subsections discuss some existing works of semantic image retrieval.

2.2.1 Implicit Semantics Mining

Machine learning methods have been widely adopted to mine semantic features from low-level features[60–65]. The learning process takes the low-level features as input, and

associates the image features to high-level human concept through prediction. Since there is no explicit concept ontology or framework defined, the training and learning process is considered as implicit semantic relation mining. In image retrieval tasks, we are most interested in finding the association between concepts and features. For retrieval and annotation purposes, image are finally classified/annotated by semantic concepts. Therefore supervised learning, which generates explicit prediction of outcome, such as high-level concept, is applicable to the retrieval tasks.

Support vector machine(SVM)[66] is one commonly used supervised learning method. Given training data, SVM views the input data as a vector space, and separates the space using hyper-plane. The principle is that, by finding the optimal separating plane, the separated data clusters have the maximum margin between them. Training samples are attached with class labels and treated as *support vectors*. In image retrieval, the training samples are the low-level features. Traditional supervised learning techniques meet new challenges in image retrieval fields, as the annotation of the ground truth for training could be both tedious and time consuming due to the need of large number of training samples. To solve this problem, Shi et al. [67] propose an adaptive segmentation method to divide images into meaningful units. Part of the units are put as training data to SVM to find optimal classifier models for 23 high-level concepts. Based on the derived classifiers, test units are attached with prediction results as one of the concepts. In [67, 68], a bootstrapping approach is proposed to annotate large testing examples by using small set of training examples. The proposed training process also utilizes segmented image units as input. The authors use two independent segmentations and extract two sets of features to support a co-training process. Therefore, the proposed methods outperform traditional supervised approaches in a way that only a small number of labelled training images are required. A similar approach is applied to WWW image retrieval in [69], where the HTML text information is combined with visual information to co-train two "orthogonal" classifiers. The result based on a 5,000 WWW image database proves that comparable retrieval performance to the traditional supervised learning methods could be achieved with less input.

Besides SVM, there are other learning methods. [62] uses a different way of light supervised training by including text cues in the training stage. Latent Dirichlet Allocation(LDA) is used to discover the latent topics from web pages. A linear and equal

weight combination of four independent scores from both text cues and visual cues are used to retrieve *animal* image on the web. In their work, local text information on the web page are considered to be important information. Bag of words model is used to represent the nearby text information. A set of highly-related words is associated with each image topic. Visual cues consist of shape, color and texture features of images. For cue combination the authors use a linear combination with equal weight for each cue. The integration of web page text information is proved to be effective. Domain knowledge can also be represented by the causal relationships and independence between nodes in the networks[70]. Bayesian network is one such network which contains directed, acyclic graphs and obtains domain knowledge through prior training. After training, the Bayesian network can be directly applied for inference. Neural network classifiers are trained to classify unlabelled image regions to semantic classes[63]. The experiment is done on a data set of natural scene images. The images are segmented into small parts and low-level features are extracted to represent each part. Later these low-level features are passed to the neural network and after the transition, the output are transformed into the concepts. Another work[65] also uses Bayesian network to classify indoor and outdoor images. Y. Gao et al.[71] propose Bayesian inference to measure the concept similarity. In their ontology model, nodes are classified into concept nodes and content nodes. Concept nodes represent the semantics of a whole image. Content nodes represent the salient object concepts in the image. Both image concept and salient object concept information are obtained from the image data set, where each image is labelled with its dominant salient object information. The images are further grouped and kept in different folders, whose names contain information of image concept. Both the semantic relationships within the concept nodes and content nodes are extracted from the holonymy/hypernymy relationships in WordNet. The links between concept nodes and content nodes are obtained through the salient object detection phrase. Once a salient object is found in certain image concept, a holonymy/meronymy link is assigned to the two concepts. Probabilistic concept reasoning is adopted for ontology reasoning. The matched salient object concepts and image concepts become the final results for image annotation. However, the scalability of the proposed method is a problem because the training of the salient object classifiers depends heavily on manually labelled data. Similar work which also uses Bayesian inference is [72]. The authors use a hybrid model,

which combines Bayesian network uncertainty reasoning techniques with ontology logic inference for image retrieval. Though it only mentions that an external ontology is used in the experiment and no further details are given, it is clear that the ontology only involves keyword concepts. In the first step, inference process is implemented as a direct retrieval of the neighboring nodes of the input keywords. These neighboring nodes are considered as a candidate concept pool. In the second step, Bayesian network is used to find the most relevant concepts among all the candidates and rank them accordingly. It is worth noticing that the authors believe that no relevance of concepts can be obtained in the first step, while it actually can be calculated through the distance of concepts in the concept graph. In the experiment, the test set includes 100 images which are annotated with 15 different keywords. The image retrieval is based on the concepts generated by the hybrid model. The result proves that the proposed method is better than keyword-based retrieval in both recall and precision. Similar to Bayesian network, decision tree[73] is considered as an effective decision support tool, which uses tree structure graph and recursively separates the input attributes into non-overlapping spaces. Decision tree is also used to derive semantic features in [61, 74, 75]. In these works, the color descriptor attributes traverse through the decision tree from root to leaves and are finally mapped to text descriptions.

These aforementioned works have shown that learning methods can mine implicit semantics from low-level features. However, the applications are limited to very few concepts(usually less than 15 concepts) and to ensure the effectiveness of image features, the experimental images contain homogeneous visual features. Whether implicit semantic mining can perform on more complex image domains is yet to be resolved.

2.2.2 Customized Semantic Template

Besides mining the semantic relationships directly from low-level features, semantic template, which explicitly defines the relations between low-level features and high-level concepts are also proposed in [76]. Templates are used to define a personalized/customized view of concepts and link image features directly to high-level concepts. The semantic visual template is generated manually to define the spatial and temporal constraints with weight of effective object features. The template is optimized through user interaction

to reflect the most appropriate model according to users' understanding. This kind of approaches also require certain degree of relevancy feedback. Instead of using manually defined semantic template, another template based approach[77] automates the template generation. The template is defined as concepts and feature centroids with corresponding weights. Once the template is obtained through image feature training, the retrieval process starts with finding the proper template, and then uses the feature centroids and weights in the template to find matched images.

There is variations of semantic templates proposed in [78], where composite region templates(CRT) is used to represent the semantics in images. Homogeneous color regions are considered as meaningful segmentations and the images are represented by counting the frequencies of segmentations in certain CRTs. The CRTs are collected together as a knowledge base. Unknown images are identified through image region spatial information matching of the CRTs in the knowledge base.

The semantic templates have similar disadvantages to the machine learning methods, as no formal knowledge definition is given. Once the template is defined, there is no further inference mechanism to derive additional facts. Meanwhile, the discriminative power of the templates relies heavily on the training images.

2.2.3 Concept Ontology

Ontology[1] provides a shared and common understanding of a domain. It helps knowledge communication between people as well as across application systems. Originally ontology engineering is extensively studied in the field of artificial intelligence and knowledge representation[79]. Many ontology engineering methodologies[80–82] are proposed together with a set of ontology development tools[83, 84]. Some of the representative manually built ontologies includes[4, 85]. Later, in today's ubiquitous use of the web, ontology plays an important role in presenting, storing, and reasoning with structured information from the web. It is employed in applications related to knowledge management, data mining, natural language processing, and information retrieval. The most important difference between ontology and other methods of knowledge representation is that ontology could be machine readable. Therefore, ontology provides solution to bridge the *semantic gap* between the low-level features and the high-level human concepts.

Attempts[3, 86, 87] have been made to validate the use of an ontology in practice to bridge this *semantic gap*. The need of a machine-understandable representation of multimedia content descriptions is addressed earlier in [86]. Based on the MPEG-7 standards, the authors manually built an ontology in RDF schema. This ontology is built purely from MPEG-7 descriptors and defines the semantic relationships between various low-level descriptors. MPEG-7 visual descriptors are put into different classes: Color, Texture, Motion and Shape. Within these classes, hierarchical structure is constructed. For instance, DCD, scalable color descriptor(SCD) and CLD are defined as subclasses under the color descriptor, which is further defined as a subclass of resource under RDF schema. The proposed ontology can be easily reused and integrated with meta-data descriptions from other domains, such as museum, educational, and medical. A set of given image meta-data is used in [3] to improve their original content-based image retrieval system. The techniques using the meta-data include a minimum-normalization over the semantic elements in each image, an object significance calculation, and a selective use of WordNet concepts to find the category of each semantic object. A step forward to middle-level ontology based on machine learning is discussed in [87]. The authors set up an image retrieval web service, whose essential part is a three-level ontology which associates the image content and human-understandable concept. The lowest level is in accordance with the original MPEG-7 ontology except that the MPEG-7 image descriptors involved are limited to DCD, CLD and Edge Histogram Descriptor(EHD). The remaining two levels define higher level concepts and their properties which are organized in a hierarchical structure. WordNet is used to extend the upper-level ontology to bridge the gap between user query and defined concepts in the ontology. The final retrieval performance is proved to be fast and effective. Though this ontology moves a step forward from the original MPEG-7 descriptor ontology to a higher level concept ontology, all the concepts are built from low-level features and the semantic elements are only from single modality. Besides, the matchmaking process is a fuzzy mapping from the concept to the image descriptors. No extra information can be obtained from this mapping process.

Other earlier works involve manually constructing ontology for each image[88, 89]. In [88], ontology is understood as an interconnected browsing space. This space is constructed according to people's common-sense in real life. A thesaurus is first selected and

activity contexts are defined to link the terms in the thesaurus. The activity contexts involved are carefully selected to ensure only closely related terms in certain domain are included. It takes months for knowledge engineers to finish the construction. Benefiting from the closely interconnected terms, end-users of the system can have a bigger browsing space as well as easy access to the collections of photographs. In [89], the authors use ontology to promote the retrieval performance for image exhibitions. The ontology in their work is specially designed for the photograph exhibition. Every image in the database is associated with instances of the ontology. Both the ontology construction and image annotation are done manually. Ontology is extended along with the annotation process. For instance, if certain image should be annotated with a famous person, who has not been defined in the current ontology, an instance under the *person* class is then created and the annotation of the image continues. The final system allows users to browse and navigate through the ontology for more accurate image retrieval. The system is also able to give recommendation of relevant images to users' query.

More improved ontology-based works are found in special domains like medial image. [90, 91] combine text and visual ontology for medical image retrieval. In [90], the retrieval system requires an input of both sample images and text descriptions of natural language. Both text information and image features are processed separately to retrieve medical images. For text part, an existing ontology MeSH(Medical Subject Headings) is used as external knowledge resource. First, semantic dimensions are defined according to the sub-trees in the ontology. Dimension relevant terms are specified as concepts in the corresponding sub-tree. The mapping is based on the Boolean conjunction of each single dimension matching result. For the image part, 39 previously defined image terms are used for matching. Small regions in images are detected with a confidence value and then represented in a grid pattern. The fusion of the two modalities includes two simple merging techniques based on their performance measure: either the best result of one modality or the average of two results from either modality is taken as the final result. The performance of this fusion method is proved to be better than using either of the two modalities. Similar work in medical image retrieval domain is also discussed in [91]. The important difference is that Description Logic(DL) is used as the matching scheme. In this work, Breast Cancer Imaging Ontology(BCIO) is constructed by commonly used

domain vocabulary. With the vocabulary, image content and meta-data are modelled as high-level concepts. Even though the work introduces the DL for concept inference, the construction of ontology, annotation of image, and detection of Region of Interests(ROI) are manually done, which is a limitation for further extension to larger image database. Animal image retrieval system driven by ontology is discussed in [92]. The constructed ontology is a straight high-level concept hierarchy extracted from WordNet. In the final system, the ontology allows a concept navigation through the concept hierarchy. For instance, when the input query is *dog*, the system extends the concept to most common sub concepts, such as *pug*, *papillon*, and *collie*. Then the users are allowed to navigate according to their interest. However, no inference or semantic matchmaking is involved in the whole process.

Ontologies are effective in providing structural concepts. However, their capabilities are not fully exploited without domain expert knowledge. Domain knowledge is of special importance in high-level modality information matchmaking and can provide more accurate result. Most of the current works apply probabilistic methods to high-level information while few works have explored the domain knowledge and use semantic matchmaking to classify the images. Some possible reasons are: (1) Expert knowledge for most image domains are out of researchers' specialty and not easily available; (2) Domain knowledge is most useful in reasoning over related image categories while most of the current researches are based on sparse image categories.

2.2.4 Image Retrieval based on Multi-Modality Information

Semantic oriented approaches as well as concept ontology are especially useful to find semantics for web images where information from both text and image modalities are available. The surrounding text information provides rich source to extract high-level concept information. On the other hand, the noisy web environment makes it necessary to find an effective way for information retrieval. In [93], the authors discover that most web users have difficulty in determining explicitly the semantics of the web pages. This is due to the fact that most web pages do not have just one particular semantic concept. In order to enable effective web information retrieval, the authors propose techniques to extract semantic coherency from global web page features and users' browsing path

information. The global page features include text features, structural features, and image features. Matrices are formed to combine these features with the users' browsing path information. The authors use latent semantic analysis to reduce the dimension of the feature space. The semantics of the web page is defined as a subset of points in the feature space. Thus the dimension reduction process reveals the semantics of web page gradually. In [94], the authors propose a web image retrieval result clustering based on salient image region pattern extraction. At first the web image is segmented into content region and context region. Given the assumption that the retrieved images may have similar content, the salient image phrases are mainly extracted from context regions. A regression model is used for training process. The approach is shown to be better than traditional image content feature clustering. A web image clustering system is proposed based on the improvement of some existing web page search result clustering algorithms[95]. As additional features are introduced by the proposed algorithm, the resulting image clusters are labelled with semantic tag. Clustering is done on the whole data set rather than part of data set. Moreover, the efficiency could be improved provided that the system has its own supporting image search engine to replace the current external *Picsearch*. The user-oriented algorithm evaluation shows that users find the proposed system more satisfactory than Google image search. An extension work[96] adopts the VisualSEEk tool[51] for large scale web image and video search. Automated agents collect all the data from the web. Both text and visual feature information are processed and cataloged for further retrieval. Text features are mainly drawn from the URL links of images and videos. The substrings of the URL links are chopped and attached with the images/videos as annotation terms. From an aggregation of the terms, a subset of key-terms are manually selected according to their frequency and descriptive power. The idea of *subject taxonomy* is proposed to arrange key-terms according to is-a hierarchy. For visual information, color histograms are adopted to assess the similarity of different images and videos. To catalogue all the data, each image or video is assigned with an information table, which includes both text and visual information. Later the system allows searching by both key-term query and content-based visual query. From the experiment the system is proved to be robust and effective in cataloging visual information on the web. The result further proves that by introducing a term/concept hierarchical structure, the visual information

can be more effectively organized. However, the work leaves questions on the efficiency and limitation of the manually built hierarchy, as well as the limited visual features used. In [97], the authors discuss the approach of combining the detectors with the hierarchical structured concepts from WordNet ontology while successfully eliminating the global inconsistency of concepts within different ontology branches. In this semantic space, the metric that characterize the concept WUP[98] distances between pairs of concepts includes is-a relationship(hypernymy/hyponymy), while other researches have found that incorporating part-of(holonymy) and member-of (meronymy) relationships enables better reasoning. When it comes to web image research which deals with rich lexical information, it is necessary to incorporate more than the aforementioned relationships. In the multi-modality fusion part, the fusion in the paper is performed in a in-series style: The ontology-enriched semantic space(OSS) first finds the matched concept(image detector), and then uses the image concept detector for image classification. If the OSS gives false alarm detectors, it will be difficult for the image concept detector to make up for the mistake and generate positive result.

These works mainly focus on finding the semantics of a group of images. The emphasis is still very much on region-level image patterns. The rich text information in the attached web page is usually ignored or only partially utilized to help the image retrieval. If we can fuse the information from both the text and image modalities together, the results from both parties should be mutual reinforced in an integrated way.

2.3 Summary

In this chapter, a literature review in the field of image retrieval is given. We start from an overview of traditional image retrieval, to various approaches of semantic image retrieval.

In traditional image retrieval, basic approaches based on either text features or image features are introduced. These approaches serve as a background to provide basic ideas of the commonly used text and image features for image retrieval together with their related processing techniques. Methods based on either single-modality information suffer from certain limitations. Text information could be misleading, especially in a noisy retrieval

environment. Features extracted from keywords could be ambiguous. On the other hand, the *semantic gap* is always an open issue in content-based image retrievals which mainly rely on image features. To go beyond the single-modality based approaches, there are also attempts to combine information from both sides, especially in the field of web image retrieval, where both text and image information becomes important cues. Even though improvement has been brought by information fusion, there is no prior knowledge involved in these approaches. The *semantic gap*, which exists between the low-level features and human concepts, should be bridged through concept-based approaches, which include knowledge construction and inference phases.

In the field of semantic image retrieval, works are classified into implicit semantic mining, customized semantic template, concept ontology and web image retrieval. For implicit semantic mining, associations between image features and human concepts are mined through supervised learning methods. However, the associations are not explicitly defined. Because of the manual annotation and inefficient training process, the concepts involved in these experiments are usually limited. The applicable domain can not be extended to larger concept set. Different from the implicit semantic mining, semantics are explicitly defined in the customized semantic templates. These templates are defined to reflect a customized view of the links between concepts and image features. The construction of the templates follows the basic idea of knowledge definition. The disadvantage is that the knowledge is not formally defined. And no inference mechanism is used to derive new knowledge. Following the semantic template methods, researchers start to build concept ontologies, which clearly define concepts and relations in different domains according to domain knowledge. Meanwhile, semantic matchmaking is introduced for concept inference. Most of these concept ontologies are manually defined, which lead to an open issue of construction efficiency and scalability. Given the benefits brought by concept ontology, the question becomes how to effectively construct the ontology and apply it in real applications.

Web image retrieval is categorized separately in the applications as this is a special field where ontology can fully utilize the information from different modalities. A review is given in semantic related web image retrieval works. These works put much emphasis on the image features, and ignore the rich text information on the web pages. As a result, only simple hierarchical relations between concepts are used to aid web image retrieval.

In review of existing works, it can be seen that few attempts have been made to build and apply multi-modality ontology in the field of image retrieval. Hence, we focus on the research in web image retrieval, where there is need for information fusion from both text and image modalities. Methods are proposed for effective multi-modality concept ontology construction. Further discussion about the ontology application with various semantic matchmaking methods is also included.

Chapter 3

Multi-Modality Ontology Construction and Related Experimental Result

In this chapter details of multi-modality construction methods for web image retrieval are described step by step. Ontology-based approaches are not totally new in the field of multimedia retrieval field. However, our multi-modality approach is one of the earliest attempts to integrate information from different modalities and a multi-modality model has been applied to a complex domain.

In order to develop the model, the following questions need to be answered: (1) How to find the proper structure to construct an ontology which can integrate information from different modalities; (2) Given an effective ontology structure, how to find source of expert knowledge to support the ontology construction? (3)Given the ontology model and knowledge source, how to implement ontology construction more efficiently? These questions will be answered in the following sections.

To convince readers that our multi-modality ontology is the right step forward, we decided to experiment on the domain of animal kingdom. This domain is believed to be challenging as demonstrated by images depict animals in a wide range of size, pose, configurations and appearances. An initial data sets of 4,000 web images is set up for experiment. Based on the manually annotated ground truth, the image content and text information is analyzed. The multi-modality ontology model is built up. The retrieval results are given as experiment results. Results show that even close animal species which

share similar visual appearances can be classified. The hidden relationships between animal classes can also be inferred from the *animal* family graph.

3.1 The Proposed Multi-Modality Ontology Structure

In this section, the structure and construction of the multi-modality ontology, which provides shared semantic interpretation of image contents, is discussed in details. The ontology has three main components, including *animal domain ontology*, *textual description ontology* and *visual description ontology*. The *animal domain ontology* is generated based on a formal animal taxonomy and provides the information of animal relationships and classifications. The *textual description ontology* captures the high-level text information. This information contains narrative animal descriptions and converts the descriptions into classes and properties. The *visual description ontology* is constructed from the visual descriptors which encapsulates low-level image features. In the following parts of this section, issues regarding how to build up the ontology and how the knowledge is maintained in each part of the ontology is discussed. An example of the ontology structure is also provided at the end of this section.

3.1.1 Animal Domain Ontology

An ontology is usually designed by the experts to provide knowledge of certain domain, which is the basis of semantic information. In the initial experimental system, the animal group *canine* is chosen as the target domain. Twenty subspecies belonging to *fox*, *wolf*, *wild dog*, *jackal* under the domain of *canine* have been collected as the experiment subjects. These are four-leg animals sharing similar appearances. A hierarchical tree structure is built for the *canine* ontology based on the formal definition of animal taxonomy. According to domain knowledge, the hyponyms relationship between two concepts is refined as subclass property in this ontology. For example, *fox* is a kind of *canine* (hyponyms), therefore *fox* is defined as a subclass of *canine* in animal domain ontology. The reason why the animal taxonomy is imported as part of the ontology system is because it is necessary to handle the relationships and classifications between animals.

The efficiency of matchmaking between the user's query and the concepts defined in the knowledge base is maximized. Taking the example of image retrieval using the term "wild dog", although there is no text description with words like *wild* or *dog* for a *dhole* image, it will be correctly retrieved, as *dhole* is defined as a subclass of *wild dog*. In this case, animal ontology enables the system to match the user query with not just a single category of animal with matching search term, but also the group or class of animal.

3.1.2 Textual Description Ontology

The text information is a straightforward cue for image retrieval. However, in most cases the text information can often be decomposed into isolated keywords and utilized in query and retrieval. The earlier keyword-based retrieval systems have proved that it is not efficient if the user query is only matched with single keyword. In this case, a large amount of false results will be returned due to polysemy in different context. To solve this problem semantic interpretations should be provided for words based on different context, which can be achieved by combining the domain knowledge. Therefore it is important to build a formal ontology whose domain knowledge can be accepted by both domain experts and end users.

In our experiment, a reliable source which is the BBC Science & Nature Animal category[1] is found to extract the class and relationship definitions. This web site serves as an expert of animal domain in the experiment and provides standard and unified descriptions in various aspects for around 620 animals. A web page snapshot is shown in Figure 3.1. Note that this web site is not irreplaceable as a source of domain knowledge. A real human expert will likely do even better in domain knowledge customization. All the animal description pages are downloaded with their page content parsed. The animal description statements are extracted to facilitate further ontology construction. Most of the general animal descriptions, like scientific name, habitat, distribution and diet, are available. In the next step, the statement contained in the narrative descriptions are manually extracted to build up the ontology. The high-level narrative information of animal descriptions is collected and encapsulated into the classes and properties in the *textual description ontology*. Several classes have been defined like *ScientificName*,

[1] http://www.bbc.co.uk/nature/wildfacts/animals_a_z.shtml

Figure 3.1: A snapshot of BBC Science & Nature Animal category web page

Diet, *Habitat*, *Distribution* and *ColorDescription*, and semantic relationships have been generated to connect different concepts including *hasName*, *hasDiet*, *hasHabitat*, *hasDistribution*, *hasColorDescription*. By defining this ontology, animal concepts are associated with their corresponding domain knowledge. More detailed semantic interpretations are provided for animal keywords, which help image retrieval system identify whether a specific keyword in a document is relevant to user query or not. Once the textual description ontology is created, the generated classes in the *animal domain ontology* will be associated with corresponding text description information by using the properties defined above. It is worth noticing that ontology building is a nontrivial task as every concept and relation needs to be defined. However, it is a piece of straightforward work once the concepts and relationships for text and image content information are decided. Taking the *canine* ontology which involves around 210 concepts and relationships as an example, it takes approximately 3 hours for one person to manually define the ontology. Most of the concepts, like concepts under general concept domain of color and distribution, can be reused once defined. The final textual description ontology encapsulates the high-level descriptions for terms and their correlations into a well-defined structure.

3.1.3 Visual Description Ontology

In the previous subsection, a textual description ontology has been manually defined. However, to apply the semantic matchmaking in image retrieval, having only text information is not sufficient. Loosely-coupled correlations exist between the web images and

text annotation. Text information may not represent the content of image correctly. In this case, the knowledge generated from the low-level features of the images can be utilized to enhance the performance of semantic matchmaking. Different from the normal approach of CBIR, the low-level features cannot be applied in semantic matchmaking directly. A specific knowledge base need to be built, where the terms and correlations are extracted from low-level features. Subsequently, the knowledge generated from high-level text information and low-level features are incorporated and used as knowledge base in semantic matchmaking.

To build this knowledge, a set of terms which are relevant to the image content are first defined. These terms are assumed to be distinguished in low-level features. After that, the knowledge from the low-level features of images is extracted using supervised learning technology. Several low-level features are extracted from the images like color correlogram[18], MPEG-7 color structure descriptor, coocurrence matrix[99] and MPEG-7 edge histogram descriptor which provide the color and texture descriptions for image content. In the current implementation[100, 101], a SVM with radial-based kernel is trained and applied in image classification. One third of the data set is used as training sample. The training set is further divided into 5 parts for cross validation. The class labels used in classification are associated with the terms defined in the knowledge. For example, in *canine* image analysis, images are first classified as *Colorful* or *Grayscale* images. Since *Grayscale* images cannot provide as much low-level information as *Colorful* images can, these two kinds of images need to be distinguished. At the same time, the images are also classified into *photograph* or *Graph* category as different methods will be used to analyze them. After that, for all images that are classified as *Colorful* images, they are classified into *Outdoor* or *Indoor* category. For all *Outdoor* images, they are further classified into *Buildings*, *Humanrelevant* or *Wildlife* category. Finally, the *Wildlife* images are classified according to whether the background contains *Snow*, *Sand*, *Stone*, or *Greenery*. To extract more information for *canine*, the images are classified based on their foreground colors corresponding to four typical *canine* fur colors and *non-canine* fur. After the classification, each image has a set of labels to describe its content, which are to be matched with the concepts defined in the knowledge base. In this way, the low-level features of images are converted to a set of terms which can be utilized by

semantic matchmaking. Therefore the low-level image attributes are incorporated into the high-level text information.

To apply the high-level information extracted from low-level image attributes in semantic matchmaking for image retrieval, the extracted information needs to be incorporated into the knowledge base. Hence, an ontology for describing these visual concepts is created. This visual concept ontology is a component of the proposed animal ontology model. To generate this part of the ontology, each image classification scheme is defined as a class in the ontology and, intuitively, the image categories under this classification scheme are defined as its subclasses. For instance, *ContentType* is defined as a class under visual ontology. *Outdoor* and *Indoor* are defined as the subclasses of *ContentType*, and *Building*, *Humanrelevant* and *Wildlife* are defined as the subclasses of *Outdoor*. The next step is associating the generative classes under animal ontology with corresponding visual concepts. Since most of the relationships between animals and visual concepts are determined by human prior knowledge and it is not easy to be generated automatically, these relationships are manually created according to the descriptions of BBC Science & Nature Animal category described earlier. A complete list of semantic relationships are extracted from low-level features as follows: *hasPixColor*, *hasPixProp*, *hasEnvironment*, *hasContent* and *hasFur*. Thus, for every animal, now there are high-level descriptions which are not only generated from text information but also from low-level image attributes. This will help us to provide more accurate semantic interpretations to high-level human conceptual terms for such an animal domain.

3.1.4 Examples of the Generated Classes

Figure 3.2 provides a structure of the *canine* ontology in our system. In this figure, ellipse and rectangles are used to represent predefined class and generated class, respectively. Some parts of the ontology are omitted in the figure due to the limited space. Two examples *red fox* and *red wolf* are included in this ontology to show how the animal concepts are defined. In the figure, it can be seen that *red fox* and *red wolf* are two generated class under the superclass *canine*. From the textual description ontology, it is clear that *red fox* has distribution in USA whereas *red wolf* has distribution in Asia. From the visual description ontology, it can be seen that the fur color of *red wolf* is

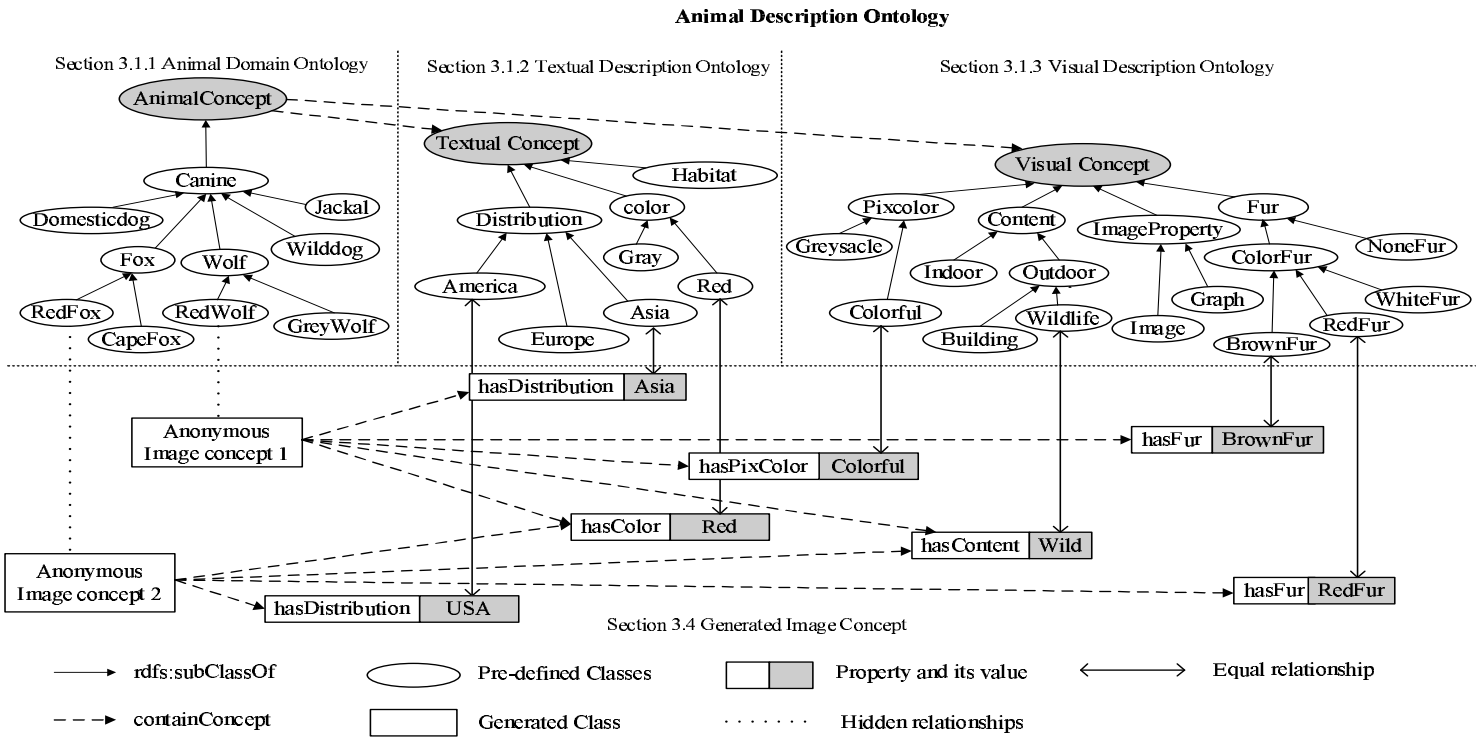


Figure 3.2: Layer structure of the proposed multi-modality ontologies

brown and the fur color of *red fox* is red, even though the name of the two classes share the same keyword of *red*. At the same time, the visual descriptors can help to filter off a substantial amount of inaccurate results. For instance, it can be reasonably inferred from an indoor background that a wild *cape fox* is not likely to exist in the image.

3.2 Adding Wikipedia Semantics for More Scalable Ontology

A multi-modality ontology model has been proposed for image retrieval in the previous section. It can be seen that one main difficulty that hedge against the development of ontology approaches is the extra work required in the construction of an ontology. Scalability is indeed a problem when a single party or a consortium tries to create a whole ontology structure. However, the problem could be solved when existing ontologies or newly created ontologies can be imported and merged seamlessly with other ontologies. The important issue here is to understand and handle the similarities/dissimilarities of concepts existing in the respective ontologies, which is of research interest to relevant groups in artificial intelligence and ontology-related areas.

With the rapid development of Internet, online categories are becoming a good choice as ontology, especially when some popular online categories also provide easy access[102, 103]. By indexing a huge number of web pages/topics, online categories cover most real world objects, activities, news and documents in a timely manner. Besides the hierarchical structure offered by these categories, web page submitters and category indexers also provide more related concepts with varied relationships, which further extend the semantic space. With the evolution of ontology-based applications, finding a proper knowledge source has become an important issue. WordNet[3], which is developed at Princeton University, has been commonly used as a lexical dictionary as it is able to group words into sets of synonyms and further link them by semantic relationships. The structured network makes it easy to derive semantic hierarchies for various domains. Some typical examples which utilize WordNet for object recognition and video search includes[92, 97, 104]. However, WordNet works fine in these experiments as only common concepts such as car, dog, grass, tree, etc. and simple relations(particularly hypernymy

and meronymy) are involved in. In other words, concepts and relationships outside the scope of WordNet will not be captured by the semantic graph or the semantic space constructed based on WordNet. For example, though WordNet elaborates and categorizes general concepts such as "bike" and "dog", it has limited coverage of less popular/more specific concepts like"mountain bike" or "bush dog". This limitation decides that WordNet will only work on some general concept domains, which certainly conflicts the idea of using ontology to share knowledge for extensive domains. WordNet is also not frequently updated to natural language vocabulary which changes almost everyday. As a result, it is not able to work on developing domains with advanced topics and concepts. In comparison to WordNet, Wikipedia certainly contains more information. For the experiment case, only 12 out of the 20 class names are covered by WordNet. Class names such as *African wild dog, bat-eared fox, black jackal, bush dog, cape fox, Ethiopian wolf, fennec fox, golden jackal* are all missing from WordNet. Such limitations make WordNet an incomplete appropriate resource for ontology learning. On the contrary, Wikipedia is more suitable for this task. The total number of words has reached 2 million and it keeps increasing significantly daily. It can cover almost all the relevant concepts in our experiment.

To differentiate the Wikipedia-based ontology created in this section, hereafter the manually built text ontology is defined as ManuOnto. The manually built multi-modality ontology is defined as ManuMMOnto. It is shown in previous section that both ManuOnto and ManuMMOnto can effectively help machine understand multimedia in a better way. This section describes our preliminary attempt to automatically construct large-scale multi-modality ontology, which is defined as AutoMMOnto, for web image classification. In particular, to enable the automation of text ontology construction, we take advantage of both structural and content features of Wikipedia, and formalize real world objects in terms of concepts and relationships. This process is named as Wikipedia2Onto. A proposed hierarchy is built with not only hyponymy(is-a) or meronymy(part-of) relationships, but also more real-life relationships. For visual part, classifiers are trained according to both global and local features, and middle-level concepts are generated from the training images. Verifying the correctness and usefulness of AutoOnto is necessary, a variant of the association rule mining algorithm is further developed to refine

the built ontology. As a result, the semantic concept hierarchy of the generated ontology is consistent with real world knowledge and can be used to map text information on the web page to detect semantic concepts.

Before we go to the details of the proposed Wikipedia based ontology construction method, the major contribution in this section is briefly summarized as follows: A method Wikipedia2Onto is proposed to build large scale concept ontology from Wikipedia in a cost effective way. The generated ontology is able to extract additional information from the web pages and increase the concept detection accuracy. An association rule mining algorithm is also proposed in Section 3.2.3 to refine relationships in the ontology. The resulting ontology is more concise with higher precision.

3.2.1 Wikipedia Concepts and Structure

Wikipedia is by far the biggest online free encyclopedia. It provides definitions for more than 2 million words and phrase concepts. This number is still growing as Wikipedia is based on online collaborative work and anyone can freely access, create and edit the page content of each concept. This open feature makes Wikipedia an up-to-date knowledge source, where even the latest concepts can be found. It also covers many concepts which are not commonly used and included in other electronic lexical dictionaries. In the following subsections, some of Wikipedia's features which make it suitable for ontology construction will be introduced.

3.2.1.1 Wikipedia Category

The underlying structure of Wikipedia can be described in two network graphs: category graph and article graph. In both graphs, nodes represent articles and edges represent links between articles. Basically, all the Wikipedia web pages are put into a subject category according to general knowledge. This structure is depicted as the category graph which has been proved to be a scale-free, small world graph by graph-theoretic analysis[105]. The category graph is formed following the taxonomy of concepts. Therefore, the links in category graph indicate either is-a or part-of relationships between the two connected concepts (a sample of the category graph is given in Figure 3.3). In this sense, the semantic relationships provided by the category graph are quite similar to the

relationships provided by WordNet. When referring to specific article, the Wikipedia classification is listed in a separate "Categories" section. Besides the category graph, there is also an article graph which indicates the cross-references between Wikipedia web pages. In particular, the articles are nodes of the graph, which are hyperlinked to corresponding Wikipedia articles. These links indicate a direct semantic relationship between the two connected concepts. Compared with WordNet which mainly organizes word concepts according to synset, Wikipedia category provides a more formal classification of concepts. As a result, the extracted concepts and relationships are closer to an ideal ontology with various semantic relationships.

3.2.1.2 Wikipedia Web Page

In Wikipedia, each web page defines one concept according to general knowledge. Disambiguation is removed by separating different senses in difference web pages. The searching in Wikipedia is straightforward as each web page has already been associated with the keywords. In most cases, the page title is the indexed keywords. The text information on the web page is divided into sections. Each section describes one aspect of the concept in details. Taking the concept *Aardwolf* as an example (see Figure 3.4), the main web page content includes physical characteristics, distribution and habitat, behavior, and interaction with humans. From the view of concepts, sections are connect to the main concept with semantic relationships depicted as section titles. A concept graph is easily drawn from this web page content structure. On the right side the web page also provides a section of scientific classification, which lists the zoology taxonomy of the animal. By integrating different concepts under the same domain *Animalia*, a big hierarchy picture can be easily constructed with the concepts positioned under corresponding branches. Compared to our manually built *Animal Domain Ontology*, the hierarchy generated from Wikipedia scientific classification is more formally defined, and is considered to contain rigid domain information.

3.2.2 Knowledge Extraction from Wikipedia Articles

In this section the construction of our multi-modality ontology is discussed. Similar to the previous manually construction process, the automatic process Wikipedia2Onto includes three steps. Firstly, the key concepts in the animal domain and the taxonomic

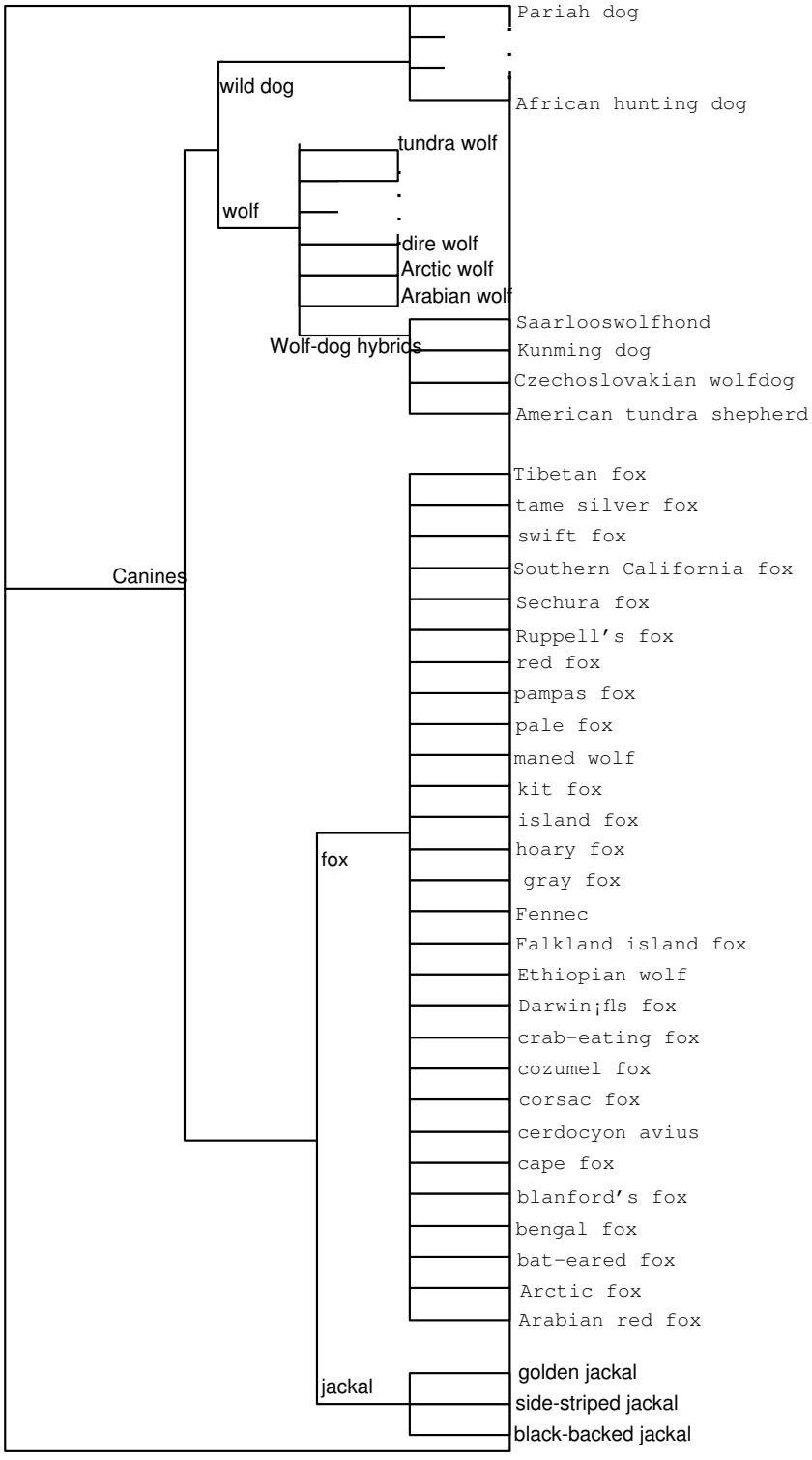
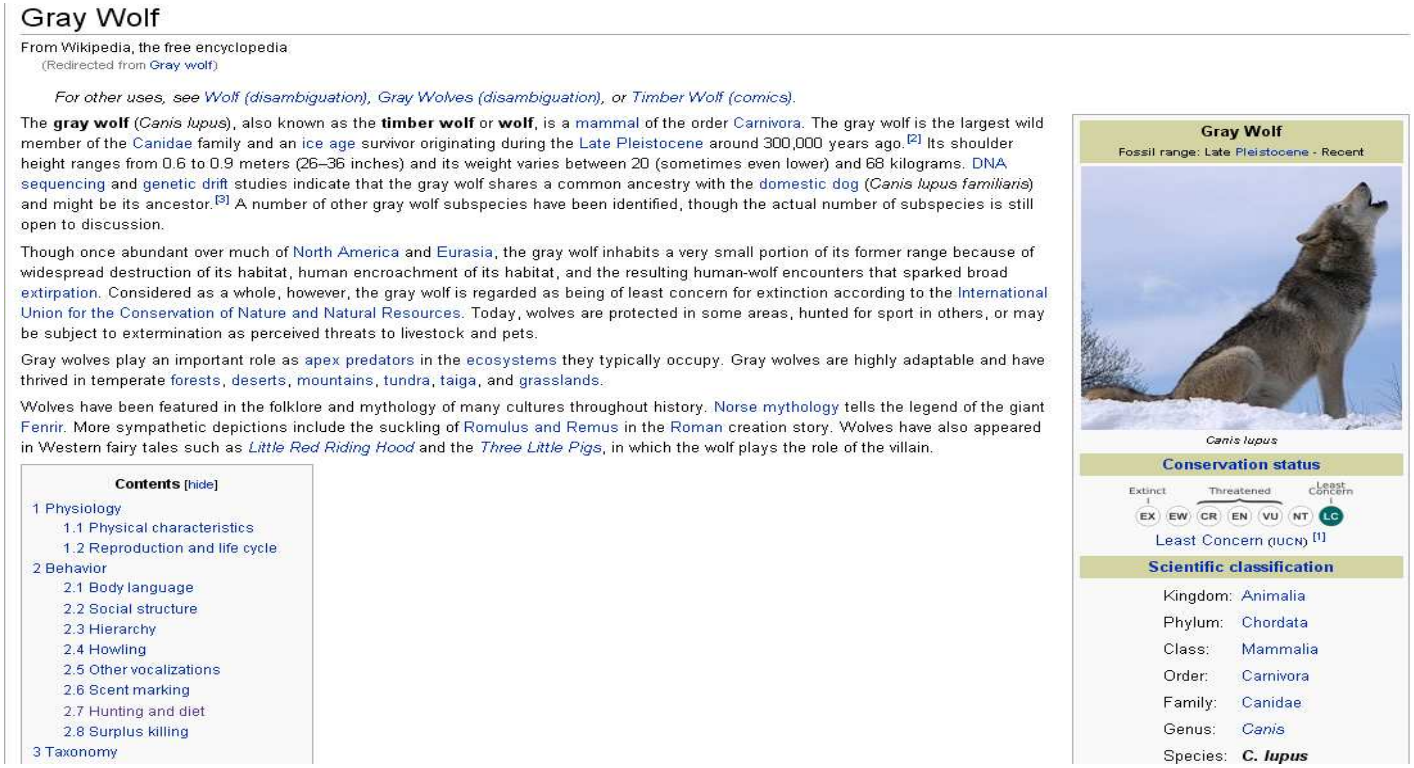


Figure 3.3: An example of the *Canines* wikipedia category.



Gray Wolf

From Wikipedia, the free encyclopedia
(Redirected from Gray wolf)

For other uses, see Wolf (disambiguation), Gray Wolves (disambiguation), or Timber Wolf (comics).

The **gray wolf** (*Canis lupus*), also known as the **timber wolf** or **wolf**, is a mammal of the order Carnivora. The gray wolf is the largest wild member of the Canidae family and an ice age survivor originating during the Late Pleistocene around 300,000 years ago.[2] Its shoulder height ranges from 0.6 to 0.9 meters (26–36 inches) and its weight varies between 20 (sometimes even lower) and 68 kilograms. DNA sequencing and genetic drift studies indicate that the gray wolf shares a common ancestry with the domestic dog (*Canis lupus familiaris*) and might be its ancestor.[3] A number of other gray wolf subspecies have been identified, though the actual number of subspecies is still open to discussion.

Though once abundant over much of North America and Eurasia, the gray wolf inhabits a very small portion of its former range because of widespread destruction of its habitat, human encroachment of its habitat, and the resulting human-wolf encounters that sparked broad extirpation. Considered as a whole, however, the gray wolf is regarded as being of least concern for extinction according to the International Union for the Conservation of Nature and Natural Resources. Today, wolves are protected in some areas, hunted for sport in others, or may be subject to extermination as perceived threats to livestock and pets.

Gray wolves play an important role as apex predators in the ecosystems they typically occupy. Gray wolves are highly adaptable and have thrived in temperate forests, deserts, mountains, tundra, taiga, and grasslands.

Wolves have been featured in the folklore and mythology of many cultures throughout history. Norse mythology tells the legend of the giant Fenrir. More sympathetic depictions include the suckling of Romulus and Remus in the Roman creation story. Wolves have also appeared in Western fairy tales such as *Little Red Riding Hood* and the *Three Little Pigs*, in which the wolf plays the role of the villain.

Contents [hide]

1 Physiology
- 1.1 Physical characteristics
- 1.2 Reproduction and life cycle

2 Behavior
- 2.1 Body language
- 2.2 Social structure
- 2.3 Hierarchy
- 2.4 Howling
- 2.5 Other vocalizations
- 2.6 Scent marking
- 2.7 Hunting and diet
- 2.8 Surplus killing

3 Taxonomy

Gray Wolf
Fossil range: Late Pleistocene - Recent

Canis lupus

Conservation status

Extinct | Threatened | Least Concern
EX EW CR EN VU NT LC

Least Concern (IUCN) [1]

Scientific classification

Kingdom: Animalia
Phylum: Chordata
Class: Mammalia
Order: Carnivora
Family: Canidae
Genus: *Canis*
Species: ***C. lupus***

(define-concept concpet_gray_wolf(or SomeAnimal(all hasName (or gray_wolf timber_wolf wolf))(all hasDistribution (or Canada Ireland Kazakhstan the_Middle_East North_Anerica Russia Europe the_United_States India Asia Finland)) (all hasDiet (or Herbivore Coyote American_Bison Deer Caribou Moose Yak Ungulate Rodent))))

Figure 3.4: An example of Wikipedia web page with corresponding extracted concept

relations are firstly extracted from Wikipedia. Then, the narrative descriptions of particular animals, including relevant concepts and non-taxonomic relations, are extracted. Finally, the visual descriptions of each concept are added. Note that the XML corpus provided by Wikipedia is not used directly for construction. Instead, a web page crawler is used to download relevant concept web pages before ontology building starts. Such an approach makes it more flexible to build ontology for specific domain. Meanwhile, a dynamic connection to Wikipedia can update the concepts as Wikipedia web pages are edited from time to time.

3.2.2.1 Wikipedia2Onto: Key Concepts and Taxonomic Relations Extraction

Wikipedia has provided an entire category of many meaningful concepts, which is formed according to hypernymy relationships between concepts. In other words, Wikipedia category provides taxonomy of general concepts in natural language, which is much more precise than the in-door manually built one. Therefore, the *Animal Domain Ontology*, which is used to describe the taxonomy information of animal concepts, can be directly obtained from Wikipedia category. However, as the Wikipedia concepts under *animal* domain have some special content features, the scientific classification entry on each concept page is used as a shortcut. This entry provides animal taxonomy in a top-down manner, from *Kingdom, Phylum, Class, Order, Family, Subfamily, Genus* to *Species*. Then the hierarchical structure is extracted from this entry and form the *Animal Domain Ontology*. For example, *Phylum* is defined as a sub-class of *Kingdom*, while *Class* is defined as a subclass of *Phylum*. Since the ontology is only defined for general web image classification, the structure construction stops at *Family* level and does not go beyond *Subfamily*. Taking *Aardwolf* as an example, this concept belongs to the family of *Hyaenidae*. So when an input query suggests a concept of *Hyaenidae*, *Aardwolf* will also be considered as a matched concept.

3.2.2.2 Wikipedia2Onto: Narrative Descriptions Extraction

In the definition of ontology, what is alluded to but not formally stated is the modelling of concept relationships. In order to show that ontology is more than a set of related

keywords, it is necessary to prove that every concept in the ontology is different from a plain word. It should be understood as concepts supported by structures. When building the *Textual Description Ontology*, the main concerns are twofold: an ontology, which depicts the real world, should contain more descriptive concepts and relationships. These relationships convey general knowledge according to domain knowledge. On the other hand, the related concepts should contain a hierarchical structure, so that in concept inference phase, additional facts could be generated. Here is an example to illustrate the above concerns: *South Africa* is where the species *cape fox* lives. Therefore, *South Africa* is linked to *cape fox* with a named relationship *hasDistribution*. Given two other relations *Zimbabwe* is a part of *South Africa* and *South Africa* is part of *Africa*, one could reasonably infer that *cape fox* can also be found in *Zimbabwe*. This possibility increases when additional information matches. Therefore, the first step is to find all the important terms. Some pre-process includes crawling Wikipedia web page of relevant concept and using HTML parser to filter irrelevant HTML codes. After that, the web page content is analyzed to extract useful concepts and relationships. It is worth noticing that at the beginning of each web page, where a short paragraph is given as a brief introduction of the particular concept, some words are emboldened as alternative name or synonymy to the main concept. By extracting these words, a synonymous set is first constructed for the original concept. A *hasName* relationship is used to link it to the original concept. This relationship extends the naming information. In the next step, by analysing HTML tags of document title, section title and links to other pages, the location of the title of each section is found. Before the details of the section content is analyzed, the section title is examined to see if it contains relevant semantic relationships, like information about Distribution, Habitat, Diet, etc.. Once the relevant keywords are discovered in the section title, we look into the details of the section and find candidate concepts for that particular relationship. Candidate concepts are defined as those that have their own Wikipedia web pages. For the normal plain text on the Wikipedia web page, it is believed to be of trivial importance, thus has less contribution to the concept detection. Based on this assumption, a set of concepts are extracted from the section for each relationship. While not all the candidate concepts are correct, an association rule mining is discussed later to improve the accuracy of the generated ontology.

After the relationships and related concepts are collected, further hierarchical is constructed among all the concepts. This step is done based on the Wikipedia category structure, which offers a systematic categorization of all the concepts. The category information is listed as a separate section at the bottom of each Wikipedia web page. In most cases one Wikipedia concept belongs to several categories, some of which serve for Wikipedia administration purposes, such as *Wikipedia administration*. These categories are removed and the rest categories which follow different categorical classification are kept. For each related category, we move one step further to find its parent category. In the current implementation five iterations are done, and construct a hierarchical structure of five levels for each concept. This step helps to formulate the information and introduce more structured concepts on top of the current ontology. To evaluate the performance of the proposed ontology system with other text aware methods, the text processing part of [62] are followed and LDA is used to find 10 latent topics from the web page text. The top 20 words are taken from each latent topic as the topic representation. However, the resulting clusters of words do not show explicit semantic meanings. The reason is presumed to be the relative smaller size of text corpus. Therefore, the ontology approach is more appropriate on the median-sized experimental data set.

3.2.2.3 Visual Descriptions Extractions

In this section the visual description features for AutoMMOnto are discussed. The experimental data set is still a median size collection of 4,000 animal web images, together with the corresponding web pages. More specifically, the data set contains 20 animal categories under the domain of *canine*. For the experiment, recognition techniques are used to build a visual vocabulary and train classifiers using SVM. We do not propose our own object detection techniques as these techniques have been extensively discussed in computer vision researches. The main objective is to show that the object detection techniques whose superiority has been proved in the latest researches[106] are followed. Harris-Laplace detector[32] are first used to detect interest points. Then SIFT descriptor is used to represent the shape information around the interest point. A 20 by 20 image patch around the centre of the interest point is generated to extract *opponent angle* color descriptor. The color descriptor is combined with SIFT descriptor. The final descriptor

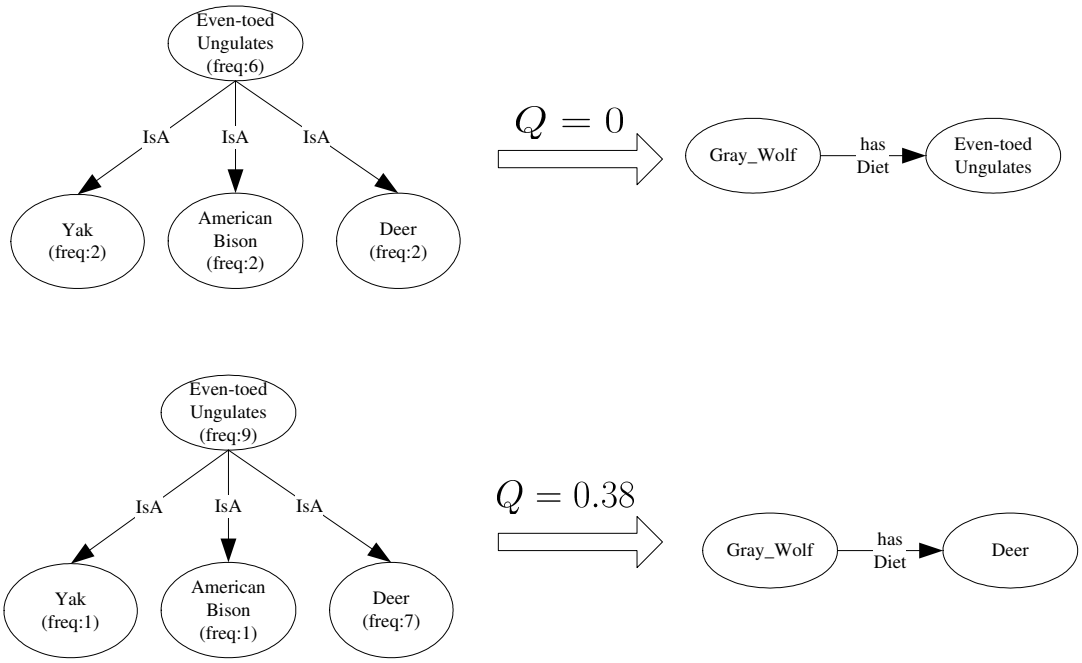


Figure 3.5: Illustration of the relationships pruning.

is a vector of dimension 164, where 128 dimensions are from SIFT descriptor and 36 dimensions are from opponent angle descriptor. In addition, a shift along the horizontal or vertical axis is made when image boundary meets the patch range. A vocabulary of 500 visual words are built based on k-means clustering result of feature vectors from all images. For each image in the data set, a histogram of visual words is calculated and then each image is represented by a vector whose dimension is 500. After feature space construction, one third of the data set is used as training sample. The training set is further divided into 5 parts for cross validation. After training, the relations between image feature concepts and the animal concepts are obtained.

3.2.3 Semantic Relationships Pruning with Association Rule Mining

After the automatic ontology construction, association rule mining is used to refine the initial ontology. Wikipedia is an online collaborative work and the content is maintained by users, therefore a certain level of inherent noise must be expected. Those relations other than hypernymy and meronymy relations are extracted and incorporated into the

ontology by analyzing the Wikipedia web pages with text analysis techniques. A small set of inaccurate relations could be extracted either due to the complexity or correctness of the texts and the strategy we used for relation extraction. For association rule mining[107], the research has evolved from a flat structure with a fixed support value to variances that include complex tree or graph structure with different support values. In order to enhance the correctness of semantic relations extracted, a variance of association rule mining method which considers the hierarchical structure of the ontology is developed. A new quality measure called Q measure for relation pruning is proposed. Here, Figure 3.5 is used to illustrate the idea of the Q measure. It can be seen that concept *Even-toed Ungulates* has three children in the ontology, namely *Deer*, *Yak*, and *American Bison*. If the relation *Gray Wolf hasDiet Even-toed Ungulates* is correct, the three relations *Gray Wolf hasDiet Deer*, *Gray Wolf hasDiet Yak*, and *Gray Wolf hasDiet American Bison* should also be correct if a minimum support level is present. Given a sufficient large number of collected documents, the three relations should have the same frequencies (i.e., the expected value of 1/3). But in realty and with a smaller number of documents, the three relations have different frequencies. Therefore a variance-like value Q could be computed by

$$Q = \sum_{i=1}^{N} [\frac{freq(C_i)}{freq(R)} - \frac{1}{N}]^2 \tag{3.1}$$

where C_i represents a child rule of a generalized rule R, and N is the number of children rules of R. For those relations with parent concepts, they would have a lower Q value although they have high frequencies. Therefore those relations can be efficiently removed by looking at the Q value and the predefined support threshold. After the pruning process, the accuracy of the remaining relations in the ontology is improved. To find the Q value and support threshold, two animal concepts *Bush Dog* and *Spotted Hyena* are randomly selected for experiment. The relation accuracy is calculated using different combinations of Q value and threshold. Fig. 3.6 contains the plots of result. It can be seen that the accuracy is more sensitive to the change of Q value than the support threshold. The final support and confidence thresholds are set as 20 and 0.6, as this set of parameters give best accuracy. After the relation pruning process, the final ontology contains 743 concepts.

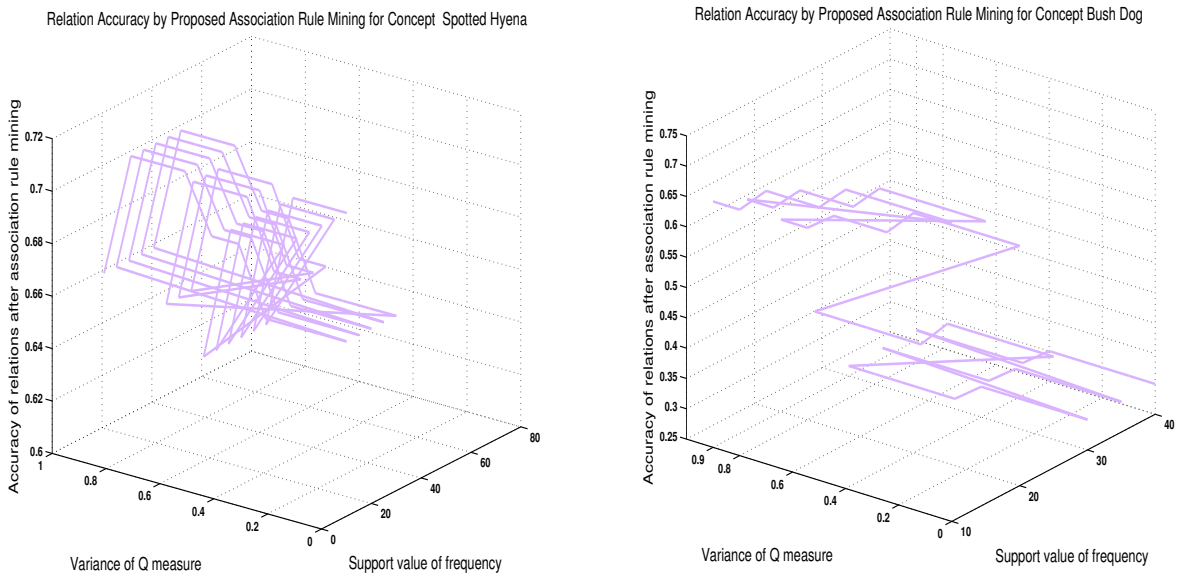


Figure 3.6: The relation pruning accuracy by proposed association rule mining

3.3 Experiment Result

In the experiment the ontology-based image retrieval systems is compared with the Google image Search, which is among the best keyword based search engines and handles over 2 billion images. The experiment data set is set up by a total of 4000 images using the top 200 Google images of each of the 20 *canine* subspecies. Google image is used in the experiment as it is accessible and other researchers can easily compare our performance with their experimental results. The reason for using only the top 200 ranking results of Google image is that this set of images are statistically and visually higher in significance and ranking. The data set of the web images and their web pages is downloaded by our image crawler. For the comparison, the Google Image Search is compared with text ontology-based retrieval and multi-modality ontology-based retrieval. For semantic matchmaking of ontologies, RACER version 1.9[108] is chosen as our reasoner since it is able to provide consistency checking of the knowledge base, computing entailed knowledge via resolution and processing queries through complex reasoning. For the image processing part, classifiers are trained according to a 164 dimensional features (SIFT with opponent color angle) and generate middle-level concepts from the training

Table 3.1: Performance of image classification

Classification	ACCR
Colorful/Graylike	0.921
Photograph/Drawing	0.842
Outdoor/Indoor	0.806
HumanRelevantScene/Buildings/Wildlife	0.794
Greenery/Sand/Stone/Snow/Others	0.814
WhiteFur/RedFur/GrayFur/BrownFur/NonAnmial	0.634

result and integrate the AutoOnto to form AutoMMOnto(Auto Multi-Modality Ontology). The MAP results of our experiment on Google (top 200 retrievals) Image search, AutoOnto and AutoMMOnto(AutoOnto+ visual descriptions) are 0.7049, 0.8942 and 0.9125 respectively. The experiment result in this chapter proves that AutoOnto captures the relationships between concepts as well as, if not better than the manually-built ontology with bigger knowledge coverage and higher efficiency. It is shown in the following subsections that the proposed method allows automatic construction of large-scale multi-modality ontology with high accuracy from a challenging web image data set.

3.3.1 Experiment on Ontology Model Construction

The following experiment aims to find a proper ontology model for web image retrieval. We start from a comparison of keyword approaches with text ontology approaches, and move forward to add more information to build a more informative ontology. Each step examines the improvement brought by the previous step. Finally we arrive at a multi-modality ontology which works well on our data set.

3.3.1.1 Keyword versus Text Vector

In this section our initial attempt at text-based image retrieval is briefly discussed. The text information is mainly from the web page, including the web page title, image captions and surrounding text. Web pages usually contain noisy information. WordNet[109] is used to produce a concise word set without much redundancy for each web page and find the top 10 most frequently occurred keywords for a particular subspecies. This set of frequent keywords is used as a text vector to match with individual image in the database. The matching degree determines the ranking of the image. The more matches

are found, the higher the ranking is and vice versa. The experiment is first done on 2 groups of images: 1.*Golden Jackal*; 2.*Arctic Fox*. The result is compared with Google Image Search, which is among the best keyword based search engines and handles over 2 billion images. The result is shown in Figure 3.7. This figure plots the number of correct images in top N retrievals versus the total number of images returned (in ranking order). Therefore an optimal result would be a line with 45 degrees to the x-axis. Note that the optimal blue line in this figure returns the N correct images in the first N ranking positions. 200-maximum N is the number of images that has been manually(visually) identified as not being in that particular *canine* subspecies(but wrongly returned in the text-based Google Image Retrieval). It is found that the text vector approach follows the Google Image Retrieval result closely. There are 2 main reasons: 1. The current text vector is constructed only from the body of the main text, whereas Google Image Retrieval also looks at image filenames for keyword. The extended list of keywords in our approach help to compensate but did not significantly improve on the overall result; 2. Due to the flat structure of the text vector and limited keywords defined. This text vector does not provide the relationships between the keyword and it can not support a hierarchical structure. Therefore, we turn to a text description ontology, which has a more refined structure and will provide semantic relationships between terms and concepts.

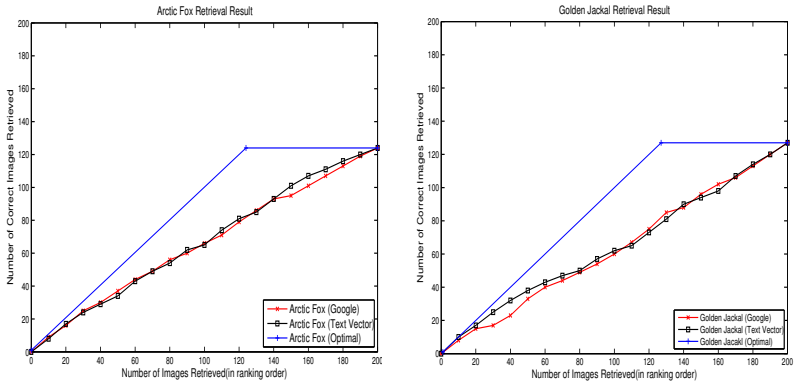


Figure 3.7: A comparison of image retrieval results between keyword and text vector approaches

3.3.1.2 Keyword versus Text Ontology

Semantic matchmaking is applied on the 200 top-ranking Google images with web pages and some results of the overall performance of different approaches are presented in Fig-

ure 3.8. In this test, the 200 images in each category include the training and test data used in previous image classification test. It can be seen that the overall performance of text description ontology retrieval is slightly better than keyword based search. However, text ontology retrieval is still hampered by the lack of text information within the web page. For example, if no related concept and relationship is extracted from the surrounding text, the generated class of this image is void. The result ranking is based on the degree of match: exact match, subsume match and disjoint. As a void class is disjoint with any pre-defined *canine* class, the image is ranked low in the final result even if it is correct.

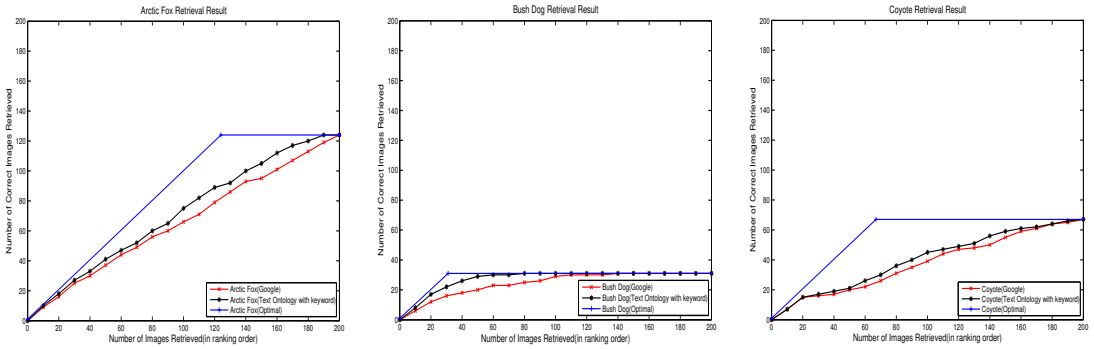


Figure 3.8: A comparison of image retrieval results between keyword and text Ontology approaches(1)

As most keyword-based search engines use single keyword to query the database, to further verify that ontology is effective in information retrieval, the animal name, which is usually entered by users as keyword for search, is removed from the ontology. The result in Figure 3.9 shows that the overall performance is still better compared to keyword-based retrievals. However, the text description ontology is still single-modality and only based on text information. When the images are annotated with loosely-coupled text, there are negative effect on the performance. Better performance are expected by using a multi-modality ontology which incorporates more information from image features.

3.3.1.3 Keyword, Text Ontology versus Multi-Modality Ontology

From the Figure 3.10, it can be seen that the multi-modality ontology-based retrieval outperforms others by returning more relevant images with higher ranking. In the best

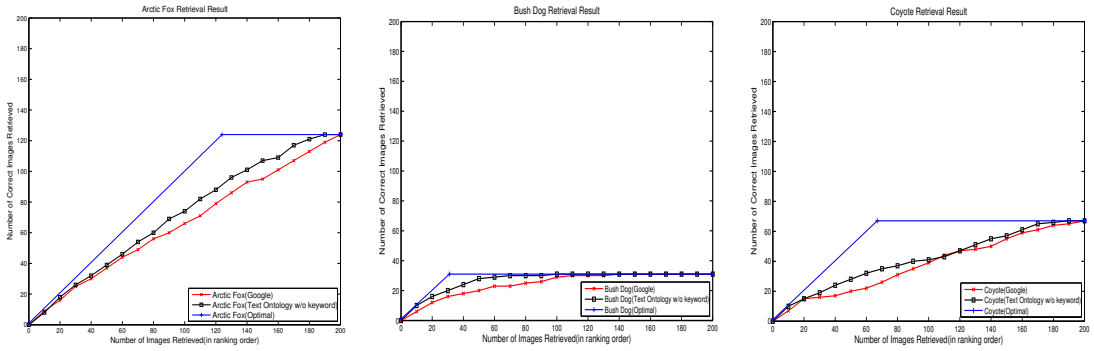


Figure 3.9: A comparison of image retrieval results between keyword and text ontology approaches(2)

case *Arctic fox*, the multi-modality ontology-based retrieval almost overlaps the optimum blue line, which returns the N correct images in the first N ranking positions. The experimental results for low-level feature extraction using SVM is shown in Table 3.1, which are used to build the multi-modality ontology. To evaluate the performance of high-level information extraction based on low-level features, the average correct classification rates (ACCR) are listed in Table 3.1. The formula to calculate ACCR is given as follows:

$$ACCR = \left(\sum_{i=1,\ldots,N} \frac{n_{i,c}}{n_{i,t}} \right) /N$$

where N is the total number of repeated classifications, $n_{i,c}$ is the number of correctly classified images in the i^{th} classification, $n_{i,t}$ is the number of total images in the i^{th} classification. In the classification, one third randomly selected data are used as training samples and test samples. Each classification are repeated 10 times and the ACCR is then calculated. The last set of classification does not achieve very good classification performance because most fur colors of *fox* are affected by the change of illumination and viewing angle,with *WhiteFur* as an exception a high ACCR of 0.826(this ACCR value is different from the one

shown in Table 3.1 as that value is the Average ACCR of all fur types). However, there are gaps between the multi-modality ontology results and the optimal results in most cases. We presume it could be due to one or more of these reasons: First the performance could be affected by the accuracy of image feature classification; Second, the lack of text information in the web pages will result in less correspondence in text ontology and multi-modality ontology.

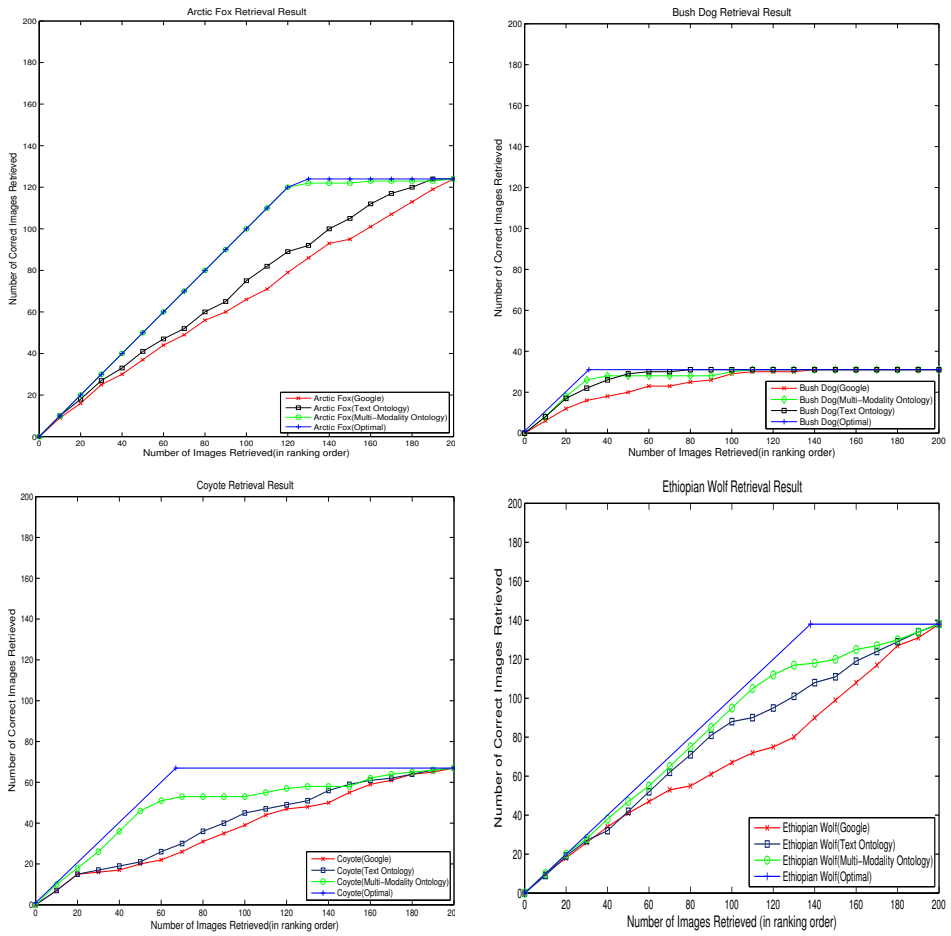


Figure 3.10: A comparison of image retrieval results between different approaches(1)

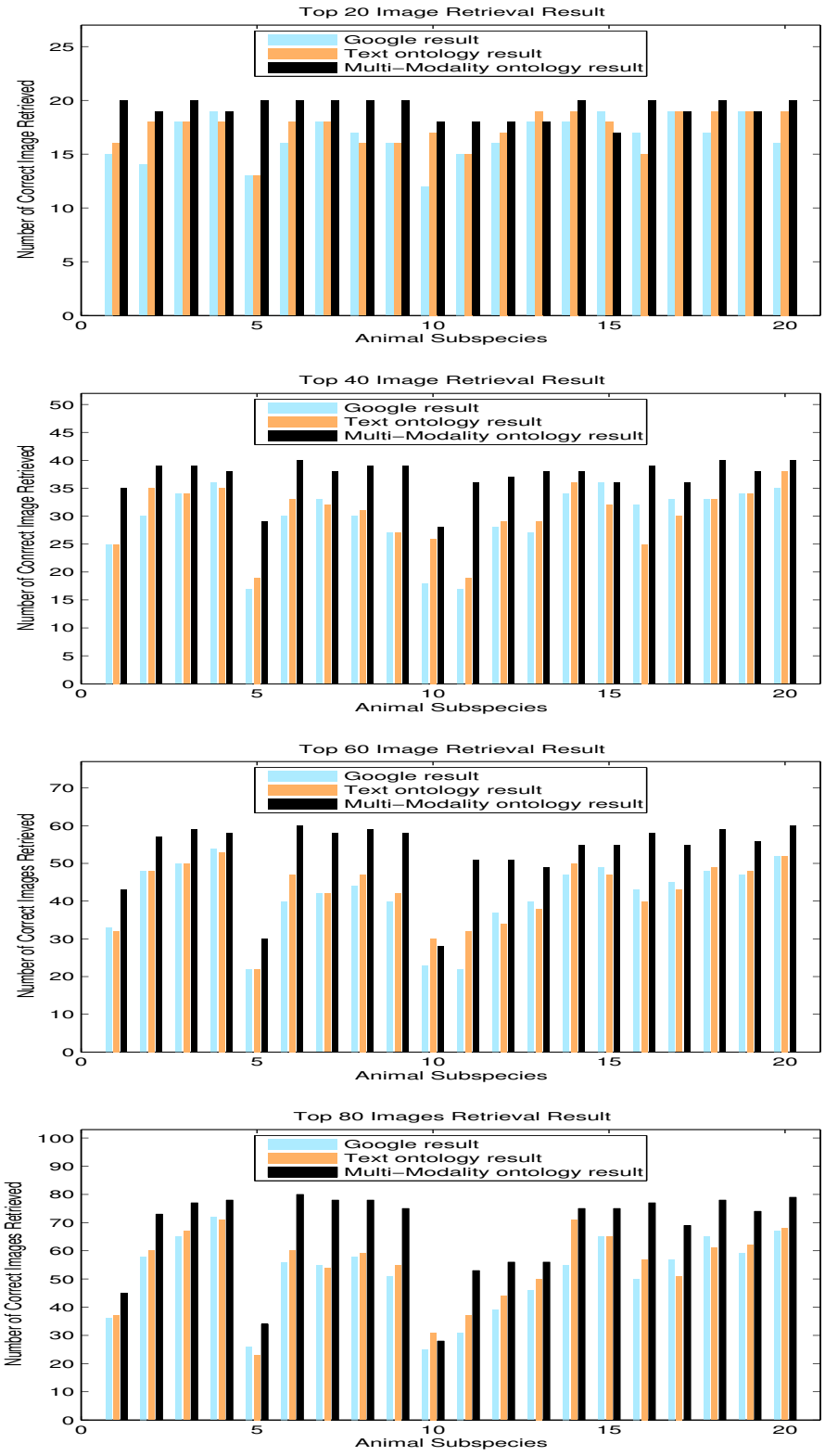


Figure 3.11: A comparison of image retrieval results between different approaches (2)

3.3.1.4 Comparison of Precision Results

In web image search and retrieval systems, precision is a very important guideline to measure the performance of a retrieval system because most users will browse only limited numbers of results. Therefore recall rate for a web image retrieval is not that crucial. The correct number of image retrieved in the top 20, 40, 60 and 80 for all 20 *canine* subspecies are shown in Figure 4.9. The 20 subspecies are: 1. *aardwolf*, 2. *African wild dog*, 3. *bat-eared fox*, 4. *black jackal*, 5. *cape fox*, 6. *arctic fox*, 7. *gray fox*, 8. *red fox*, 9. *kit fox*, 10. *bush dog*, 11. *coyote*, 12. *dhole*, 13. *dingo*, 14. *Ethiopian wolf*, 15. *fennec fox*, 16. *golden jackal*, 17. *gray wolf*, 18. *maned wolf*, 19. *red wolf* and 20. *spotted hyena*. From the figure, it can be seen that for top 20 image retrieval result, nearly all images retrieved by the multi-modality ontology are correct. In most cases, ontology-based image retrieval can achieve better retrieval precision than keyword-based image search. In the experiment, only generic image classification mechanism, which is not particularly designed for the target domain is implemented. Table 3.1 shows that if the images are retrieved only based on image classification results, the result fails to outperform the normal text based image retrieval mechanism. However, by combining the high-level text information with low-level image features, the retrieval precision is further improved by about 5 to 30 percent.

3.3.2 Comparison of ManuOnto and AutoOnto

In this experiment, we mainly focus on the comparison of manually built ontology and automatically built ontology. Manually built ontology includes ManuOnto(Manually built text Ontology) and ManuMMOnto(Manually built Multi-Modality Ontology). ManuOnto uses concepts that are manually extracted from the descriptions of BBC Science & Nature Animal category. ManuOnto is further enhanced by adding concepts from the image feature classification results to form the ManuMMOnto. Similarly, automatically built ontology includes AutoOnto(Automatically built text Ontology) and AutoMMOnto(Automatically built Multi-Modality Ontology). AutoOnto is built from concepts that are automatically extracted from Wikipedia through the Wikipedia2Onto process, while AutoMMOnto adds low-level image concepts on top of AutoOnto. In the experiment, both the DL reasoner(semantic matchmaker) RACER[108] and an enhanced

ranking algorithm(Section 4.2) are used for matchmaking. Matchmaking is defined as a process that requires the user specified domain concept repository to take an image's detected concept as input and to return all the matched domain concepts which are compatible with the concept generated from the input concept. From the matchmaking result it can be decided which predefined concept the generated image concept corresponds to and what relationship can be found between two given concepts. In this step, reasoners are used to derive additional facts which are entailed on any optional ontologies and predefined rules. They process and reason over the knowledge encoded in the ontology language. The details of the matchmaking result will be shown in Chapter 4. The matched concepts are attached with the web image as semantic labels.

The experiment result is shown in Table 3.2 & 3.3. The classification result is a list of web images which are ranked according to their match degree. Average Precision(AP) is used to evaluate the performance. As a most frequently used summary measure of a ranked retrieval, AP is defined as: $AP = \frac{1}{min(R,k)} \sum_{j=1}^{k} P(r_i) I_j$, where R is the total number of correct images in the ground truth, k is the is number of current retrievals, and $I_j = 1$ if image ranked at j^{th} position is correct and $I_j = 0$ otherwise. $P(r_i) = \frac{R_j}{j}$ is the interpolated precision. And {r,P(r)} are the available recall-precision pairs from the retrieval results. By using AP, the PR curve can be characterized in a scalar. A better retrieval performance, with a PR curve staying at the upper-right corner of the PR plane, will have a higher AP, and vice versa. In the current experiment, $j = 200$ is set. Mean Average Precision(MAP) is also shown as MAP is sensitive to the entire ranking with both recall and precision reflected in this measurement. The table below shows the Average Precision(AP) values for each class using different approaches. The input query keywords are the names of the classes. Google Image Retrieval result is given as the baseline. From the result it can be concluded that text ontology improves the retrieval performance by formulating the text information into structured concepts. Table 3.3 shows the experimental results of web images retrieval based on ManuMMOnto and AutoMMOnto. From this table, it is observed that the AutoMMOnto approach gives comparable performance to the ManuMMOnto approach. In most classes, AutoMMOnto generates even better results by extracting more concepts from the web page text. The MAP of the Google, ManuMMOnto and AutoMMOnto are 0.7049, 0.8942 and

Table 3.2: Performance of image classification on single-modality text ontology

Class	Aardwolf	Cape Fox	Bush Dog	Arctic Fox	Ethiopian Wolf
Google	0.5801	0.4958	0.4695	0.715	0.7516
ManuOnto	0.6209	0.5446	0.7881	0.7905	0.844
AutoOnto	0.6472	0.5231	0.8422	0.7983	0.7973
Class	Coyote	Gray Wolf	Gray Fox	Fennec Fox	Spotted Hyena
Google	0.5042	0.7513	0.7183	0.8181	0.8365
ManuOnto	0.5421	0.7196	0.6336	0.8145	0.8683
AutoOnto	0.496	0.7316	0.6465	0.8214	0.9024
Class	Dhole	Red Fox	Maned Wolf	Black Jackal	Bat-Eared Fox
Google	0.6342	0.744	0.7949	0.8872	0.7967
ManuOnto	0.6598	0.781	0.8565	0.8805	0.7914
AutoOnto	0.6835	0.7522	0.8193	0.8959	0.8396
Class	Dingo	Kit Fox	Red Wolf	Golden Jackal	African Wild Dog
Google	0.67	0.6698	0.7669	0.7092	0.7844
ManuOnto	0.6799	0.6844	0.715	0.7252	0.7723
AutoOnto	0.7196	0.6791	0.8175	0.7528	0.7869

0.9125, respectively. The result of MAP also shows an overall improvement. It is worth adding, AutoMMOnto requires minimal level of human involvement: only the main domain concepts, which are the image classes in our case, are given by users according to experimental domain to build up the whole concept hierarchy in the domain. The result is encouraging, as it proves that it is viable to build large-scale concept ontology from Wikipedia automatically for effective web image retrieval.

3.4 Summary

In this chapter, the issue of how to encapsulate knowledge in an effective ontology model has been discussed. The model is constructed step by step through a series of experiments. The final model, which is derived from both the domain knowledge of text and image features, is proved to be effective. It has also been proved that a semantically rich ontology addresses the need for complete description of image retrieval and also improves the precision of retrieval. To address the issue of ontology construction efficiency and scalability, we have proposed Wikipedia2Onto - an approach that uses the content and structure features of the online encyclopedia Wikipedia to build large-scale concept

Table 3.3: Performance of image classification on multi-modality ontology

Class	Aardwolf	Cape Fox	Bush Dog	Arctic Fox	Ethiopian Wolf
Google	0.5801	0.4958	0.4695	0.715	0.7516
ManuMMOnto	0.8332	0.8911	0.8087	0.9955	0.9218
AutoMMOnto	0.8552	0.8835	0.9302	0.9938	0.9447
Class	Coyote	Gray Wolf	Gray Fox	Fennec Fox	Spotted Hyena
Google	0.5042	0.7513	0.7183	0.8181	0.8365
ManuMMOnto	0.9058	0.8267	0.935	0.857	0.9391
AutoMMOnto	0.884	0.8561	0.9766	0.8981	0.942
Class	Dhole	Red Fox	Maned Wolf	Black Jackal	Bat-Eared Fox
Google	0.6342	0.744	0.7949	0.8872	0.7967
ManuMMOnto	0.8184	0.966	0.9508	0.9498	0.9134
AutoMMOnto	0.8535	0.9526	0.938	0.9555	0.9333
Class	Dingo	Kit Fox	Red Wolf	Golden Jackal	African Wild Dog
Google	0.67	0.6698	0.7669	0.7092	0.7844
ManuMMOnto	0.8108	0.9483	0.8537	0.8962	0.8627
AutoMMOnto	0.8334	0.8821	0.9034	0.9166	0.9185

ontology automatically. The constructed ontology - AutoMMOnto has extracted more descriptive semantic relationships than the hypernymy/hyponymy or meronymy relationships which are the only semantic relationships contained in many existing ontologies. The proposed ontology construction approach has detected 743 concepts and with the association rule mining algorithm, a high accuracy has been achieved for the concepts and the corresponding relations.

Chapter 4

Semantic Matchmaking with Multi-Modality Ontology and Experiment Result

This chapter discusses the following issues: (1)How to use the knowledge contained in an existing ontology for information retrieval? (2)What is a scalable and effective way to do so? Semantic matchmaking is the process to make inference on the ontology knowledge and generate additional facts. This chapter discusses the semantic matchmaking models used for the multi-modality ontology proposed in the previous chapter. Traditional DL based semantic reasoners can be used with the ontology to improve the retrieval performance by re-ranking Google's retrieval result. However, there are some limitations. Most semantic matchmaking approaches only provide three results: *Exact Match*, *Subsume Match* and *Disjoint Match*. *Exact Match* is considered to be the best matching result, since the input query concept is exactly the same as the predefined concept. *Subsume Match* is chosen as the next preferred match, as the input query concept subsumes several predefined concepts, which means it could be annotated as one of several concepts. *Disjoint Match* comes last because the input query concept does not belong to any predefined concept. However, there is further need to quantize the measurement of the similarity between concepts using a global method. We start from the traditional Description Logic(DL) based reasoners, to a more scalable ontology inference model aided by Spreading Activation Theory(SAT). Based on the proposed matchmaking model, an overall structure of the image retrieval system is also shown as an implementation.

4.1 Description Logic Based Semantic Matchmaking

As an initial attempt, reasoners are used for semantic matchmaking. Reasoners are based on DL and they are able to provide consistency checking of the knowledge base. They also compute entailed knowledge via resolution and process queries through complex reasoning. The following subsections will introduce the underlying mechanism of DL and the matchmaking algorithm adopted in the experiment.

4.1.1 Description Logic

DL is the logic specifically designed for knowledge representation in terms of classes and relationships between classes. They are considered as the most important knowledge representation formalism. They unify and give a logical basis to the traditional frame-based systems, semantic networks, semantic web ontology languages, object-oriented representations, semantic data models, and type systems. In DL, the domain of interest is modelled by means of *concepts* and *relationships*, which indicate classes of objects and relations between objects, respectively. Numerous works have investigated the relationship between expressive power and computational complexity of reasoning[52]. The research on these logics has led to a number of automated reasoning systems (See section 4.1.2 for more details). Here is a brief review on the DL $\mathcal{ALCN}$ which is used in the current reasoning system.

The basic elements of $\mathcal{ALCN}$ are *concepts* and binary *relations*. Let A denotes atomic *concepts* in $\mathcal{ALCN}$. $\mathcal{ALCN}$ concepts, denoted by C, can be represented by the following constructs:

$$\begin{aligned} C \doteq \quad & \bot \mid \top \mid A \mid \neg C \mid C_1 \sqcap C_2 \mid C_1 \sqcup C_2 \\ & \mid \exists R.C \mid \forall R.C \mid (\leq nR) \mid (\geq nR) \end{aligned} \tag{4.1}$$

Its semantics are given by an interpretation $I = (\triangle^I; \cdot^I)$, which consists of a set $\triangle^I$, called the domain of I, and a valuation $\cdot^I$ which maps every concept to a subset of $\triangle^I$ and every role to a subset of $\triangle^I \times \triangle^I$. Here some abbreviations are used in the constructs: $\bot$ stands for $A \sqcap \neg A$, where A is any atomic concept. Therefore $\bot$ is the bottom concept of the world described, which means it contains no individuals. $\top$ for $\neg\bot$. It stands for

the universal concept of the world described, which means it is a union of all individuals. $C_1 \sqcup C_2$ for $\neg(\neg C_1 \sqcap \neg C_2)$; $\forall R.C$ for $\neg\exists R.\neg C$;

Among the constructs used in forming concept C, boolean constructs can be seen, i.e., *negation* ($\neg$), *conjunction* ($\sqcap$) and *disjunction* ($\sqcup$), number constructs, that is, *exists restricts*, *value restricts* and the *number restrictions*. $\exists R.C$ denotes that at least one of concept C's object participates to the relation R as its range object; The *number restrictions* ($\leq nR$) denotes that there are at most n objects participated in the relationship R. ($\geq nR$) denotes that there are at least n objects participated in the relationship R.

4.1.2 Semantic Reasoner

Reasoners are used to derive additional facts which are entailed on any optional ontologies and predefined rules. They process and reason over the knowledge encoded in the ontology language. Currently there are various reasoners with respective tool suite developed by different organizations, from semantic web research community to DL community. For instance, some of the representative reasoners are RACER[108], FaCT[110], Pellet[111] and KAON[112]. A comparison of them is shown in Table 4.1. Some works [113–115] in the semantic web area have studied the application of domain knowledge in the web services discovery or grid system. In the multimedia retrieval system, the semantic retrieval still relies on the human annotation of the multimedia objects. This is mainly because the semantics between high level semantics and low level features are quite different. Research works have studied either side of this problem, but few of them successfully combines them together. In our system, low-level features are aligned to some middle level concepts, which represent the low level descriptions. After that, these concepts are combined together with the high-level semantics of the image, which are generated from on the image's surrounding text. All the domain knowledge is represented in the form of DL for the knowledge base retrieval purpose.

In our system, RACER version 1.9 is used as the reasoner given its capability in comparison with others. RACER can be accessed by standard HTTP or TCP protocols. It comprises two parts: A HTTP server named RacerPro and a Graphic User Interface called RacerPorter. Lisp style language, which is a multi-paradigm, reflective programming language, is used to construct and search the data in ontologies. A typical

Table 4.1: Comparison of various reasoners

	RACER	FaCT	Pellet	KAON[116]
Algorithm	Tableau[112]	Tableau	Tableau	Tableau
Logic	DL	DL	DL	
Language	Lisp Style	Lisp Style	Java	
Interface	DIG Java GUI	DIG Cmd-Line	DIG, Java	Java API
query	nRQL		RDQL	spaRQL
Limitation	HTTP overhead	No Abox Support	Performance	

knowledge base in RACER consists of the set of TBox and associated ABox. The TBox contains sentences describing concept hierarchies (i.e., relations between concepts) while the ABox contains "ground" sentences stating to where individuals belong in the hierarchy(i.e., individuals and the relations between individuals). RACER provides strong support to inference over ABox. Its functionality of providing a mirror data substrate is of special value for our system[117]. This function automatically creates, for every object in a given ABox, a corresponding and appropriately labelled substrate data object. As a result, given any two data objects, their relationships can be retrieved according to the ontology.

4.1.3 Matchmaking Algorithm

The *canine* domain knowledge is predefined in the knowledge base. Matchmaking is defined as a process that returns all the matched domain concepts which is compatible with the concept generated from the input image I. The matchmaking result decides which predefined concept the generated image concept corresponds to. Let θ be the users' specified portion of domain concept repository, the matchmaking algorithm should return match(I), which is given by:

$$match(I, \theta) = \{A \in \theta | \neg((A \sqcap I) \sqsubseteq \bot)\}$$

The matchmaking algorithm is defined as follows: For the input generated image concept I and the predefined concept C in knowledge base, the matching degrees, from the best to the worst, are: *Exact*($I \equiv C$), *Subsume* ($C \sqsubseteq I$), *PlugIn*($I \sqsubseteq C$), *Intersection* ($\neg(I \sqcap C \sqsubseteq \bot)$), and *Disjoint* ($I \sqcap C \sqsubseteq \bot$).

Algorithm 1 Matching Algorithm

```
1: for all concepts in image concept knowledge base do
2:    if concept-equivalent(inputConcept, concept) then
3:       put EXACT match
4:    else if concept-subsumes(inputConcept, concept) then
5:       put SUBSUME match
6:    else if concept-subsumes(concept, inputConcept) then
7:       put PLUGIN match
8:    else if concept-subsumes(¬inputConcept, concept) then
9:       put DISJOINT match
10:   else
11:      put INTERSECTION match
12:   end if
13: end for
```

The input concept I is created by the text analysis and the low-level feature extraction phase. The predefined concept C is from the ontology constructed by domain expert. For the input image concept I and the predefined concept C, *Exact*($I \equiv C$) matches are considered to be the best matching, since the input concept is exactly the same to the predefined concept. *Subsume* ($C \sqsubseteq I$) matches are chosen as the next preferred match, since it is expected that the predefined concept conforms to the input concept. *PlugIn*($I \sqsubseteq C$), *Intersection* ($\neg(I \sqcap C \sqsubseteq \bot)$), and *Disjoint* ($I \sqcap C \sqsubseteq \bot$) matches are considered as concepts with lower relevancy to the input query, as an intersection indicates that the text and visual information of the web image leads to multiple concepts and a disjoint indicates that no information can link to the query concepts. Therefore, in the final retrieval result, web images annotated as intersection and disjoint have lower rank to images annotated as exact and subsume match. The main matchmaking process is shown in Algorithm 1. Here an example of the matchmaking is shown. If image concept I is defined as *hasName.CapeFox*, *hasDistribution.Zimbabwe* and *hasDiet.Insect*, it is inferred to be subsumed by concept *cape fox*. If image concept J contains both high-level feature *hasName.CapeFox* and low-level feature *hasFur.WhiteFur*, this concept is

inferred to be an *Intersection* of *cape fox* and *Arctic fox*. If image concept K has features as *hasName.GrayWolf* and hasFur.GrayFur, it is inferred to be an *Disjoint* of *Cape Fox*. As a result, in a search of *Cape Fox* images, the retrieval ranking is $I > J > K$.

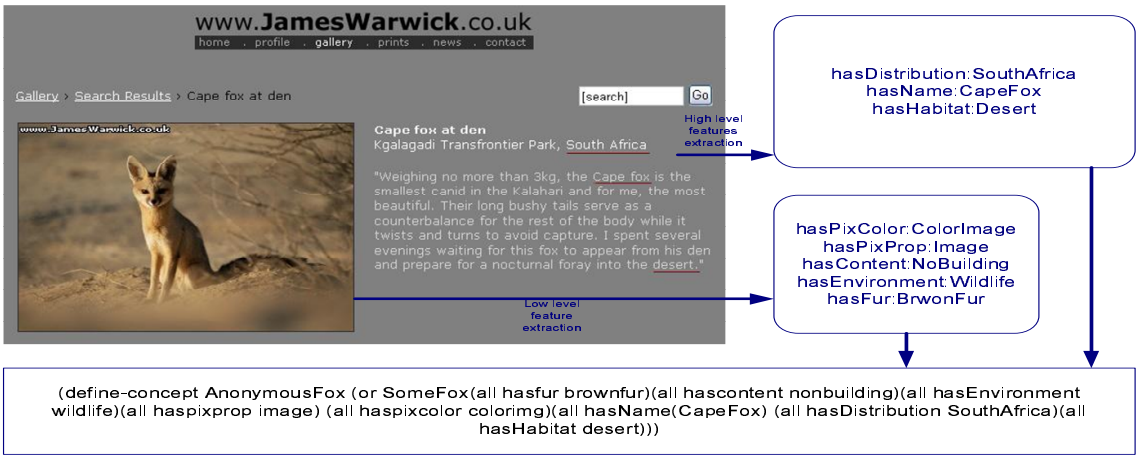


Figure 4.1: A *cape fox* image sample

Figure 4.1 is given to illustrate the matchmaking process: The high-level text features from the surrounding text have already been obtained as *hasName.Capefox*, *hasDistribution.SouthAfrica* and *hasHabitat.Desert*. Combining this information with those from low-level features, a definition for the *AnonymousFox* in the image is generated.

$$\begin{aligned} AnonymousFox \doteq\ & Fox \sqcup \forall hasName.CapeFox \\ & \sqcup \forall hasDistribution.SouthAfrica \\ & \sqcup \forall hasHabitat.Desert \\ & \sqcup \forall hasPixColor.Colorimg \\ & \sqcup \forall hasPixProp.image \\ & \sqcup \forall hasEnvironment.Wildlife \\ & \sqcup \forall hasFur.Brownfur \\ & \sqcup \forall hasPixContent.Nobuilding \end{aligned}$$

From the above definition, it can be inferred that this *AnonymousFox* is a subclass of *cape fox*. Concept satisfiability and subsumption judgment are the primary tasks of DL reasoner. Firstly the RACER system is used to classify the taxonomy hierarchy for

all domain knowledge concepts. The classified taxonomy will be used to speed up the concept subsumption judgment. When the input image arrives, RACER judges the input image concept's relationship with the predefined domain concepts. Those with predefined concepts C equivalent to I are considered as *Exact* matches. C that subsume but not equal to I are considered to be *Subsume* matches.

The matchmaking time of each image concept is not influenced by the size of image database, but only by the size of predefined knowledge base. Since the images are matched one by one against the knowledge base whose size is fixed once constructed, the computational complexity for the system is linear with the size of image database. As a result, with a given ontology the system has potentially good scalability, and the performance can be further improved by parallel matchmaking. A major advantage of the ontology matchmaking approach is its extensibility and reusability. The current system can be easily extended to other domains by defining new domain knowledge in the knowledge base. The system will immediately work on the new domain and generate corresponding results. The overlapping concepts between the original and new domains can be easily reused.

4.2 Matchmaking with Enhanced Ranking Algorithm

In the previous model, when displaying the retrieval result, a ranking is generated according to the aforementioned criteria. A further ranking is given according to the degree of *Subsume Match*. The more concepts an input query concept subsumes, the lower its ranking in the final result. This is easy to understand, as a concept subsumes more concepts, the possibility that it belongs to any of these concepts is smaller. In most cases, this ranking mechanism provides satisfying result. However, a properly defined measurement is still preferred to further quantize the similarity between concepts.

The importance of different features is taken into consideration. In literature, T. L. Berg et al.[62] used a linear and equal weight combination of 4 independent cue scores from both visual cues and text cues. Y. Gao et al.[71] propose Bayesian inference to measure the concept similarity. Y. Keiji et al.[118] use probabilistic method to process the

text information. However, different features should not have same priority in real practice. Probabilistic methods are based purely on mathematics and do not take advantage of the ontology knowledge.

In order to solve the concept similarity measurement problem, some other information which was not used in the previous model is exploited: the ranking of semantic relationships in concept. As mentioned in previous chapter, ontology is composed of relationships between concepts, which come from different modalities. The knowledge base is constructed by predefined concepts. These concepts include *animal* concepts, which are used to matchmake with and annotate to the web image, and general concepts such as *color*, *distribution*, etc.. The general concepts are defined to construct the *animal* concepts and these two kinds of concepts are connected by semantic relationships, which are also defined in the knowledge base. During the image analyzing process, a generated concept is automatically constructed for each image according to the information extracted from text analysis and image feature analysis. This generated concept, which is also known as an anonymous concept, is reasoned over the *animal* concepts in the knowledge base to check if it matches any predefined *animal* concept. Different from [62], semantic relationships should not have equal weight according to human knowledge. Taking the *animal* concept in our experiment as an example, *animal* concept has *hasName* and *hasColor* relationships, which link the *animal* concept to *Name* and *Color* concepts. In the case of *Gray wolf*, there is *Gray wolf.hasName* (*Gray wolf* or *timber wolf* or *Rockey mountain wolf*). There is also *Gray wolf.hasColor* (*brown* or *gray* or *white*). *hasName* relationship should have higher priority than *hasColor* relationship, as different animals may share the same color, but not the same name. The aim is to build a similarity measure system according to different rankings(priorities) of semantic relationships. In the following sections, the work on how to improve the existing matchmaking model is introduced.

4.2.1 Image Concept Extraction

The concepts in the ontology (which contains the domain knowledge) are defined as predefined concepts. On the other hand the concepts extracted from images are image concepts and contain both image feature and text information. In this section, the

features extracted from image content and surrounding text are discussed. Three main steps are involved to generate a feature histogram for each image. In the first step, the image features and text features are extracted and used to form the image concept for each image. The second step is to use a bin for each image feature or text feature. Features from the two modalities are handled differently. In the final step, when constructing a combined histogram from different modalities, a higher weight (as a multiplier) is assigned to the bins of the histogram containing the images features. This combined weighted feature histogram is used in the calculation of ranking correlation for each image. By using Spearman's correlation, the degree of concept similarity can be quantized and the images can be indexed accordingly.

4.2.2 Binary Image Features

Binary histogram is used to represent each image features. The image features used in the visual ontology part are extracted from low levels. The multi-modality ontology is used to connect them to human understandable high-level concepts. Considering the data set size and performance of offline indexing, relevant image features is detected from both of the image foreground and background. These image features include *PixColor*, *PixProperty*, *Content*, *Environment* and *ObjectFurTexture*. The first two features analyze the image property, as to identify if images are photographs of real animal. The next two features mainly analyze the background information of each image. A decision is made on whether the image is an indoor or outdoor shoot and whether the image is photographed in the wild. The last image feature provides information about the texture of the foreground object. The reason why a binary histogram is used is that the possible values for these features are all pairwise. For example, *PixProperty* can either be an image or graphic picture; *ObjectFurTexture* can be any colorful animal fur or non-animal fur; those values that match the predefined concept descriptions are considered as positive. For both cases the former values indicate positive feature while the latter indicate negative. Therefore the following definitions are given:

$$H(F_I) = \begin{cases} 1 \text{ for positive feature} \\ 0 \text{ for negative feature} \end{cases} \quad F_I = 0, 1, ..., NF - 1 \tag{4.2}$$

where $H(F_I)$ is the value of bin and NF is the number of image features.

The binary image feature is a set of indicators of middle-level feature description. To extract the middle-level feature from images, a SVM classification was proposed in Chapter 3 to generate middle-level features from low-level image features. However, it is not easy to deal with multiple features in a single classification. The classification efficiency need to be further improved. Therefore, the different approaches in kernel-based classification is investigated.

Kernel functions measure the similarity of the input data, which could be image feature vectors. Kernel-based classification is an effective way to represent decision boundaries of features and it is widely used in discriminative object classification and recognition. It first appeared in SVM as a classification algorithm. Some representative kernel functions are linear, polynomial and Gaussian Radial Basis Functions. A very important question is how to find an appropriate kernel function. This question is even more important in image classification field where the number of dimensions of the image feature space is high. Researches are trying to find better kernels to improve the classification performance. In recent reports[119–121], pyramid match kernel provides us a possibility to improve our current middle-level feature extraction. The multi-resolution histogram is not a new concept. K. Grauman and T. Darrell proved that the intersection of two multi-resolution histograms is positive-definite and thus can be used as a kernel function. They designed a pyramid match kernel and this kernel function has several advantages:

- This kernel function is no longer a simple distance measure function. Therefore, it can capture certain level of context (semantic) information.
- Since it is a histogram-based approach, this kernel function can deal with any data that could be mapped to a histogram.
- Compared with other kernel function, this kernel has lower computational complexity.

Basically, pyramid matching is to find an approximate correspondence between two sets of n-dimensional vectors. To use pyramid matching kernel, series of histogram is first built for a feature vector which can be achieved by placing a sequence of increasingly coarser

grids over the feature space. After that, pyramid matching calculates the *histogram intersection* of two histograms at different resolution, which is defined as:

$$I(H_X^l, H_Y^l) = \sum_{i=1}^{D} min(H_X^l(i), H_Y^l(i)) \tag{4.3}$$

where H_X^l indicates the histogram of feature vector X at resolution l. Level $l+1$ is finer than level l. Therefore the matches found at level l include all the matches found at level $l+1$. Therefore, a weight $\frac{1}{2^{L-l}}$ is assigned to each level l over totaly level of L. Finally, a pyramid match kernel is defined as:

$$k^L(X, Y) = I^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}}(I^l - I^{l+1}) \tag{4.4}$$

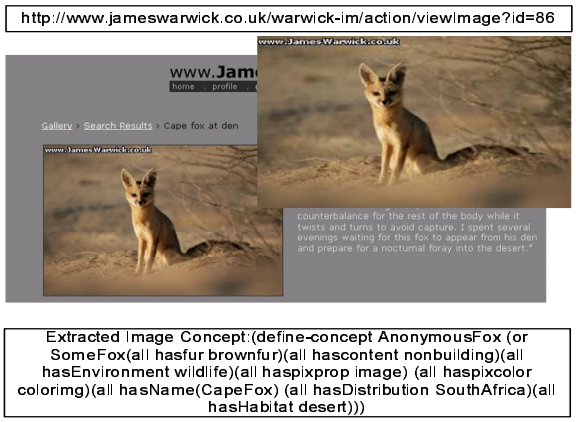
In the implementation, a particular case of utilizing pyramid match kernel is carried out, where feature vector X and Y are two elements of n-dimensional feature space. That means X and Y are also n-dimensional feature vector. More specifically, they could be two instances of a feature descriptor. Various kinds of low-level feature descriptors are involved in our middle-level feature generation stage. Since a pyramid match kernel can deal with an orderless image representation, it allows a precise matching of two collections of features in a high-dimensional appearance space.

4.2.3 Text Features

Text features include *AnimalName*, *BodyColor*, *Habitat*, *Distribution* and *Diet*. In the previous matchmaking model, for text feature part, only the feature values are considered and utilized in the semantic matchmaking. A new matchmaking mechanism is introduced by the different semantic relationship frequencies extracted from web pages. The general idea is that the occurrence of a particular word in one web page indicates the priority of that word, which further decides its degree of relevance to the web page subject. Therefore after the text features are extracted from the web page, a frequency record of the text features is kept. A histogram is also created for each text feature. The number of occurrences fall into the separate bins and the bin value H(f) is defined as follows:

$$H(F_T) = N_{i,j} \tag{4.5}$$

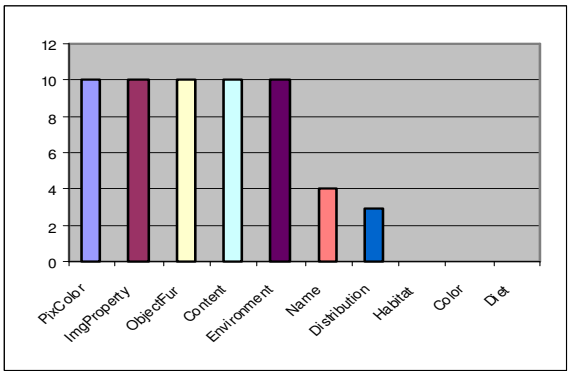
where $N_{i,j}$ is the total number of occurrence for text feature i in the surrounding text of image j.



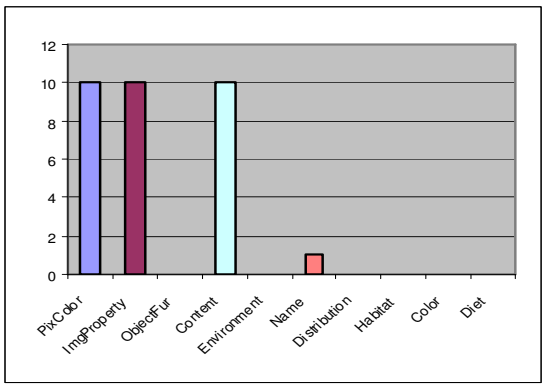
4.2.a: An example of weighted feature histogram and Spearman's ranking correlation 1. The features extracted are: hasObjectFur brownfur; hasContent nonbuilding; hasEnvironment wildlife; hasPixProp image; hasPixColor colorimg; hasName Capefox; hasDistribution SouthAfrica and hasHabitat desert.

4.2.b: An example of weighted feature histogram and Spearman's ranking correlation 2. The features extracted are: hasObjectFur nonanimal; hasContent nonbuilding; hasEnvironment nonwildlife; hasPixProp image; hasPixColor colorimg and hasName Capefox.

Figure 4.2: Examples of weighted feature histogram and Spearman's ranking correlation Part I: Web image examples



4.3.a: An example of weighted feature histogram and Spearman's ranking correlation 1. The weighted feature histogram and final similarity of Figure 4.2.a.



4.3.b: An example of weighted feature histogram and Spearman's ranking correlation 2. The weighted feature histogram and final similarity of Figure 4.2.b.

Figure 4.3: Examples of weighted feature histogram and Spearman's ranking correlation Part II: The weighted feature histograms and final similarities

4.2.4 Weighted Feature Histogram

Since surrounding text of a web image usually contains a lot of noise, the features from image content analysis are included as helpful cues. In the experiment data set, there are cases where the surrounding text are highly related to the input query animal, while the images themselves are about cartoon characters, people or other irrelevant subjects. Therefore, more emphasis is put on image features to filter out these false samples. To make sure these low-level image features will affect the final degree of similarity more, a higher weight is assigned to the image feature histogram than text feature histogram when a combination of the two is made. The final histogram H is defined as:

$$H = [\alpha \times H(F_I), H(F_T)] \tag{4.6}$$

From experience α is usually assigned a larger value to guarantee the priority of the image features. A vector with number of dimensions equals to the number of features is constructed. The n^{th} element in this vector corresponds to the ranking of the n^{th} feature in the histogram. This ranking is calculated according to the value of bins in the histogram. The larger the bin value is, the higher the corresponding ranking is. The underlying assumption is that if certain semantic relationship appears more frequently than others in a web image, the information it contains is more relevant to the image subject, and should have higher priority. This vector is named as *image concept ranking vector* as compared to the *predefined concept ranking vector* of the *predefined concept*. Different from the *image concept ranking vector*, which is generated according to the histogram values of each image, *predefined concept ranking vector* is predefined according to general knowledge. The *image concept ranking vector* is later used in the proposed ranking mechanism as the representation of that particular image.

4.2.5 Ranking Correlation with the Weighted Feature Histogram

In order to quantize the measurement of the similarity between different concepts under the same semantic match-making result, Spearman's rank correlation with the weighted feature histogram is used to measure the similarity between the *image concept* and the

predefined concept in the knowledge base. This is done by calculating the ranking correlation between the *image concept ranking vector* and the *predefined concept ranking vector*. Figure 4.2.a and Figure 4.2.b are simplified examples in our experiment.

Suppose now there are 10 relationships: *hasPixColor, hasPixProperty, hasObjectFur, hasContent* and *hasEnvironment*, which comes from the image feature modality; *hasName*, *hasDistribution*, *hasHabitat*, *hasColor* and *hasDiet* are from the text information modality. Note that this sequence also corresponds to the sequence of the columns in the coming ranking vectors. A *predefined concept ranking vector* is defined according to the importance of each feature. An example of such an ideal case is an image with extracted frequency as [1,1,1,1,1,6,5,4,3,2](The first 5 elements are binary image features whose frequency could be either 1 or 0. The rest 5 elements are text features whose frequency match a descending importance). For illustration purpose all the image features are considered of equal ranking with a high weight of 10, so that the final result will be affected more by the relationships from image feature modality as expected. By putting an weight on the first 5 image feature frequencies, the frequency vector becomes [10,10,10,10,10, 6,5,4,3,2]. Based on this weighted frequency vector, ranking vector is calculated, with rank of frequency values that are the same as the mean of what their ranks would otherwise be. Since this is the ideal ranking vector, it is used as the *predefined concept ranking vector*.In this case the *predefined concept ranking vector* is [3,3,3,3,3,6,7,8,9,10]. This vector is later used to compare with generated *image concept ranking vector* for relevancy calculation.

Two images with their corresponding weighted histograms are shown in Figure 4.3.a and Figure 4.3.b. In the first web image, the generated *image concept ranking vector* is [3,3,3,3,3,6,7,8,9.5,9.5] and the generated *image concept ranking vector* for the second web image is [2,2,7.5,2,7.5,4,7.5,7.5,7.5,7.5] respectively. This ranking is calculated according to the value of histogram.

Spearman's ranking correlation is combined with the multi-modality ontology model to refine the result of semantic matchmaking. Spearman's ranking correlation is used to measure the relationship between two variables regardless of their frequency distribution. The rank correlation coefficient is denoted by ρ and given by:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{4.7}$$

where d_i is the ranking difference between the i^{th} entry of two histograms, and n is the number of elements in the histogram. In our case, the variables are vectors representing the semantic relationships at ordinal level, which makes Spearman's ranking correlation applicable. The values of ρ for the two images in the given example are 0.99697 and 0.65909. This result is the same as our expectation. Therefore, even if the matchmaking results are the same for the two images, the fist web page introduces a higher similarity.

4.3 Larger Scale Ontology Inference Aided by Spreading Activation Theory

We have discussed the issue that given an ontology, how to use DL based semantic reasoner for image retrieval. An enhance algorithm has also been proposed to further measure the concept similarity. The next issue is that, when bigger concept ontology is involved, the semantic matchmaking phase could become a bottleneck, as it takes longer time and much more memory to process the increasing concepts in the ontology. To solve the scalability issue, we move away from the traditional reasoners, and study the structure of the ontology for other inference models. The multi-modality ontology has a similar structure to semantic network, which also consists of concepts and relations. As the spreading activation inference model has been proved to be effective for inference in the semantic network of the cognitive field, a new ontology inference model based on the spreading activation procedure is proposed in the following subsections.

4.3.1 Spreading Activation Procedure

In the area of cognitive science, one popular form of knowledge storage in long term memory is semantic network[122]. Semantic network represents knowledge in the form of nodes and arcs. It is a declarative graphic representation that is used to present and reason over knowledge. While both semantic network and ontology share similar structure, the main difference is that, ontology is often specialized to certain domains. It represents concepts within the domain and the relations between the concepts. Semantic network does not have such constraint on domains. Information processing in a semantic network typically follows the SAT. The activation value of each and every node spreads

to its neighboring nodes. Given an initial input activating specific nodes of the network, after the spreading activation process finishes, each and every concept in the network will be activated with certain values depending on its relations to neighboring nodes. As an ontology is structurally similar to a semantic network and SAT has been proved to be efficient for inference in an ontology in the previous work[123], it is adopted as a natural choice of inference in the proposed multi-modality ontology.

The mechanism of the SAT is hereby defined formally as below: Given a source node x and a destination node y, the activation propagation process follows the formula:

$$I_y(t_{i+1}) = O_x(t_i) \times w_{xy} \times (1 - \alpha), \quad \alpha \in (0, 1) \tag{4.8}$$

where $I_y(t_{i+1})$ is the input of node y at time t_{i+1}, $O_x(t_i)$ is the output of node x at time t_i, w_{xy} is the link between nodes y and x, and α is a decay factor to represent the energy loss in the spreading activation process. A simplified SAT is that the output of the node y at time t_i is the input of the node y at time t_i, $O_y(t_i) = I_y(t_i)$. Thus, the entire spreading activation process can be summarized into the following formula:

$$O = [\mathcal{E} - (1 - \alpha)w^T]^{-1} I, \tag{4.9}$$

where $I = [I_1, \ldots, I_n]^T$ is the initial input to the network, w is the matrix representation of the user ontology whose element w_{ij} represents the link between concepts c_i and c_j, α is the decay factor, $\mathcal{E}$ is an $n \times n$ identity matrix of order n, and $O = [O_1, \ldots, O_n]^T$ is the final output vector of the spreading activation process in which O_i is the value of concept c_i obtained from the spreading activation process.

4.3.2 SAT based Ontology Inference

In our case, the surrounding text and low level features of each image are made instantiations of particular concepts in the multi-modality ontology. Given these associated concepts with their frequencies, a vector $I = [I_1, I_2, \ldots, I_n]^T$ is formed as the input of the spreading activation process to inference an image's relevance to c_q, where I_n, the input to the concept c_n, is calculated by

$$I_n = \frac{freq(c_n)}{\sum_{all\ c_n} freq(c_n)}, \tag{4.10}$$

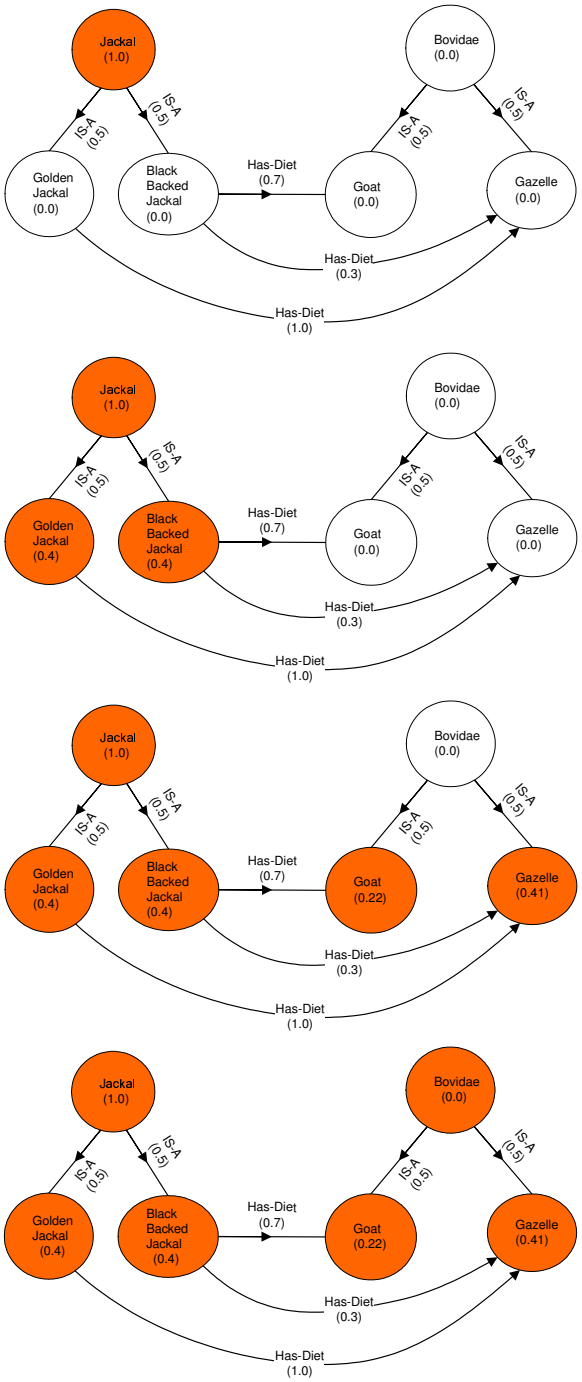


Figure 4.4: Stages of the spreading activation process(from top to bottom).

where $freq(c_n)$ represents the frequency of the concept c_n obtained in the image's surrounding texts or low level features. Upon receiving the input vector I, the spreading activation procedure is performed on the semantic network w to infer the relevance of the image to c_q. In our system, the configuration of the matrix representation w is described as follows. All the semantic relations are first extracted from a set of sample data set. The element $w'_{ij}s$ value of the matrix w is the frequency of the semantic relation r_{ij} in the data set. Then, the matrix is normalized using the following formula:

$$w_{ij} = \frac{freq(r_{ij})}{\sum_{all\, j} freq(r_{ij})}, \tag{4.11}$$

where $freq(r_{ij})$ represents the frequency of the relation r_{ij} in the data set.

Using the spreading activation formula (4.9), the activation value obtained for concept c_q is calculated. A high value represents that this image is more relevant to this concept c_q. These images are thus ranked according to their relevance values of c_q from the spreading activation procedure and return the result to the users. In experiment, after the users submit the queries to the system, the relevance of each image to the particular concept c_q of the query need to be computed. Images are then ranked according to their corresponding relevance values of c_q and returned to the users. In the system, the SAT based inference procedure[124] is used to compute the relevance values.

An illustration of the spreading activation procedure in the ontology is given in Figure 4.4. The node *Jackal* has been activated with an activation value of 1.0. Its activation then propagates across the entire semantic network following the spreading activation procedure. When the network stabilizes, the nodes in the network will be activated with certain activation values such as those shown in the fourth stage in Figure 4.4. Note that the activation value of each node does not depend solely on its distance from the initial node. For instance, the concept *Gazelle* obtains a higher activation value than that of *Goat*, which means *Gazelle* is considered more related to *Jackal* in this semantic network.

4.4 Experiment Result

In this section, given the proposed semantic matchmaking model, together with the ontology model proposed in Chapter 3, prototype web image systems is shown as an

implementation. System modules and work flow are discussed in details. A variation of the system called OntoEnhanced web image retrieval system is also given. Two major parts are changed: (1) The part of ontology construction part is changed from Manually Multi-Modality Ontology(ManuMMOnto) construction to Automatic Multi-Modality Ontology(AutoMMOnto) construction; (2)The inference model part is changed from the DL based reasoner to SAT based inference model. Later the retrieval result from both inference models is compared. Web images (and their corresponding web pages) are crawled from the top 200 hits returned by Google Image Search. The experimental data includes 20 different classes of animals under the *canine* family.

4.4.1 Experimental Systems

The following subsections explain the work flow model, system structure, and the modification part introduced by the AutoMMOnto construction and SAT inference model in the proposed web image retrieval system.

4.4.1.1 Work Flow Model

In this subsection, the work flow model of the image retrieval system is introduced. The detail structure is shown in Figure 4.5. There are three parallel modules in the first phase: Domain Knowledge Building Module, Text Context Analysis Module and Semantic Interpretations for Image Content Module. These modules communicate through the information flow. The second phase is composed of Semantic Matchmaking Module and Image Retrieval Module. These modules are wrapped as core of the Image Retrieval System. The main motivation of separating the model into two phases is to model each of the knowledge preprocessing and image matchmaking with retrieval aspects in a single frame. In this manner the framework enables the preparation of knowledge base and analysis of image information before proceeding to the image retrieval implementation. Meanwhile, the upper parallel structure supports concurrent data process for both high-level text information and low-level multimedia information. In the following paragraphs, each of the modules will be explained in details.

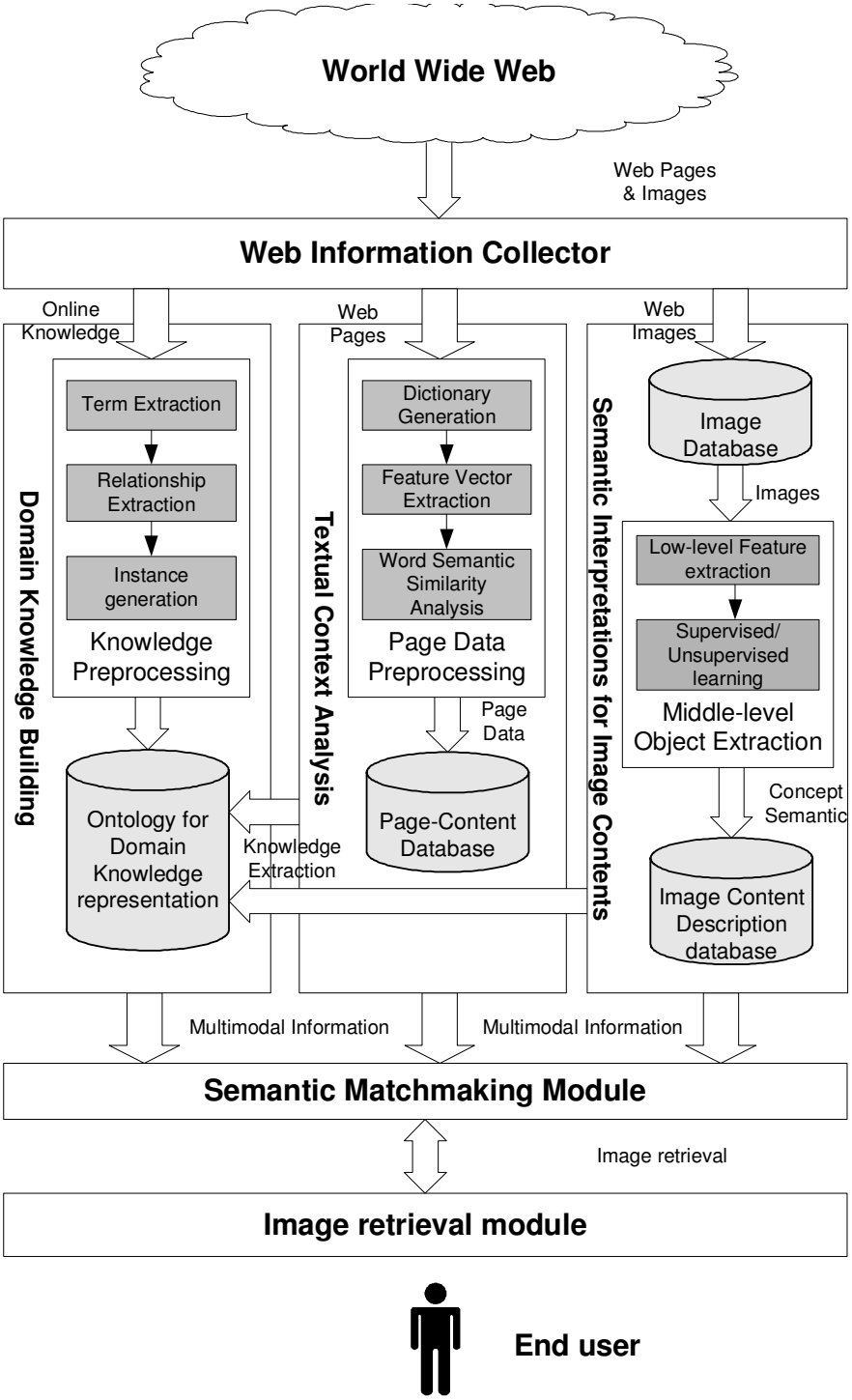


Figure 4.5: Work flow of image retrieval system

Domain Knowledge Building The first module discusses is the Domain Knowledge Building module. In this module, the online knowledge is imported from the BBC Science & Nature Animal category. Three steps are involved: Term Extraction, Relationship Extraction and Concept Generation. The knowledge regarding specific animal species is classified according to a unified scheme, from where the terms and the relationships are extracted. For instance, for terms *cape fox* and *South Africa*, a relationship of *hasDistribution* is extracted. Afterwards, concepts are defined and generated according to the extracted knowledge.

Text Content Analysis After building the ontology, related web information is collected through Text Content Analysis Module and Semantic Interpretations for Image Content Module. The Text Context Analysis Module first collects the surrounding text from the web pages. The web texts contain much noise and most of them are loosely-coupled with the images in the web pages. It is challenging to retrieve the exact information which can help build up the text ontology to classify the image. Some works[62, 118] use probabilistic method, a popular approach, to process the text information. However, there is no guarantee that only information with clear semantic relationship is extracted. In the experiment, the explicit semantic relationships which are previously defined in the knowledge base are reused and the RACER[108] semantic matchmaker is implemented to extract the concepts and relationships for further reasoning. The outputs of the text content analysis constitute part of the final anonymous concept ontology.

Semantic Interpretation for Image Content The Semantic Interpretations for Image Content Module focuses on the processing of low-level features in images. The data source is the pre-crawled images, based on which low-level feature extraction and supervised learning are done. Through this learning process the images are classified according to their low-lever features. For example, the classification result will show whether an image is an outdoor scene with a gray object in the foreground. Again, the output results of this module are stored in an image content description database. In addition, these three modules are connected and communicate through information flows.

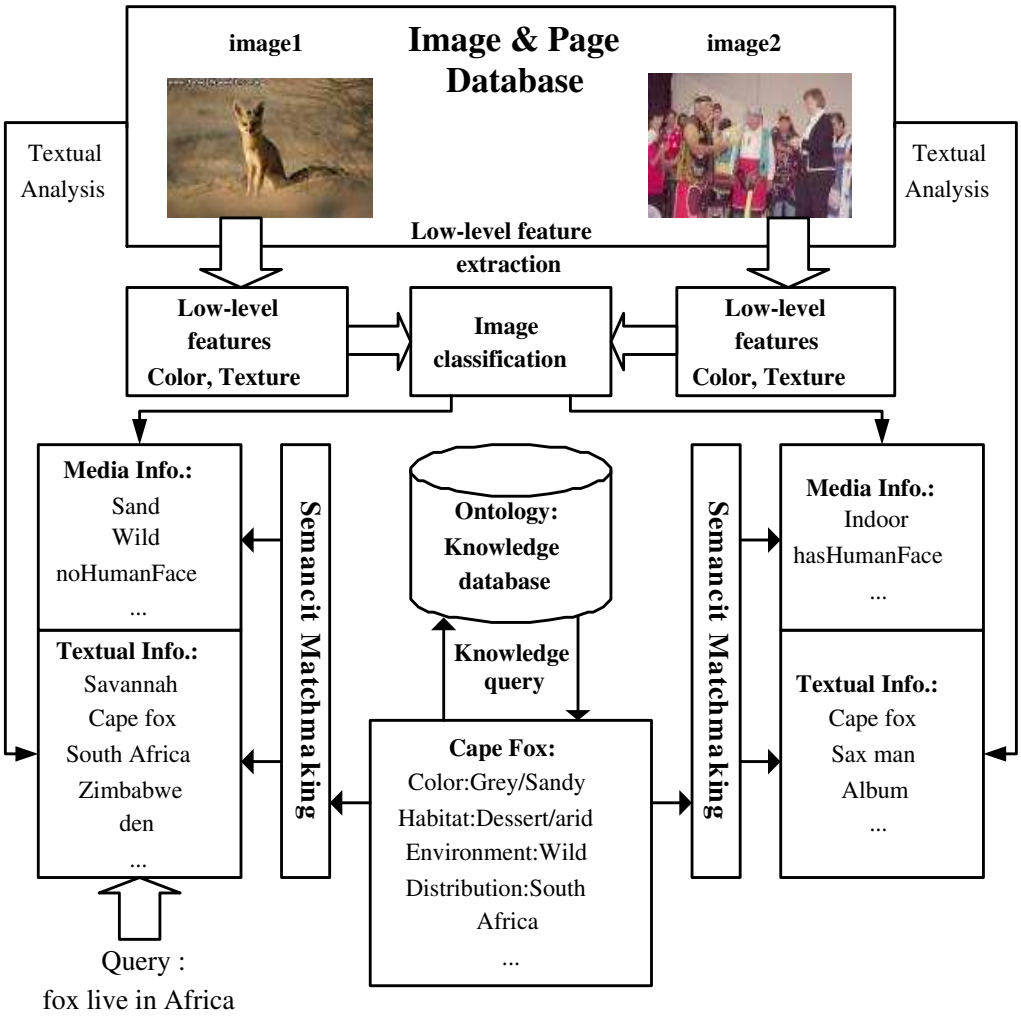


Figure 4.6: System structure of image retrieval system

Semantic Matchmaking and Image Retrieval After the data preparation phase, the next phase is the semantic matchmaking and image retrieval phase. As concepts of *canine* species have already been generated, the anonymous concepts from the images are used as the input and reason over the underlying knowledge base. Through the matchmaking it is able to judge whether the input image is a particular *canine* species or a super class of several subspecies. The former case is called an exact match and the latter is a subsume match. For instance, a *fox* in some images may be labelled with contradicting terms as *hasDistribution.SouthAfrica* and *hasFur.WhiteFur*. Therefore, the only conclusion is that this image may contains a *cape fox* or an *arctic fox*, as the first species normally has distribution in South Africa while the second species, in most cases, has white fur. By digging and incorporating information from both text and image aspects, the inference model avoids making an arbitrary decision, which is the virtue of semantic matchmaking. Those exact matches are ranked highly in our list and they are followed by the subsume matches.

4.4.1.2 Overall Structure of the Image Retrieval System

The overall structure of the proposed image retrieval system is shown in Figure 4.6. There are two image examples given to illustrate how the web images are stored in the image database and how they are retrieved according to users' input. It can be seen that *image 1* is a *cape fox* image while *image2* is an image describing human activities, though there are keywords like "cape fox" in the surrounding text of *image2*. After all the image content analysis and text analysis, *image1* is classified as a *cape fox* image while *image 2* is classified as a non-animal image and both the images are stored in the database with their labelled image concepts. The retrieval process of our system can be described as follows: The user keys in free text indicating his query target. The coordinator program invokes a query expansion service and passes the expanded concept to the knowledge base. The concepts contained in the knowledge base are matched with the ontologies and those images that satisfy the query are retrieved from the database. Finally matching results are displayed to the user. Here is a simplified example of this process: The user gives the free text "fox live in Africa" as input query to the image retrieval system. Terms in this query are sent to the reasoner and semantic relationships

are extracted. An inference is drawn that the user is looking for images of *fox* whose area of distribution is Africa. This further leads to a target animal *cape fox*. This target, which is originally represented by several terms, is now expanded to the *cape fox* concept which has all the corresponding relationship definitions inside. This query concept is the same as the predefined *cape fox* concept in the knowledge base, which is the multi-modality ontology including all the predefined concepts and relationships.

On the other hand, an image crawler program has already been used to download the target web images along with corresponding web pages as the database. image content analysis module respectively. In the text content analysis, a stop list is first created for the web pages to filter off all the redundant information such as prepositional phrases and character sequences of numbers and punctuation. These information are considered as redundant and do not provide useful information in constructing the ontology. The next step is to do the stemming work, which reduces inflected words to their root forms. The cleaned terms in the web pages are sent into the reasoner to find the relations between each other. Based on the pre-defined knowledge base, the reasoner automatically extracts the words and relations and groups them into triples of subject, predicates and object. These triples are later transformed by our program to ABox and form part of the *AnonymousConcepts*. In image content analysis, a set of labels are assigned to every image. According to the foreground and background of an image, labels like *hasContent.wildlifescene*, *hasPixProperty.graph*, and *hasPixColor.grey* are attached to the image. These labels are assigned to corresponding semantic relations, which further constitute part of the anonymous concept. The final generated instance is a semantic representation for the image and its surrounding text. After semantic matchmaking the relationships between the anonymous concepts and the concepts in the ontology are revealed, and images like *image 1* which satisfies the query concept are retrieved from the database. To illustrate the matching concepts in a semantic matchmaker, some simplified

examples in the image concept knowledge base are shown as follows:

$$
\begin{aligned}
Fox &\doteq Animal \sqcap \forall hasName.FoxName \\
CapeFox &\doteq Fox \sqcup \forall hasEnironment.wild \\
&\quad \sqcup \forall hasName.(CapeFoxName \\
&\quad \sqcup TreeFoxName) \sqcup Vulpeschama) \\
&\quad \sqcup \forall hasDiet.(insect \sqcup mammal \\
&\quad \sqcup fruit \sqcup carrion) \\
&\quad \sqcup \forall hasHabitat.savannah \\
&\quad \sqcup \forall hasDistribution.Africa \\
&\quad \sqcup \forall hasColor.Grey \\
&\quad \sqcup \forall hasPixColor.Colorimg \\
&\quad \sqcup \forall hasPixProp.Image \\
&\quad \sqcup \forall hasEnvironment.Wild \\
&\quad \sqcup \forall hasFur.Grayfur \\
&\quad \sqcup \forall hasPixContent.Nonbuilding
\end{aligned}
$$

Note that the objective part of the semantic relations, such as Color, Environment, Distribution, etc., all have their structured ontologies defined according to general knowledge. It can be inferred that if an *anonymous fox* is defined as *hasDistribution. SouthAfrica*, the image will likely be a *cape fox* image as *South Africa* is a subclass of *Africa*, and *cape fox* is usually found in *Africa*.

4.4.1.3 The OntoEnhanced Web Image Retrieval Prototype System

This section presents a variation web image retrieval system as OntoEnhanced. The difference of the proposed OntoEnhanced system includes the AutoMMOnto construction and SAT inference model. The major structure and work flow is very similar to the description in Section 4.4.1.2, except the part of ontology construction and inference. A basic system structure is shown in Figure 4.7. This system includes two components, namely an offline ontology construction part and an online image retrieval part.

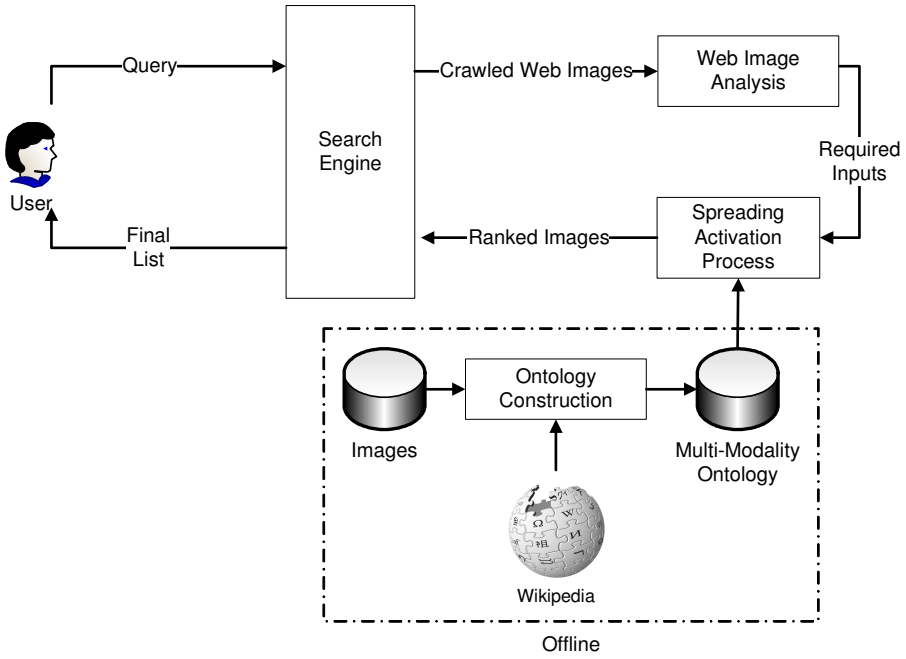


Figure 4.7: Illustration of the prototype system for OntoEnhanced web image retrieval.

The offline part generates the multi-modality ontology as follows: First, a set of relevant concepts to the target domain with their associated semantic relations, including taxonomy and non-taxonomy relations, are extracted from Wikipedia. Then, other concepts related to the low-level features of the images, together with their relations are obtained from a set of training images. Finally, the concepts and relations extracted in the two steps are combined to form the final multi-modality ontology. A detailed description of the ontology construction part has been given in Section 3.2.

For the online image retrieval part, the users first submit queries to the system. The queries would be either particular concepts in the multi-modality ontology of free texts as ontology can easily convert the free text into the concepts[123, 125]. Then, the search engine fetches all the crawled images in the database, analyses these images, and produces the required inputs for the spreading activation procedure. After inference is done on the multi-modality ontology, the relevance of each image to the submitted query is calculated. Finally, these images are ranked according to their relevance values and returned to the users.

4.4.2 Comparison of Different Matchmaking Results

The comparison of different matchmaking result includes two parts. The first part shows the improvement brought by the enhanced ranking algorithm to the DL based reasoners. The second part compares the result generated by enhanced ranking algorithm with the SAT based inference model.

4.4.2.1 Matchmaking with Enhanced Ranking Algorithm

It can be seen from Figure 4.8 that the retrieval performance is improved by the enhanced ranking algorithm. The baseline ranking is taken from Google Image Retrieval result and Figure 4.8 shows details of the returned ranking results for six *canine* classes. In this figure, the *M-M Ontology* is short for *Multi-Modality Ontology*. The enhanced multi-modality ontology with its independent ranking mechanism outperforms the one without the ranking mechanism.

An overall comparison of precision is also given . In web image search and retrieval systems, precision is a very importation guideline to measure the performance of a retrieval system because most users can browse limited numbers of results and therefore recall rate for a web image retrieval is not crucial. In order to give a whole picture of the performance, the correct number of images retrieved in the top 20, 40, 60 and 80 for all twenty *canine* subspecies are shown. From the Figure 4.9, it can be seen that in the precision test results for all categories, most groups have better or comparable results with the previous multi-modality model result. It can also be seen that almost all results returned by the multi-modality ontology with ranking correlation are correct for the top 20 images. In most cases, either of the ontology-based image retrievals can achieve better retrieval precision than keyword based Image Search. Note that the total number of correct image in the ground truth of *bush dog* is 31, which makes the *bush dog* result relative low in the Top 80 Image Retrieval Result. According to Figure 4.9, by using multi-modality ontology the retrieval precision is improved by about 5 to 30 percent. A snapshot of six groups of top retrieval results is also shown in Figure 4.10.

4.4.2.2 SAT based Inference Model

In our experiment, the size of the semantic network includes around 743 concepts and 872 relations. As comparison, the advantage of SAT over RACER is that it can support

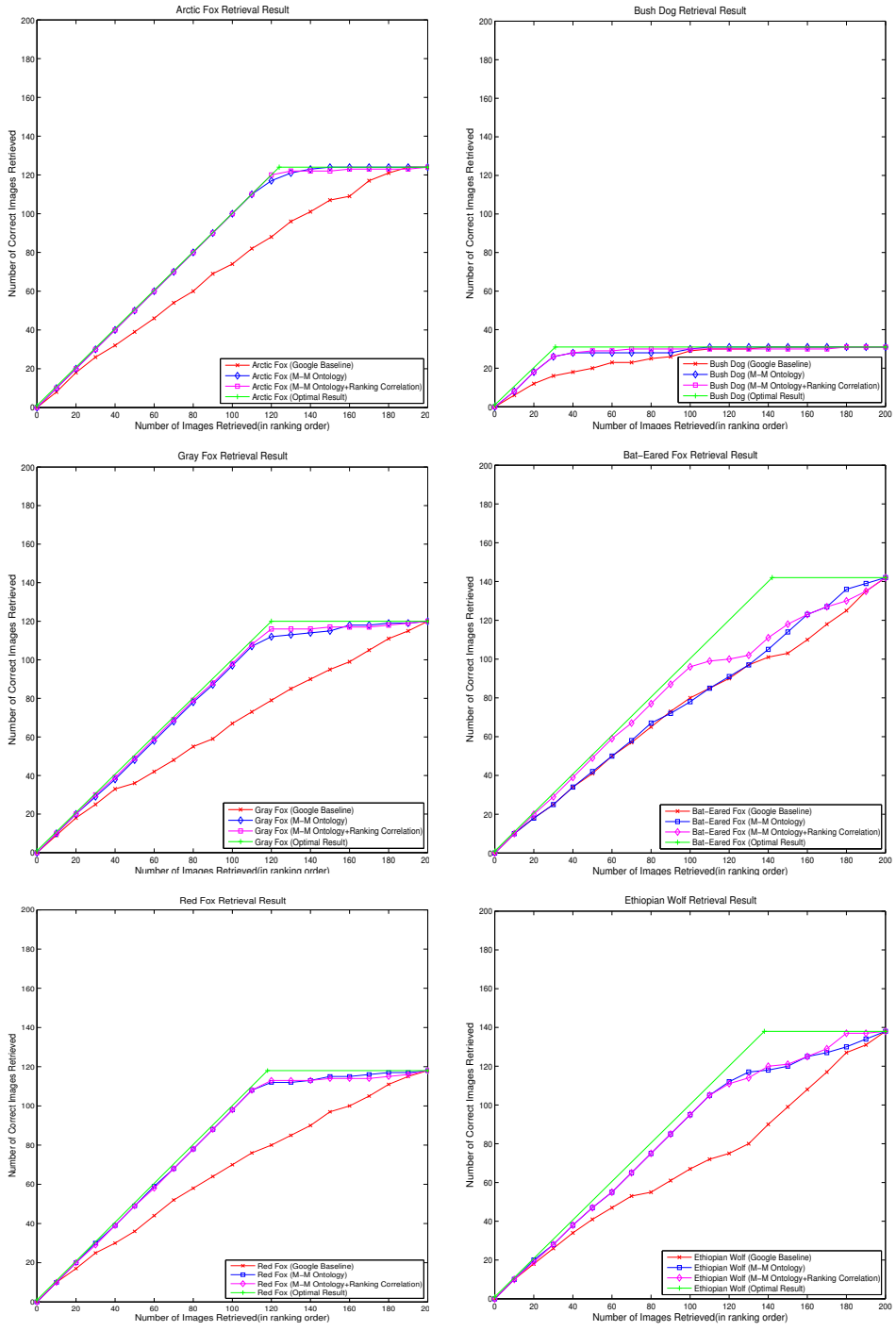


Figure 4.8: A comparison of image retrieval results between multi-modality ontology and multi-modality ontology with ranking correlation approaches(1)

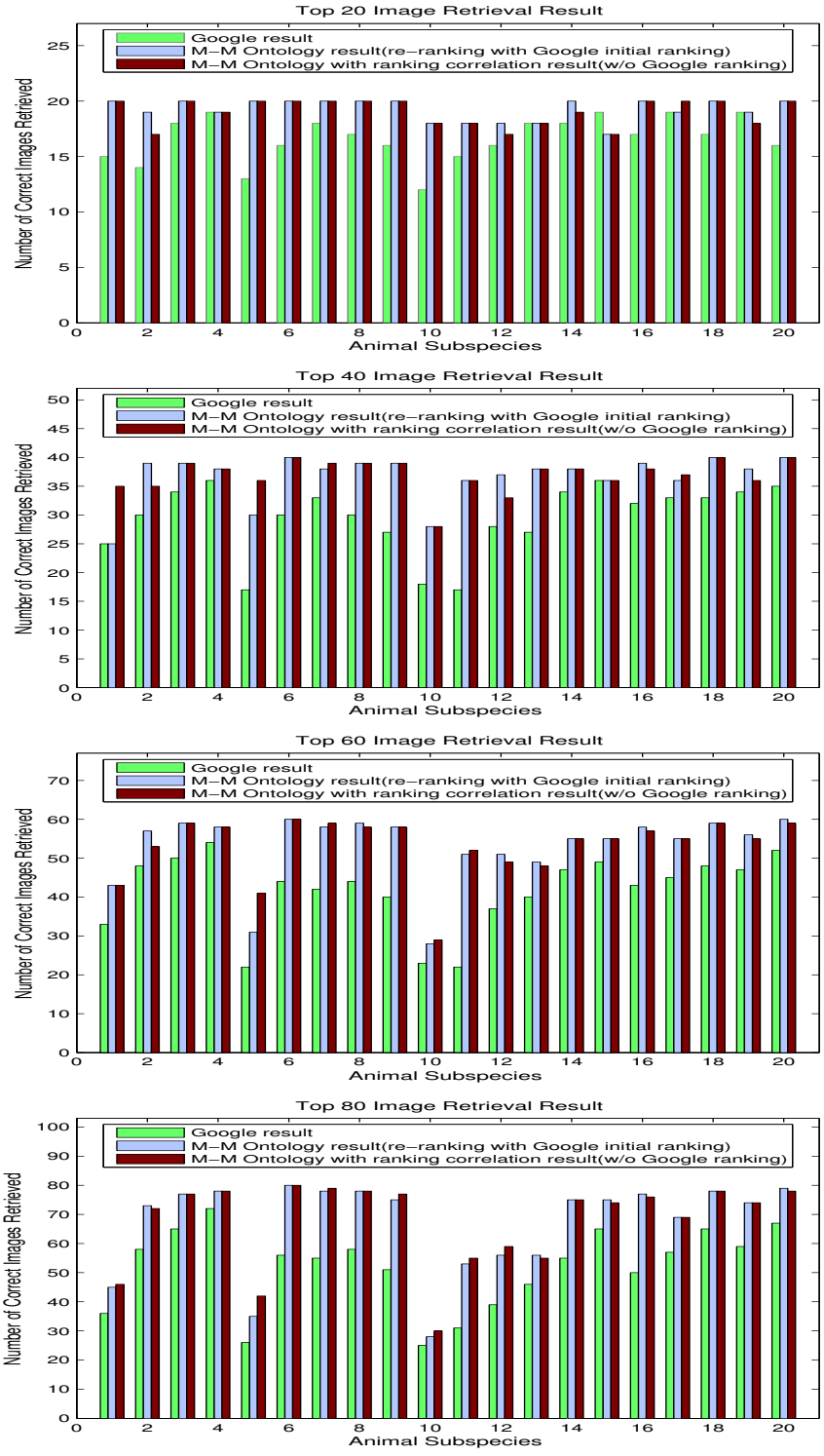


Figure 4.9: A comparison of image retrieval results between different approaches (2): The 20 subspecies are: 1.*Aardwolf*, 2.*African wild dog*, 3.*bat-eared fox*, 4.*black backed jackal*, 5.*cape fox*, 6.*Arctic fox*, 7.*grey fox*, 8.*red fox*, 9.*kit fox*, 10.*bush dog*, 11.*coyote*, 12.*dhole*, 13.*dingo*, 14.*Ethiopian wolf*, 15.*fennec fox*, 16.*golden jackal*, 17.*grey wolf*, 18.*maned wolf*, 19.*red wolf*, and 20.*spotted hyena*.

Table 4.2: Performance of image classification on different inference models

Class	Aardwolf	Cape Fox	Bush Dog	Arctic Fox	Ethiopian Wolf
Google	0.5801	0.4958	0.4695	0.715	0.7516
SAT	0.7167	0.5914	0.8611	0.9404	0.862
Class	Coyote	Gray Wolf	Gray Fox	Fennec Fox	Spotted Hyena
Google	0.5042	0.7513	0.7183	0.8181	0.8365
SAT	0.7694	0.8003	0.8767	0.8237	0.8591
Class	Dhole	Red Fox	Maned Wolf	Black Jackal	Bat-Eared Fox
Google	0.6342	0.744	0.7949	0.8872	0.7967
SAT	0.7546	0.889	0.8662	0.9271	0.8856
Class	Dingo	Kit Fox	Red Wolf	Golden Jackal	African Wild Dog
Google	0.67	0.6698	0.7669	0.7092	0.7844
SAT	0.7109	0.7981	0.8768	0.8793	0.8753

matching of ontology with large scale concepts. When using RACER, there is no bound of the memory consumption. The matchmaking program will gives an out of memory error when the loaded ontology contain many concepts. On a computer with a CPU of 2.67GHz and 2 GB of Ram, we do the matchmaking test on a single query to see the performance of both RACER and SAT on a HP xw4400 Workstation, with CPU of 2.67GHz and 2 GB of Ram, with Windows XP operating system and no other applications running. One GB of memory has been assigned to the program. RACER shows out of memory error in this case. For the SAT program, it takes 220 milliseconds to process. This result reveals that RACER is good at processing domain ontologies with limited concepts. When the ontology is moving towards larger scale, SAT is a better choice especially when memory usage is a concern. To give an evaluation of the overall performance of the SAT based inference model approach, the performance of the proposed method is evaluated using AP, which is defined as the averages (interpolated) precisions at certain recalls. A better retrieval performance, with a Precision-Recall(PR) curve staying at the upper-right corner of the PR plane, will have a higher AP, and vice versa. In the current experiment, $j = 200$ is set. In Table 4.2, the AP from both Google retrievals and ontology enhanced web image retrieval (OntoEnhanced) are shown. It is clear that the proposed method performs consistently well.

4.5 Summary

In this chapter, the different approaches of semantic matchmaking are studied. In view of the existing problems, an approach of using DL based semantic reasoner is first introduced. Later the SAT based procedure for inference on the ontology is described. Two prototype systems that combines the ontology and inference models for web image retrieval are implemented. The experimental results demonstrate the efficacy of both semantic matchmaking models. It is shown that the proposed OntoEnhanced approach with SAT as underlying inference model will help to improve the retrieval performance for images of various domains. The proposed approach largely dispenses with the conflict between cost and precision in ontology-based applications.

Figure 4.10: Some top retrievals from original Google Image Search and our proposed multi-modality approach

Chapter 5

Building Large Scale Concept Thesaurus

So far in previous chapters domain-dependent multi-modality ontology with corresponding matchmaking models have been proposed. In Section 3.2 Wikipedia has been proved to be an effective knowledge source to help generate ontology automatically. The result has also proved that the generated ontology provides good performance in web image retrieval task. This chapter proposes a new approach to build a large-scale concept ontology in the form of *concept thesaurus* from Wikipedia corpus dump XML files[1]. This thesaurus, which contains extensive concepts with various relations between them, will be used to discover salient semantics and calculate the similarity between concepts. The proposed approach allows incorporation of semantic concepts and relations from different domains and levels, thus provides us with richer information for concept matching. This proposed *concept thesaurus* aims to provide extensive vocabularies for describing extended domains of real-world web.

Example web images, with diverse object, foreground, background, illumination effects, and noisy text descriptions are shown in Figure 5.1. It is not easy to model the image content purely from low-level features due to the well-known *semantic gap*. It is necessary to use a formal semantic representation so that machine can understand web image in a better way. Towards this goal, many researches focus on finding the proper semantic representation for image information. The previously proposed multi-modality ontology is such an example. Besides building one's own ontology, with the emergent

[1] http://download.wikimedia.org/enwiki/

Figure 5.1: Sample images from our data set(from top to down, left to right): Cape Fox, Ethiopian Wolf, Gray Fox, Leopard.

of large collaborative encyclopedias projects, much effort[92, 97, 104] has been devoted to mining semantics from lexical dictionaries and online encyclopedias for information classification and retrieval. However, many of the above works do not fully leverage the *semantics* that could be provided by *concept ontology*. The main reasons are that both the concept coverage and relation diversity are limited by the most frequently used external knowledge sources, e.g., WordNet, which should be improved in order to address the needs of real-world media understanding. Additionally, an ideal ontology should adopt concept detection, disambiguation, and polysemy. Relations other than the hierarchical structure between parent class and child class should also be exploited. Therefore some researches[126, 127] promote the use of large online encyclopedia and automatically construct ontology from the Wikipedia article pages. While it is a promising solution, it raises other challenges like scalability and causality issues, the details of which will be given in later discussions. In view of the *concept thesaurus* construction, the following requirements should be fulfilled:

- The concept coverage of the thesaurus should be large enough to cover extensive domains. The definition of concept should follow general knowledge.
- The thesaurus should define semantic relations so that it can be used to discover connections between concepts.
- The thesaurus should support concept detection and concept disambiguation.

When applying the thesaurus to web image retrieval, algorithm will be proposed to fully utilize the semantics in the thesaurus. The idea is to mine semantic concepts from the web image and calculate the similarity between the concepts for retrieval purpose. Therefore, the proposal mainly involves two phases. The first phase is *concept thesaurus* building. Vocabularies are collected based on Wikipedia articles. For each concept, the set of related concepts are extracted and categorized according to the relations. The purpose is to set up a network between the concepts. The second phase is concept distance calculation based on the generated thesaurus. Concepts are first detected. Each web image is then represented by a concept vector. A semantic space is constructed according to the generated vectors. The distance between the concept vector are calculated to measure the concept similarity. In this work, in order verify the effectiveness of the thesaurus model, the experimental web image database is extended to 13,856 web images including 26 classes of animals.

5.1 Thesaurus Structure Extracted from Wikipedia Corpus

This section presents the efforts in building the Wikipedia-based *concept thesaurus*. The knowledge source to extract information is the Wikipedia dump file. The dump file is offered as free copies to Wikipedia users. As Wikipedia is a global web based multilingual free content encyclopedia, there are multiple language dumps available. In this work the English Wikipedia(*enwiki*) is downloaded and used. The downloaded database backup dumps contains all the Wikipedia text source and metadata wrapped in an XML format. The detailed contents include page content, page links, category links and image links between pages, image metadata such as image URLs and history, and some miscellaneous information for the web site maintenance and status records. The downloaded file has a size of 12.9GB after unzipped and therefore it is not possible to treat it like normal text files. In order to read the content of the XML file, a Wikipedia dump reader program is developed to analyze the file by parts. The next step is to build the thesaurus. The main approach includes concept extraction, refinement, and relation detection.

<page>
<title>Arctic fox</title>
<id>2208</id>
<revision>
<id>165197478</id>
<timestamp>2007-10-17T15:27:39Z</timestamp>
<contributor>
<ip>194.144.101.5</ip>
</contributor>
<text xml:space="preserve">{{Taxobox | color = Purple | name = Arctic Fox | trend = stable | image = Polarfuchs 1 2004-11-17.jpg | image_width = 200px | regnum = [[Animal]]ia | phylum = [[Chordate|Chordata]] | classis = [[Mammal]]ia | ordo = [[Carnivora]] | familia = [[Canidae]] | genus = """Alopex""" | genus_authority = [[Johann Jakob Kaup|Kaup]], 1829 | species = """A. lagopus""" | binomial = "Alopex lagopus" | binomial_authority = ([[Carolus Linnaeus|Linnaeus]], [[Systema Naturae|1758]]) | range_map = Distribution arctic fox.jpg | range_map_width = 200px | range_map_caption = Arctic Fox range | synonyms="'Vulpes lagopus'" }} The '''Arctic Fox''' "(Alopex lagopus)", also known as the "polar fox" and "White fox", is a [[Fox]] of the order [[Carnivora]]. It is a small [[fox]] native to cold [[Arctic]] regions of the [[Northern Hemisphere]]. It is common in all three [[tundra]] biomes. Although some authorities have suggested placing it in the [[genus]] "[[Vulpes]]", it has long been considered the sole member
.....
{{wikispecies|Alopex lagopus}} [[Category:Arctic land animals]] [[Category:Foxes]] [[Category:Mammals of Canada]] [[Category:Mammals of the United States]] [[Category:Fauna of Western United States]] [[Category:Mammals of Europe]] [[Category:Fauna of Greenland]] [[Category:Fauna of Russia]] {{Link FA|is}} {{Link FA|no}} {{Link FA|nn}} [[bg:Полярна лисица]] [[cs:Alopex]] [[da:Polarræv]] [[de:Polarfuchs]] [[el:Αρκτική αλεπού]] [[es:Alopex lagopus]] [[eo:Arkta vulpo]] [[fa:روباه قطبی]] [[fr:Renard polaire]] [[is:Heimskautarefur]] [[it:Alopex lagopus]] [[he:שועל שלג]] [[la:Alopex lagopus]] [[lt:Poliarinė lapė]] [[nl:Poolvos]] [[ja:□□□□□□□□]] [[no:Fjellrev]] [[nn:Fjellrev]] [[pl:Lis polarny]] [[pt:Raposa-do-ártico]] [[ru:Песец]] [[sl:Polarna lisica]] [[sr:Поларна лисица]] [[fi:Naali]] [[sv:Fjällräv]] [[tr:Kutup tilkisi]] [[zh-yue:□□□]] [[zh:□□□]]</text>
</revision>
</page>

Figure 5.2: An illustration of the Wikipedia dump file content for concept *Arctic fox*.(Extracted from the dump file enwiki-20071018-pages-articles.xml)

5.1.1 Concept Extraction and Refinement

By analyzing the dump file content, it can be seen that each entry represents a Wikipedia article content. The article is considered as one single Wikipedia concept. An abridged snapshot of the concept *Arctic fox* in the dump file content is shown in Figure 5.2. For each entry, the "title" part indicates the concept name. Some miscellaneous information includes the Wiki ID of the concept, time stamp when the article is created, and the ip of the contributor. For this concept, the page text content part starts with a taxobox, which contains taxonomy information for concepts under the *animal* domain. The following text is extracted from the original Wikipedia article. Related concepts to the article concept are highlighted by XML tags. At the end of the entry, category information is given according to the general categorization system in Wikipedia. The category information is provided in different languages and in the following work the main focus is on the categorization in English.

The thesaurus construction starts from concept extraction. The "title" part of each Wiki entry are first retrieved and indexed as Wikipedia concepts. As an online collaborative work, Wikipedia is edited every minutes by a large number of users. Noisy information is inevitably involved. Not all the concepts are meaningful to the proposed *concept thesaurus*. Among these extracted concepts, those serve for Wikipedia administration purpose such as *Wikipedia administration*, together with some miscellaneous concepts are removed, such as those with punctuation, pure number concepts, and so on. The final concept number arrives at 5,836,166.

5.1.2 Relation Detection

The Wikipedia main category has provided a fine taxonomy for all the concepts. Following the original category, a hierarchical structure is easily obtained. As we have argued, more relations should be considered to include in the final ontology. Therefore, for each concept entry in the thesaurus, concepts that are under the following four semantic relations are collected: synonymy, polysemy, parent concept/category, and associated concept.

The synonymy is built from the special links from the Wikipedia web page of the entry concept. In Wikipedia, all the synonymies of the concepts are associated with the original

concept though a *redirect* link. Hence it is straightforward to find the synonymies. There is one issue worth noticing. In the previous Wikipedia ontology automatic construction experiment, synonymies are also collected by analyzing the HTML tags of the downloaded Wikipedia web page. It is found that words in the first sentence which are emboldened are normally alternative names or synonymies to the entry concept. However, same rule is not followed in the *concept thesaurus* construction. There are mainly two reasons: (1) The emboldened HTML tag information is not included in the XML dump file; (2) Not all of the emboldened words are Wikipedia concepts. Therefore, they are not shown as entry concepts in the XML dump files. Manual work will be inevitably involved to define those concepts separately. As a result, the *redirect* link information become the only source to decide synonymy of the concept. Taken the concept *gray wolf* as an example, a redirection is made from the Wikipedia concept *wolf*, *timber wolf* and *grey wolf*. Therefore I define the three concepts as synonymies to each other.

Polysemy is extracted from the disambiguation information. Concept disambiguation has been discussed intensively in ontology construction as many concepts have multiple meanings in different contexts. Wikipedia also provides an easy way for *concept disambiguation*. If a concept has several meanings, a search will first return the web page whose content describes the most common explanation of the concept. Meanwhile, a disambiguation page is also given as supplement reference. This makes it easy to discriminate the different meanings among different concepts. For example, the concept *Red Fox* has several polysemies, such as *Emmett McLemore*, which is the name of an American football player whose nickname is *Red Fox*. Another example is the concept *Jaguar*, whose polysemies include *Jaguar Car*, which is a brand of luxury motor cars.

The basic structure for Wikipedia to organize its concepts is a hierarchical categorization. This structure is depicted as the *category graph* which has been proved to be a scale-free, small world graph by graph-theoretic analysis[105]. The *category graph* is formed following the taxonomy of concepts. Therefore the links in *category graph* indicate either hypernymy/hyponymy(ISA) or meronymy(PARTOF) relationships between the two connected concepts. In this sense, the semantic relationships provided by the *category graph* is quite similar to the main relationships provided by most lexicon dictionaries, such as WordNet. The concepts extracted from Wikipedia category information

lists all the parent concepts of the entry concept. Compared to hierarchies which are generated from other lexical resources, the hierarchy generated from Wikipedia *Scientific classification* is more formally defined, and thus is considered to contain rigid domain information. Normally each Wikipedia concept belongs to several categories. The miscellaneous categories such as *Special:Uncategorizedpages* and *Special:Mostlinkedcategories* are removed as they are mainly kept for Wikipedia maintenance purpose. An example of concepts linked by the parent concept/category relationship is the child concept *Arctic Fox* with the parent concepts *Foxes* and *Arctic land animals*.

The associated concepts are extracted from the paragraphs on the web page. The hyperlink tags are used to identify these concepts. The benefit of introducing the associated concepts are as follows: 1. Since the associated concepts are existing Wikipedia concepts, most of them are meaningfully associated concepts. No further pruning or selection of concepts are needed. 2. Compared with WordNet which works on concept synnet and contains mostly hierarchical relations, this set of associated concepts leads to an extension of relations. The associated concepts cover relations from various aspects. An example is given to illustrate the benefit. Taking the Wikipedia concept of *Aardwolf* as an example, its associated concept set contains *mammal*, *Southern African*, *Striped Hyena*, *bush land*, *muzzle*, etc., most of which are highly related concepts, with relations ranging from synonymy to habitat.

After concept extraction and relation detection, for fast thesaurus searching, all the concept entries are indexed and stored. The index is constructed based on relations. For an input query concept, its synonymy and polysemy (if any), its parent and child concepts, together with a set of associated concepts can be easily retrieved from the thesaurus. The child concepts is retrieved through an inverse query of the indexed thesaurus. Given an concept, all the available child concepts for the input are returned. This indexed concept thesaurus is by nature a large-scale concept ontology. Compared to the commonly used lexical dictionary WordNet, which mainly organizes word concepts according to synset, the constructed thesaurus provides a more formal classification of words and phrases. As a result, the extracted concepts and relationships are closer to a formal ontology with various semantic relationships.

5.2 Concept Distance Calculation based on Wikipedia Thesaurus

After the *concept thesaurus* construction, the next issue is to apply this *concept thesaurus* to the web image retrieval task. This section explains the thesaurus enhanced model of concept distance calculation.

5.2.1 Semantic Concept Detection from Text in Web Page

The first step is to deal with the text information on the web page. In traditional text based approaches, multi-word concepts are broken into single words. The semantic meaning contained in these words become blur, which further affect the text feature generated. We overcome this shortage by using the thesaurus to find the multi-words concept in the web pages. Intuitively, by finding the concepts from adjacent terms, concept ambiguity will be removed. For example, taking a text corpus with "jaguar car" as an example, by human judgment we know that if we are looking for the animal *jaguar*, this could be a miss as in most cases the web page and the image are about a car brand Jaguar. With the concept thesaurus, machine is able to consider in a similar way. A moving window is used to match adjacent words with the thesaurus. Starting from the beginning of the term sequence, words are first merged into candidate concepts by different window size. A maximum window size of 4 is set, and iteratively decrease this value until a matched concept is found from the thesaurus. If there is no match when the window size becomes 0, the term is considered as a concept by itself. The index moved forward to the next term. This process continues until the whole web page text is scanned. By this process the potential concepts are found to replace the original separated terms. Fig5.3 shows such an example. Concept *Jaguar Cars* are found as a concept under the categories of *British brands, Car manufacturers*, among other categories. Some of its associated concepts include *Ford Motor Company, automaker, car manufacturer*.

After the semantic concept extraction phase, a dictionary is generated according to the global and local frequencies of the concepts in the improved *concept* corpus and transform each web page to a concept vector. This vector is the basis for further semantic enhanced similarity calculation.

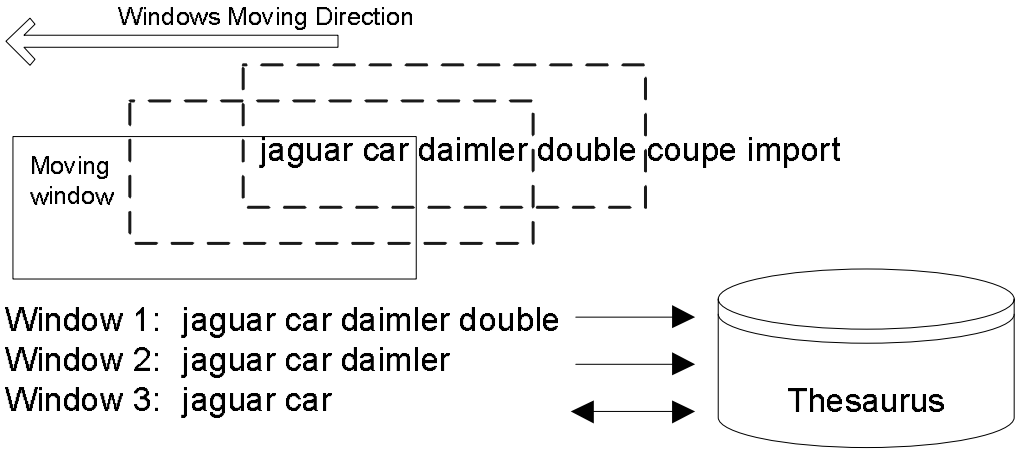


Figure 5.3: An illustration of the moving window for semantic concept extraction.

5.2.2 Semantic Salient Similarity Calculation

The next step is to use the *concept thesaurus* as the knowledge base for concept distance calculation. Now each web page is converted to a concept vector whose entry value is the concept frequency weighted by its inverse document frequency. In most cases, the size of the concept dictionary D is around 1,000, which is considerably large. The relations between concepts could offer positive influence if more emphasis is put on the frequencies of the related concepts. The question becomes how to find the related concepts. Using the limited and noisy web page samples, it is not an easy task to find the important concepts purely from clustering algorithms. The entry information in the thesaurus for each concept allows direct identification of related concepts, while the different semantic relations provide ways to assign weights to the related terms. CF is defined as the document-concept matrix, where each row represents one document, and each column represents a concept in the dictionary. cf_{ij} is the weighted frequency of concept j in document i.

$$CF = \begin{bmatrix} cf_{1,1} & cf_{1,2} & \cdots & cf_{1,d-1} & cf_{1,d} \\ cf_{2,1} & cf_{2,2} & \cdots & cf_{2,d-1} & cf_{2,d} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ cf_{n-1,1} & cf_{n-1,2} & \cdots & cf_{n-1,d-1} & cf_{n-1,d} \\ cf_{n,1} & cf_{n,2} & \cdots & cf_{n,d-1} & cf_{n,d} \end{bmatrix} \tag{5.1}$$

Given one query concept, the concept is first sent to the *concept thesaurus*. T denotes the set of related concepts, which includes the synonymy, polysemy, category and

associated concepts. A *thesaurus vector* V is constructed by mapping this concept set T to the concepts in the dictionary.

$$V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{d-1} \\ v_d \end{bmatrix} ; \text{For any } v_i = \begin{cases} 1, & D_i \in \text{T} \\ 0, & D_i \notin \text{T} \end{cases} \tag{5.2}$$

V is of the same length as the generated dictionary D. The entry of the *thesaurus vector* is either 1 or 0 (with non-zero entries indicating a mismatch of D element in T) depending on whether the concept can be found in T.

W is a diagonal weight matrix giving the concept weight according to the relations. The weight is designed according to the semantic significance of relations. For example, concept related with synonymies are having a higher weight, while the associated concepts are given lowest weight.

$$W = \begin{bmatrix} w_{1,1} & 0 & \cdots & 0 & 0 \\ 0 & w_{2,2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & w_{d-1,d-1} & 0 \\ 0 & 0 & \cdots & 0 & w_{d,d} \end{bmatrix} \tag{5.3}$$

Thus, a semantic salience matrix S is obtained as:

$$S = W \times V \tag{5.4}$$

A transformation of the original document-concept vector is performed as:

$$F = CF \times S \tag{5.5}$$

From this transformation, the concept frequency information provided by less relevant concept is omitted, while more emphasis is put on the related concepts defined by the thesaurus. This is different from the approaches reported by other research work[126], which uses Wikipedia knowledge to update the original document-term matrix to a less sparse matrix by including the similarity from related terms. The main concern is that, by expanding the relations to more indirect related terms, rather than putting emphasis on

the existing related terms, the causality may reduce the precision of the retrieval result. To be more specific, if the concept contained in the original text corpus is expanded with a more general concept, which leads to a bigger positive returned document set, even though the recall of the query concept could be improved, the precision is sacrificed by the introduced noisy.

This trade-off can be further explained by a simple example: In case the user is looking for web images of *gray wolf*, and both *timber wolf* and *Canada* are returned by the thesaurus as relevant concepts. While it is safe to conclude that an occurrence of the concept *timber wolf* could indicate a positive match, the same conclusion can not be made on the concept *Canada*. As *Canada* is a more general concept than *timber wolf*, while taking the occurrence of *Canada* as a positive indication, more false alarm are inevitably included. It is from this consideration that more emphasis is put on existing concepts to make them salient compared to less relevant information.

5.2.3 Extending with Visual Similarity

Due to the noisy web environment, in most cases, searches based purely on text cues may give wrong results and the retrieval contains images with obvious false visual features. Such images includes pencil drawings, maps, indoor scene, etc.. However, due to the surrounding text, the images are also considered as highly relevant and are thus ranked among top retrievals. By leveraging the visual information from images, the image similarity could be more robust to noise, and the retrieval performance would be further improved. In the experiment, interest points first are detected using Harris-Laplace detector[32] which is scale invariant and detects corner-like regions in the images as interest point. In the next step, both color and shape information around the interest points are extracted. Since the experimental web images mostly consist of real world images without much professional image processing, the image samples have less saturated colors and diffuse lighting. A 36-dimension *opponent angle*[128] is used as the local color descriptor. *Opponent color* consists of two chromatic channels which range from red to green and blue to yellow, respectively. Given x as the space coordinates, the derivatives of the opponent colors are defined as:

$$O1_x = \frac{1}{\sqrt{2}}(R_x - G_x); O2_x = \frac{1}{\sqrt{6}}(R_x + G_x - 2B_x)$$

Opponent angle ang_x^o is defined as:

$$ang_x^o = arctan(\frac{O1_x}{O2_x})$$

It has been proved in [129] that *opponent angle* is invariant to specular variations and is thus more suitable to describe color information for real world image. Each interest point is taken as a center of a 20 by 20 patch, and the *opponent color* is extract for each image patch. If the image patch is outside the boundary of the image, the patch is shifted to ensure the feature extraction. In addition, SIFT descriptor is used to represent the shape information around the interest point. By normalizing and combining both *opponent angle* and SIFT features, a feature vector of 164 dimensions is generated. From empirical experience, k-means is used to generate 500 clusters and a vocabulary of 500 visual words is built based on the feature vector. Therefore features of each image are categorized into 500 bins. A feature histogram is defined for each image as follows:

$$h(i) = \frac{n_i}{N} i = 1, 2, ..., K \tag{5.6}$$

where for each image, n_i is the number of features falling into bin i, N is the total number of features, and K is the number of visual words. As a result each image is represented by a 500 dimension vector.

In the next step, the classifier are trained through cross validation. The image features classify the images into groups based on whether they are color or black-and-white image, whether they are graphs, drawings or photos, and whether buildings or human faces are contained in the images, etc.. Intuitively a perfect *Arctic fox* image should be a colorful photo, and human faces are unlikely to appear, and the image would preferable has a snow background. For classification, one third randomly selected images from the image set are used as training and testing samples for cross validation. Based on the classification result an optimal classifier is found for each visual concept. Using this classifiers, each image is classified with a set of decision values, which contributes to the final similarity. This decision value is taken as the measure of similarity from the visual feature part. After the visual similarity is generated, the final similarity is calculated based on the similarity from both text and image cues. A total text similarity is first calculated from text modality. For the image modality, the decision value from SVM classification result

is taken as a measure of the visual similarity, as greater absolute decision values are indications of greater confidence in the prediction. The final similarity is the sum of the total similarities from both cues.

5.3 Experiment and Result

This section discusses the experiment. data set, process and final result.The following result are given: 1) Result based on the traditional tf-idf approach[9]; 2) Result based on the thesaurus enhanced approach; 3) The result based on the thesaurus plus image features approach.

5.3.1 Experimental Data

The data set is extended to include 26 classes of animals. Altogether 13,856 web images and their associated web pages are downloaded from Yahoo! image search, which indexes over 1.5 billion images.[2] Table 5.1 shows the image category distribution information. Taking Aardwolf as an example, among all the 290 web images downloaded by using the animal name as search keyword, there are 87 images which contain Aardwolf.

A group of technicians and research students are invited to manually label the ground truth. The annotation contains information about background scene, foreground object class, object position.

5.3.2 Experiment Result

The baseline generation is based on the traditional tf-idf model, which is a statistical model of the distribution of the words in documents, weighted by the importance of the words. The importance is decided by the term's local and global frequency in the document and the text corpus. The preprocessing of the web page data includes stop words removal and word stemming. After the preprocessing step, each web page is presented by a set of words. A dictionary is generated according to statistics from both global and local term frequency. From empirical experience, those terms with a global frequency less than 10 and more than 10,0000 are removed. A very small global frequency

[2]Our data collection was done in October 2007.

Category	Positive/Total	Category	Positive/Total	Category	Positive/Total
Aardwolf	87/290	African Wild Dog	323/466	Arctic Fox	470/802
Bat-Eared Fox	183/278	Black Backed Jackal	256/340	Bush Dog	52/596
Cape Fox	71/434	Cheetah	182/735	Coyote	234/720
Dhole	238/775	Dingo	271/704	Ethiopian Wolf	88/130
Fennec Fox	170/287	Golden Jackal	85/154	Gray Fox	120/649
Red Fox	378/630	Gray Wolf	311/624	Jaguar	98/608
Kit Fox	174/543	Tiger	59/367	Lion	336/766
Leopard	284/768	Maned Wolf	167/236	Red Wolf	107/567
Snow Leopard	339/581	Spotted Hyena	357/448		

Table 5.1: Image category distribution information

indicates a term that a rarely used, while a large global frequency indicates a term which is too general. According to the generated dictionary, the text information for each web page is transformed into a vector, with each entry representing the weighted value of the term frequency. By random sampling one third of the positive samples, the average is taken as the concept center and the distance between each sample to the center is calculated. The final ranking is given according to this distance.

The second experiment aims at testing if semantics from web pages can be grasped through thesaurus enhanced semantic salient similarity calculation. It is interesting to see how the retrieval performance could be improved by considering relations between concepts, rather then words. The following example proves that the proposed method works well for the web image data set. Taken the two web pages in Figure5.4 as a simple example. From the initial data collection, the first web page on the left is a negative image with a high ranking, while the second web page is positive with a ranking below 400. Even though the content of the second web page indicates close relation to the concept *African Wild Dog*, it is still ranked lower according to the traditional keyword matching criteria. Through the thesaurus enhanced semantic salient similarity calculation, now a correct prediction of the true semantics is made, and the second image is considered more related to the concept *African Wild Dog*. The results further reveals that the proposed semantic salient similarity calculation can detect and highlight related concepts based on the query, and calculate the overall similarity based on domain knowledge. It is worth mentioning that by using encyclopedia, some interesting facts may be discovered. For example, most people are not aware of the fact that *Cape hunting dog* is the synonymy of *African Wild Dog*, while in the experiment this relation is counted into the semantic similarity calculation automatically. The final step is to add the visual similarity from image features and combine the similarities from two modalities together. Before the combination step, a few groups of images are tested purely based on low-level features. Some groups of precision-recall results are shown in Figure5.5. More results on average precision improvement is shown in Table5.2, which shows that the combination consistently outperforms separate efforts. For the thesaurus result, the improvement will be affected by the available information extracted from the thesaurus. Taken *Aardwolf* class as an example, the total number of thesaurus terms extracted

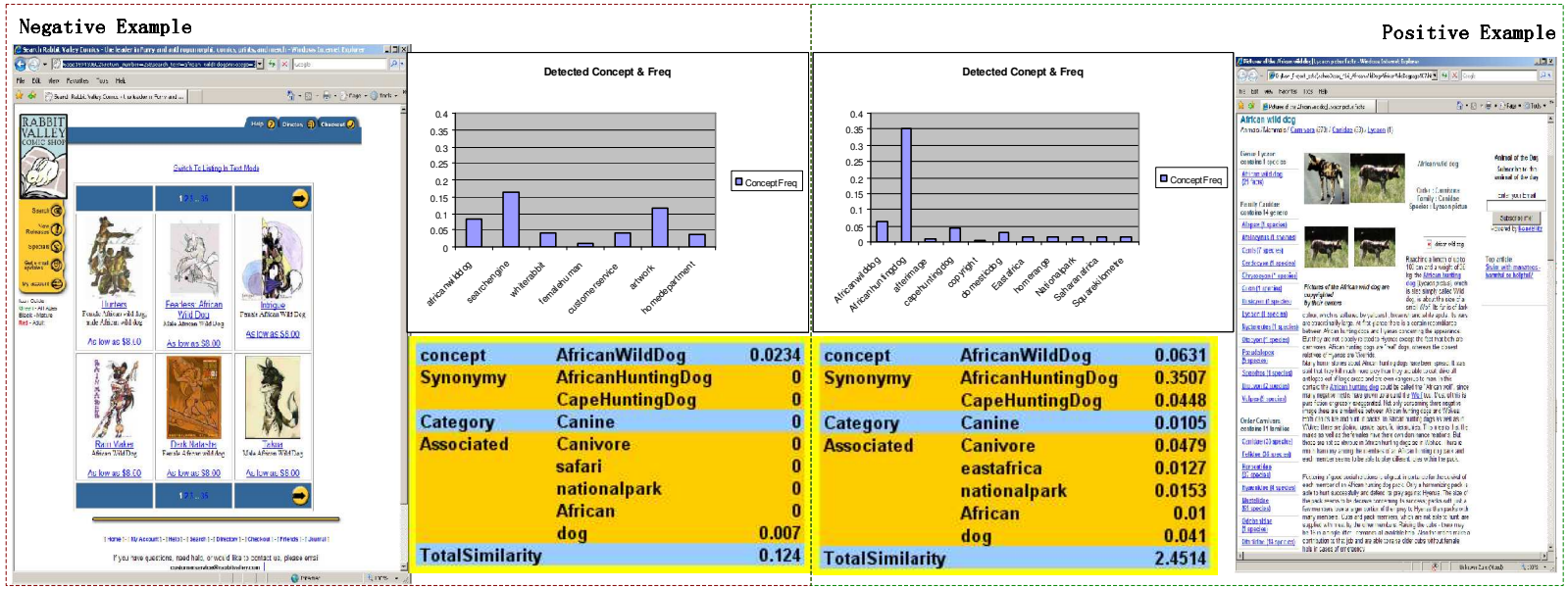


Figure 5.4: An example of thesaurus enhanced similarity calculation from web page text. Concepts from multiple adjacent words are detected. A new concept thesaurus is built, from where the semantic salient similarity is calculated.

Table 5.2: Average Precision of image classification on thesaurus enhanced result

Class	**Aardwolf**	**Cape Fox**	**Bush Dog**	**Arctic Fox**
TfIdf	0.461	0.6646	0.5132	0.754
Thesaurus	0.5384	0.7414	0.8286	0.8463
Thesaurus + visu	0.6126	0.8336	0.9039	0.9225
Class	**Leopard**	**Dhole**	**Red Fox**	**Maned Wolf**
TfIdf	0.6174	0.6992	0.734	0.8503
Thesaurus	0.659	0.7652	0.7794	0.8873
Thesaurus + visu	0.8652	0.8961	0.8592	0.8904
Class	**Black Jackal**	**Lion**	**Dingo**	**Gray Wolf**
TfIdf	0.7657	0.6427	0.4913	0.6776
Thesaurus	0.8335	0.685	0.6339	0.7157
Theasaurus + visu	0.8512	0.8643	0.8605	0.8402
Class	**African Wild Dog**	**Bat-Eared Fox**	**Cheetah**	**Coyote**
TfIdf	0.7636	0.7455	0.5532	0.5447
Thesaurus	0.793	0.7886	0.6317	0.5918
Thesaurus + visu	0.8999	0.8897	0.8379	0.7974
Class	**Ethiopian Wolf**	**Fennec Fox**	**Golden Jackal**	**Gray Fox**
TfIdf	0.7704	0.6482	0.5886	0.533
Thesaurus	0.8602	0.7267	0.7143	0.6759
Thesaurus + visu	0.9031	0.8173	0.9567	0.8824
Class	**Jaguar**	**Kit Fox**	**Red Wolf**	**Snow Leopard**
TfIdf	0.5198	0.528	0.556	0.6959
Thesaurus	0.6841	0.6646	0.6501	0.7592
Thesaurus + visu	0.8098	0.8212	0.8604	0.8821
Class	**Tiger**	**Spotted Hyena**		
TfIdf	0.4017	0.7812		
Thesaurus	0.655	0.8513		
Thesaurus + visu	0.8588	0.9318		

including synonymy, polysemy, parent concept/category and associated, is 76, while the average number per class is around 300. This is decided by the Wikipedia content of this concept. Therefore, the retrieval performance for that category is not as good as the rest. For the thesaurus+visu result, when the object is with distinct color and texture(e.g.: Arctic Fox, Leopard), adding low-level features could contribute more to a better ranking.

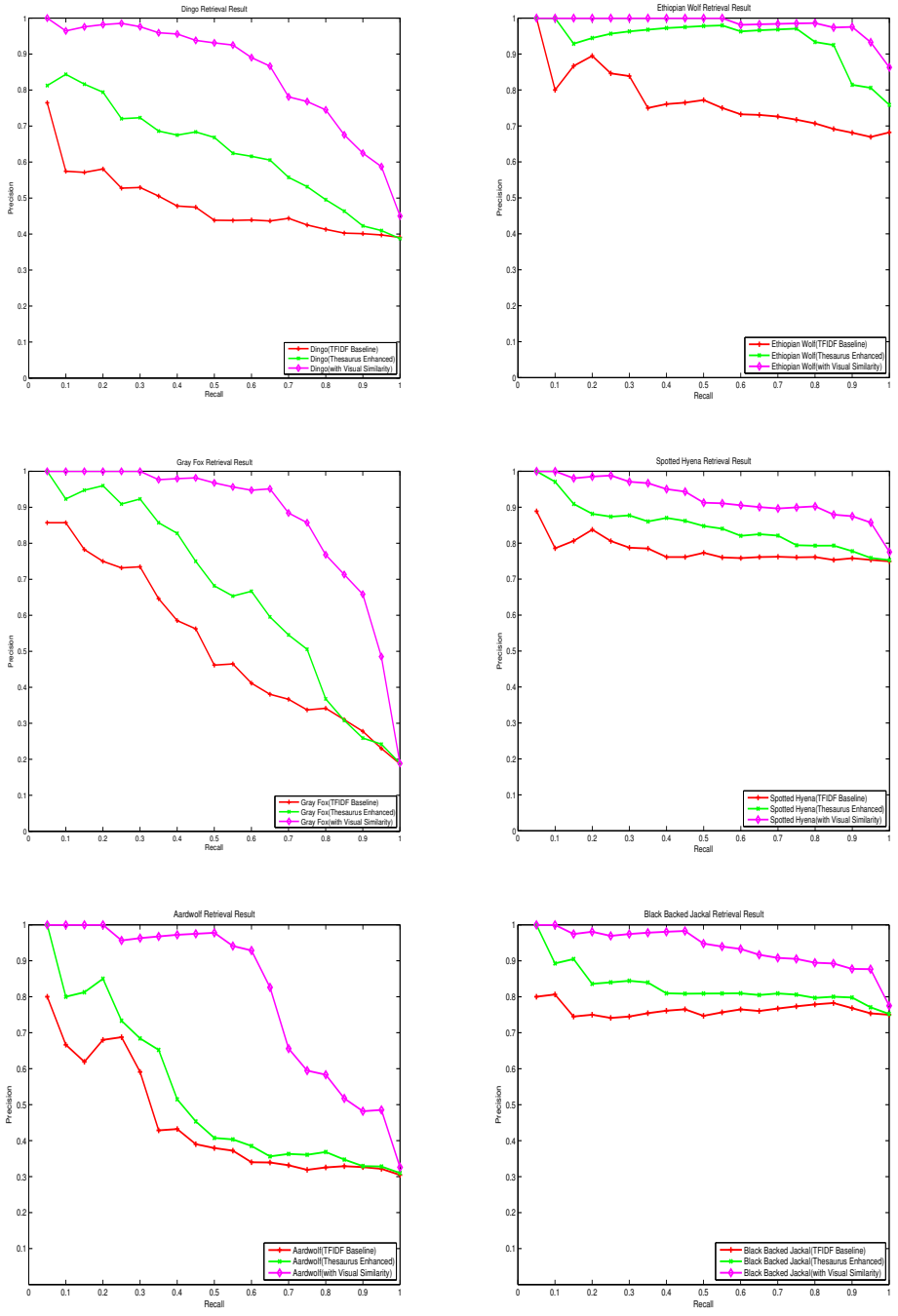


Figure 5.5: A comparison of image retrieval results between of TfIdf, Thesaurus Enhanced and Thesaurus with Visual Feature results

5.4 Summary

In this chapter ways of using an automatically built large-scale *concept thesaurus* for web image retrieval have been proposed. The thesaurus is proposed as an extensive concept ontology to cover large-scale concepts automatically. Different from the previous automatic ontology construction from Wikipedia, the *enwiki* dump file is used as the knowledge source. It is more efficient than downloading the Wikipedia web page for each concept. Instead of analyzing the HTML tags for every web page, the XML tags in the whole dump file is used to extract four main relationships between concepts. These relationships contain synonymy, polysemy, parent concept/category, and associated concept. Compared to the previous constructed ontologies, this approach enable the concept coverage to go beyond the domain-dependent scale and capture rich semantic concepts and relations. An algorithm is also proposed to utilize this thesaurus model in web image retrieval. In the retrieval experiment, the concept detection method first mines the concepts from the text corpus. Base on both text and image features, the concept distance calculation based on the thesaurus model is proved to work well in extended image database. By using the proposed *concept thesaurus* in a web image database which includes around 13,000 images, the scalability of the proposed method has been proved.

Chapter 6

Conclusions and Future Work

In this doctoral dissertation, the efforts on finding an effective ontology based model for multimedia information retrieval have been presented. The main focus is the problem of image retrieval in the dynamic and noisy web environment. Towards the solution of the problem a multi-modality ontology model, which utilizes both text and image features available from web image has been proposed. Meanwhile, ways for more effective ontology construction for larger scales have also been provided. The proposed ontology model, together with the proposed semantic matchmaking process, form the final prototype system for web image retrieval. Related problems of large-scale concept ontology have been investigated. A large-scale concept thesaurus is built and applied in image retrieval with bigger database. In this chapter, a summary of all the results obtained in this dissertation is first given in Section 6.1. Potential research issues are further discussed in Section 6.2.

6.1 Research Summary

In Chapter 3 a multi-modality ontology has been constructed step by step. The original idea of building such an ontology is inspired by the rich yet noisy text and image information provided by web image. The main idea is mining the semantic meaningful concepts from the web image, while filtering out the noisy information. The proposal starts from the plain text vector approach, and moves to construct a text ontology with hierarchical links between concepts. Since ontology are usually designed by expert and constructed to represent domain knowledge, proper knowledge source needs to be found

for the construction. Nowadays information technology development has made it easy for knowledge sharing. It is not difficult to find online categories and encyclopedias to provide knowledge needed for ontology construction. In the experiment, the animal domain knowledge is found from the BBC Science & Nature Animal category. The initial text ontology model is manually built, where the concept and relation definitions are key in using a Lisp style language. The final text ontology consists of *Animal Domain Ontology*, which depicts the taxonomy relations between different animal classes, and *Text Description Ontology*, which provides descriptive information of the particular animal classes, such as habitat, diet, etc.. The *Animal Domain Ontology* mainly contains is-a relationships, while *Text Description Ontology* contains various descriptive relationships. The concepts which are linked to the animal classes have their own hierarchical structure. The experiment has proved that a hierarchical concept structure with encoded domain knowledge could help to improve image retrieval in the noisy web environment.

Real world images are different from the image in most computer vision research databases. Research data usually contains single objects, homogeneous foreground objects and backgrounds, while web image mostly consists of real world images with heterogeneous foreground and background. Therefore it is important to analyze surrounding text information to help web image retrieval. However, on the other hand, the annotations and labels made by web users contain too much noisy. It is not safe to make a conclusion purely based on text information. It is believed that the image features from the image itself can be used to improve the retrieval accuracy. In the experiment, middle-level concepts are defined for image features. According to the image feature concepts, the images are classified into several classes, including *Colorful* or *Grayscale*, *Outdoor* or *Indoor*, *Human Relevant* or *Wildlife*, etc.. The images are also classified based on the foreground color and texture. Each image has a set of image feature concept labels to describe its content. The *Visual Description Ontology* is built by linking these image feature concepts to the animal classes. The final ontology model fuses information from both text and image modalities seamlessly.

The multi-modality ontology model has been proved to be effective. However, there are still open issues about its construction. The initial model is built with manually defined concepts, which could lead to scalability problem when more concepts are involved.

In view of the scalability issue, an automatic construction has been proposed by leveraging the knowledge from online encyclopedia Wikipedia. Every concept in Wikipedia is classified under the main category according to general knowledge. Each concept is hyperlinked to other concepts through various relations. The underlying category structure of Wikipedia provides taxonomy information for the *Animal Domain Ontology*, while the cross-references between concepts provide information for the *Text Description Ontology*. Various relations are extracted by analyzing the HTML tags on each Wikipedia concept page. The ontology model is automatically generated.

In Chapter 4, the semantic matchmaking process which is used for ontology inference is discussed. Reasoners based on description logic are first introduced. Description logic based reasoners can check the consistency of the knowledge base. They are used to process concepts and relations between concepts. They also derive new facts which are entailed in existing ontology. Since an online web image retrieval system is built, RACER version 1.9 which supports HTTP and TCP protocols is used. During the matchmaking process, the user first gives a query concept. Then the reasoner constructs an image concept according to each web image's text and image features. The image concept is matchmaked with the query concepts, and the results are generated according to the matching degrees: *Exact Match*, *Subsume Match* and *Disjoint Match*. Since only limited degrees are given by the reasoners, later an enhanced algorithm is proposed according to the different priorities of semantic relations to further quantize the similarity between concepts. The general idea is that the priority of the concept and relation is related to its occurrence in the test data. Therefore, an assumption is made that a semantic relation which appears more frequently than others contains more relevant information to the image subject. A record of the extracted relation frequencies is kept, and Spearman's ranking correlation is used to measure the final similarity within each of the three semantic matchmaking results. Through this further quantization of the original reasoner result, better tuned retrieval results are obtained. The enhanced matchmaking algorithm works fine with the ontology model. In the next step extended ontology which contains much more concepts are discussed. According to the experiment result, when the ontology contains more concepts, the memory required by the reasoner to process the inference also increases. The description logic based reasoner is a potential bottleneck for

ontology application in large-scale domain. Other inference models which could support larger ontology is proposed.. Considering that ontology, which consists of concepts and relations, is similar to a semantic network, it makes sense to treat ontology as a semantic network and use inference models of the cognitive field for ontology inference. In semantic network, concepts are nodes while the relations are represented by links between nodes. An inference model is proposed based on the SAT. An initial input is generated from the test image and activates specific nodes of the network. The activation spreads to the rest of nodes in the network and each node will be activated according to its relations to its neighboring nodes. A higher activation value means a higher relevancy. The final ranking of the retrieved images is generated according to the activation value. The proposed SAT based inference model produces comparable result to the description logic based model. This model works well with more than 700 concepts and is thus more effective and memory efficient. Following the proposed ontology and matchmaking models, the image retrieval prototype systems are also shown in Chapter 4. Even though the workflow varies between different ontology and mathmaking models, the basic system structure contains an offline ontology construction module and an online semantic matchmaking module. Both the experiment results and the system model prove the feasibility of the ontology-based image retrieval approaches.

Inspired by the success of using the online encyclopedia Wikipedia to build domain-dependent ontology, an attempt is made to go beyond the domain-dependent scale and build a domain independent large-scale *concept thesaurus*. In Chapter 5, such a thesaurus is built based on the Wikipedia dump file. This dump file covers all the Wikipedia concepts and presents the page contents of the concepts in an XML format. Through the analysis of the dump file format, an entry for each concepts is built in the thesaurus according to the four semantic relations: synonymy, polysemy, parent concept/category, and associated concept. All the concept entries are indexed and the final total number of the concepts, excluding those for Wikipedia administration purpose or miscellaneous concepts, reaches a number of 5,836,166. For an input query concept, its synonymy, polysemy, parent and child concepts, together with a set of associated concepts can be retrieved. The thesaurus enhanced similarity calculation to the web image retrieval includes two steps: semantic concept detection from web page text and semantic salient

similarity calculation. Concepts under each semantic relations are extracted from the web page text and the similarity is calculated according to concept frequency information. A final similarity value gives the relevancy of the web page text to the input query. A further enhancement is introduced by using visual similarity from the web images. The decision value from the image feature training process is taken as a measure of the visual similarity. Later the visual similarity is combined with the text similarity. The final retrieved images are ranked according to the combined similarity. Experiment in this chapter contains a bigger database which includes 13,856 web images and their associated web pages. The animal classes are increased to 26 classes. The experiment result has prove that the *concept thesaurus* can detect semantic salient concepts and also capture the semantic relations among various concepts.

6.2 Future Works

This section is about several potential research directions to be addressed as future works. Some possible improvement work together with some new challenges are listed as follows:

- For concept ontology construction, much of the emphasis is put on concept ontology and concept thesaurus construction. The part of visual features, even though which is also studied as part of the whole model, has more or less been used as an extension or add-on to accomplish the framework. State-of-art of image processing techniques is used to derive middle level concepts to represent information from the image modality. This part can be further exploited to improve the retrieval performance. Web images mainly contain real world images with big variations on both foreground and background. Currently in the experiment, the concept thesaurus for a particular animal domain has been discussed. Same image features are extracted from all the images. Actually, the concept thesaurus can also be applied to extensive domains, like plant, nature scene, famous landmarks, etc.. A potential approach is that by studying the image characteristics of different concept domains, an adaptive combination of visual features could be adopted to present images of different concepts. In that case, visual features will be selected according to their discriminative power and significe level to certain concepts. For example,

color features could be more important to classify sea and beach scenes, while texture features are more meaningful features to classify animal images. Moreover, when it comes to the phase of text and visual information fusion, a more adaptive weighing mechanism could be used to put different emphasis on information from the two modalities. When the images of certain concepts tend to present very heterogeneous visual features, text information is more reliable. On the other hand, while dealing with ambiguous concepts, visual features could be more helpful to differentiate the different meanings. Concept jaguar, being both a car brand and an animal name, is such an example. In a word, concept ontology or thesaurus can be exploit in a more adaptive way.

- The concept thesaurus in Chapter 5 can be further developed to include more semantic relations. In the current thesaurus, in each concept entry, the associate concepts are not further categorized according to their real relations to the entry concept. This is a major difference between the concept ontology and the concept thesaurus. In the concept ontology construction, for the association part, the concepts are linked with specific semantic relations, such as "hasDiet", "hasDistribution", etc.. These relations are either predefined or extracted from section titles of Wikipedia article. The similarity calculation based on the explicit relations are more refined and reflect the human cognition of concept. A further study of the Wikipedia dump file to classify the associated concepts under various relations will be helpful to better model the concepts according to general knowledge and preserve more information for effective retrieval. Updating the thesaurus is also necessary to keep up with the changing vocabulary in real life. Novel application of the more complete thesaurus should be proposed and developed accordingly.

- The proposed concept ontology approaches can also be used in other multimedia information retrieval tasks. Currently the main experiment is for the image retrieval tasks. It is also an interesting topic to study how concept ontology works for video retrievals. The popular video sharing web sites such as YouTube provides opportunities of such research. Videos on those web sites can be downloaded, together with the video title, introduction given by authors, comments from other

uses and statistics provided by the web site. The text ontology part can be constructed from the text annotation and statistics of the video in a similar way to the proposed image retrieval approaches, while the visual feature part can be derived through analysis of the key frames of the videos. The related issues include analyzing the available information and structure, building ontology to represent video information, and applying the ontology model to video retrieval tasks. With the increasing volume of shared videos on the Internet, the proposed research can provide solutions to many emergent online video analysis and retrieval issues.

- In this dissertation, an original proposal and application of ontology model for web image retrieval has been discussed. Even though there are many existing ontologies developed by various research parties, those ontologies do not serve for purpose of media information retrieval. Meanwhile, there are other researchers working towards a similar goal of concept-based media information retrieval. In the future, there might be needs to merge the available resources and ontologies into one single framework so that joint force can contribute to media retrieval tasks. A future direction could be developing a formal ontology standard, which supports descriptions and inference of both domain knowledge and image/video feature information. Such standard should define compact and effective domain concepts and efficient inference scheme in a systematic way. This could be a challenging job given the vast range of required concepts, different multimedia applications and broad number of multimedia features. Upon successful completion of the standard, ontology-based multimedia information retrieval researches can follow an open criteria and compare with each other in a unified platform.

Publication

(i) Huan Wang, Xing Jiang, Liang-Tien Chia and Ah-Hwee Tan, "Building Concept Ontology Automatically, Experimenting with Web Image Retrieval", Accepted by Special Issue on Semantic Information Technologies of Informatica, 2010(to appear).

(ii) Shenghua Gao, Xiangang Cheng, Huan Wang, and Liang-Tien Chia, "Concept Model-Based Unsupervised Web Image Re-Ranking", Accepted by IEEE International Conference on Image Processing, 2009(to appear).

(iii) Huan Wang, Xing Jiang, Liang-Tien Chia and Ah-Hwee Tan,"Wikipedia2Onto — Adding Wikipedia Semantics to Web Image Retrieval", In Proceedings of the WebSci'09: Society On-Line, 2009.

(iv) Huan Wang, Xing Jiang, Liang-Tien Chia and Ah-Hwee Tan,"Ontology Enhanced Web Image Retrieval: Aided by Wikipedia & Spreading Activation Theory", In Proceeding of the 1^{st} ACM International Conference on Multimedia information Retrieval, pp. 195-201, 2008.

(v) Huan Wang, Song Liu and Liang-Tien Chia, "Image Retrieval with a Multi-Modality Ontology Multimedia System", Multimedia Syst., vol. 13, no.5-6, pp. 379-390, 2008.

(vi) Huan Wang, Song Liu and Liang-Tien Chia, "Image Retrieval ++ — Web Image Retrieval with An Enhanced Multi-Modality Ontology", Multimedia Tools and Applications, vol. 39, no. 2, pp. 189-215, 2008.

(vii) Huan Wang, Liang-Tien Chia and Song Liu, "Semantic Retrieval with Enhanced Matchmaking and Multi-Modality Ontology", In Proceeding of the 2007 IEEE International Conference on Multimedia and Expo, pp. 516 - 519, 2007.

(viii) Huan Wang, Song Liu and Liang-Tien Chia, "Does Ontology Help in Image Retrieval?: A Comparison between Keyword, Text ontology and Multi-Modality Ontology Approaches", In Proceedings of the 14^{th} Annual ACM International Conference on Multimedia, pp. 109 - 112, 2006.

(ix) Huan Wang, Song Liu and Liang-Tien Chia, "Does Ontology Help in Image Retrieval?", published by Asia-Pacific Workshop on Visual Information Processing 2006(Best Paper Award).

References

[1] T. R. Gruber, "A translation approach to portable ontologies," *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1379, 2000.

[3] Y. A. Aslandogan, C. Thier, C. T. Yu, J. Zou, and N. Rishe, "Using semantic contents and wordnet in image retrieval," in *Proceedings of the Twentyth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 286–295, 1997.

[4] I. Niles and A. Pease, "Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology," in *Proceedings of the International Conference on Information and Knowledge Engineering. Volume 2*, pp. 412–416, 2003.

[5] V. Snásel, P. Moravec, and J. Pokorný, "Wordnet ontology based model for web retrieval," in *Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration*, pp. 220–225, 2005.

[6] Z. Zhou, Y. Wang, and J. Gu, "A novel method of extracting domain ontology based on wordnet," in *Proceedings of the International Conference on Computer Science and Software Engineering*, pp. 376–381, 2008.

[7] H. Tamura and N. Yokoya, "Image database systems: A survey.," *Pattern Recognition*, vol. 17, no. 1, pp. 29–43, 1984.

[8] S.-K. Chang and A. Hsu, "Image information systems: Where do we go from here?," *IEEE Transactions on Knowledge and Data Engineering*, vol. 4, no. 5, pp. 431–442, 1992.

[9] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, USA: McGraw-Hill, Inc., 1986.

[10] J. Laaksonen, M. Koskela, and E. Oja, "Picsom-self-organizing image retrieval with mpeg-7 content descriptors," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 841–853, 2002.

[11] M. Swain and D. Ballard, "Color indexing," *International journal of computer vision*, vol. 7, no. 1, pp. 11–32, 1991.

[12] J. Smith and S. Chang, "Automated binary texture feature sets for image retrieval," in *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 2239–2242, 1996.

[13] M. Ioka, *A method of defining the similarity of images on the basis of color information*. IBM Research, Tokyo Research Laboratory, 1989.

[14] J. Smith and S. Chang, "Transform features for texture classification and discrimination inlarge image databases," in *Proceedings of the 1994 IEEE International Conference on Image Processing*, vol. 3, pp. 407–411.

[15] C. Zahn, R. Roskies, *et al.*, "Fourier descriptors for plane closed curves," *IEEE Transactions on computers*, vol. 21, no. 3, pp. 269–281, 1972.

[16] X.-J. Wang, W.-Y. Ma, L. Zhang, and X. Li, "Multi-graph enabled active learning for multimodal web image retrieval," in *Proceedings of the Seventh ACM SIGMM International workshop on Multimedia information retrieval*, pp. 65–72, 2005.

[17] B. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Transactions on circuits and systems for video technology*, vol. 11, no. 6, pp. 703–715, 2001.

[18] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Image indexing using color correlograms," in *Proceedings of the 1997 Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 762–768, 1997.

[19] M. Stricker and M. Orengo, "Similarity of color images," *Storage and retrieval for image and video databases III*, pp. 381–392, 1995.

[20] J. Smith and S. Chang, "Single color extraction and image query," in *Proceedings of the 1995 International Conference on Image Processing*, vol. 3, pp. 528–531, 1995.

[21] F. Liu and R. Picard, "Periodicity, directionality, and randomness: Wold features forimage modeling and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 722–733, 1996.

[22] R. Haralick, "Statistical and structural approaches to texture," in *Proceedings of the IEEE*, vol. 67, pp. 786–804, 1979.

[23] A. Jain and F. Farrokhnia, "Unsupervised texture segmentation using gabor filters," *Pattern recognition*, vol. 24, no. 12, pp. 1167–1186, 1991.

[24] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man and cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.

[25] A. Kundu and J. Chen, "Texture classification using qmf bank-based subband decomposition," *CVGIP: Graphical models and image processing*, vol. 54, no. 5, pp. 369–384, 1992.

[26] M. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[27] A. Khotanzad and Y. H. Hong, "Invariant image recognition by zernike moments," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 489–497, 1990.

[28] C.-C. Chen, "Improved moment invariants for shape discrimination," *Pattern Recognition*, vol. 26, no. 5, pp. 683–686, 1993.

[29] B. Mehtre, M. Kankanhalli, and W. Lee, "Shape measures for content based image retrieval: a comparison," *Information Processing and Management*, vol. 33, no. 3, pp. 319–337, 1997.

[30] F. Mokhtarian and M. Bober, *Curvature scale space representation: theory, applications, and MPEG-7 standardization*. Kluwer Academic Publishers, 2003.

[31] G. Dorkó and C. Schmid, "Selection of scale-invariant parts for object class recognition," in *Proceedings of Ninth IEEE International Conference on Computer Vision*, pp. 634–640, 2003.

[32] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[33] S. Nowozin, K. Tsuda, T. Uno, T. Kudo, and G. H. Bakir, "Weighted substructure mining for image analysis," in *Proceedings of the Thirteenth IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[34] K. Saenko and T. Darrell, "Unsupervised learning of visual sense models for polysemous words," in *Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems*, pp. 1393–1400, 2008.

[35] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE Transaction on Multimedia*, vol. 12, no. 1, pp. 42–53, 2010.

[36] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[37] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," in *Proceedings of the Tenth IEEE Computer Society Conference on Computer Vision and Pattern Recognition Part II*, pp. 506–513, 2004.

[38] M. Grabner, H. Grabner, and H. Bischof, "Fast approximated sift," in *Proceedings of the Seventh Asian Conference on Computer Vision Part I*, pp. 918–927, 2006.

[39] A. E. Abdel-Hakim and A. A. Farag, "Csift: A sift descriptor with color invariant characteristics," in *Proceedings of the Twelfth CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition Part II*, pp. 1978–1983, 2006.

[40] S. Mehrotra, Y. Rui, M. Ortega, and T. S. Huang, "Supporting content-based queries over images in mars," in *Proceedings of the 1997 International Conference on Multimedia Computing and Systems*, pp. 632–633, 1997.

[41] T. Gevers and A. Smeulders, "The pictoseek www image search system," in *Proceedings of the 1999 IEEE International Conference on Multimedia Computing and Systems*, vol. 1, pp. 264–269, 1999.

[42] G. Wei, D. Li, and I. Sethi, "Web-wise: Compressed image retrieval over the web," pp. 33–46, 1998.

[43] E. Chang, J. Wang, C. Li, and G. Wiederhold, "Rime: A replicated image detector for the world-wide web," in *Proceedings of SPIE Symposium of Voice, Video, and Data Communications*, pp. 58–67, 1998.

[44] X. Zhou and T. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia systems*, vol. 8, no. 6, pp. 536–544, 2003.

[45] T. Quack, U. Mönich, L. Thiele, and B. S. Manjunath, "Cortina: a system for large-scale, content-based web image retrieval," in *Proceedings of the Twelfth annual ACM international conference on Multimedia*, pp. 508–511, 2004.

[46] M. Rahman, P. Bhattacharya, and B. Desai, "A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback," *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 1, pp. 58–69, 2007.

[47] D. Morrison, S. Marchand-Maillet, and E. Bruno, "Semantic clustering of images using patterns of relevance feedback," in *Proceedings of the International Workshop on Content-Based Multimedia Indexing*, pp. 323–329, 2008.

[48] Z. Li, S. Shi, and L. Zhang, "Improving relevance judgment of web search results with image excerpts," in *Proceeding of the Seventeenth International Conference on World Wide Web*, pp. 21–30, 2008.

[49] H. Nezamabadi-pour and E. Kabir, "Concept learning by fuzzy k-nn classification and relevance feedback for efficient image retrieval," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5948–5954, 2009.

[50] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image database," *International Journal of Computer Vision*, no. 3, pp. 233–254, 1996.

[51] J. R. Smith and S.-F. Chang, "Visualseek: A fully automated content-based image query system," in *Proceedings of the Fourth ACM international conference on Multimedia*, pp. 87–98, 1996.

[52] Y. A. Aslandogan, C. Thier, C. T. Yu, J. Zou, and N. Rishe, "Using semantic contents and wordnet in image retrieval," in *Proceedings of Twentyth annual international ACM SIGIR conference on Research and development in information*, pp. 286–295, 1997.

[53] Z. Wang, J. Li, and G. Wiederhold, "Simplicity: Semantics-sensitive integrated matching for picture libraries," in *Proceedings of the Fourth International Conference on Advances in Visual Information Systems*, pp. 360–371, 2000.

[54] S. Hammiche, S. Benbernou, M.-S. Hacid, and A. Vakali, "Semantic retrieval of multimedia data," in *Proceedings of the Second ACM International Workshop on Multimedia Databases*, pp. 36–44, 2004.

[55] M. Ortega-Binderberger, S. Mehrotra, K. Chakrabarti, and K. Porkaew, "Webmars: A multimedia search engine for full document retrieval and cross media browsing," in *Proceedings of the 2000 Workshop Multimedia Information System*, 2000.

[56] A. B. Benitez and S.-F. Chang, "Semantic knowledge construction from annotated image collections," in *Proceedings of IEEE International Conference on Multimedia*, pp. 26–29, 2002.

[57] E. Cheng, F. Jing, L. Zhang, and H. Jin, "Scalable relevance feedback using click-through data for web image retrieval," in *Proceedings of the Fourteenth annual ACM international conference on Multimedia*, pp. 173–176, 2006.

[58] G. Smith and H. Ashman, "Evaluating implicit judgments from image search interactions," in *Proceedings of the Web Science Conference: Society On-Line*, 2009.

[59] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, 2007.

[60] R. Zhao and W. Grosky, "Narrowing the semantic gap-improved text-based web document retrieval using visual features," *IEEE Transactions on Multimedia*, vol. 4, no. 2, pp. 189–200, 2002.

[61] I. Sethi, I. Coman, and D. Stan, "Mining association rules between low-level image features and high-level concepts," in *Proceedings of SPIE Data Mining and Knowledge Discovery*, pp. 279–290, 2001.

[62] T. L. Berg and D. A. Forsyth, "Animals on the web," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1463–1470, 2006.

[63] C. Town and D. Sinclair, "Content based image retrieval using semantic visual categories,"

[64] A. Vailaya, M. A. T. Figueiredo, A. K. Jain, and H. Zhang, "Image classification for content-based indexing," *IEEE Transactions on Image Processing*, vol. 10, no. 1, pp. 117–130, 2001.

[65] J. Luo and A. E. Savakis, "Indoor vs outdoor classification of consumer photographs using low-level and semantic features," in *Proceedings of the 2001 International Conference on Image Processing(2)*, pp. 745–748, 2001.

[66] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[67] R. Shi, H. Feng, T.-S. Chua, and C.-H. Lee, "An adaptive image content representation and segmentation approach to automatic image annotation," in *Proceedings of The Third International Conference on Image and Video Retrieval*, pp. 545–554, 2004.

[68] H. Feng and T.-S. Chua, "A bootstrapping approach to annotating large image collection," in *Proceedings of the Fifth ACM International Workshop on Multimedia Information Retrieval*, pp. 55–62, 2003.

[69] H. Feng, R. Shi, and T.-S. Chua, "A bootstrapping framework for annotating and retrieving www images," in *Proceedings of the Twelfth ACM International Conference on Multimedia*, pp. 960–967, 2004.

[70] J. Pearl and G. Shafer, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[71] Y. Gao and J. Fan, "Incorporating concept ontology to enable probabilistic concept reasoning for multi-level image annotation," in *Proceedings of the Eighth ACM International Workshop on Multimedia Information Retrieval*, pp. 79–88, 2006.

[72] L. Fan and B. Li, "A hybrid model of image retrieval based on ontology technology and probabilistic ranking," pp. 477–80, 2006.

[73] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.

[74] S. MacArthur, C. Brodley, and C. Shyu, "Relevance feedback decision trees in content-based image retrieval," in *Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries*, pp. 68–72, 2000.

[75] W.-C. Low and T.-S. Chua, "Color-based relevance feedback for image retrieval," in *Proceedings of the International Workshop on Multimedia Database Management Systems*, pp. 116–123, 1998.

[76] S.-F. Chang, W. Chen, and H. Sundaram, "Semantic visual templates: Linking visual features to semantics," in *Proceedings of the 1998 International Conference on Image Processing*, pp. 531–535, 1998.

[77] Y. Zhuang, X. Liu, and Y. Pan, "Apply semantic template to support content-based image retrieval," in *Proceedings of the SPIE, Storage and Retrieval for Media Databases*, pp. 442–49, 1999.

[78] J. Smith and Y. Heights, "Decoding image semantics using composite region templates," in *Proceedings of 1998 IEEE Workshop on Content-Based Access of Image and Video Libraries*, pp. 9–13, 1998.

[79] A. Gomez-Perez, M. Fernández-López, and O. Corcho, *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Verlag, 2004.

[80] C. Elkan and R. Greiner, "Building large knowledge-based systems: Representation and inference in the cyc project," *Artificial Intelligence*, vol. 61, no. 1, pp. 41–52, 1993.

[81] A. Bernaras, I. Laresgoiti, and J. M. Corera, "Building and reusing ontologies for electrical network applications," in *Proceedings of the Twelfth European Conference on Artificial Intelligence*, pp. 298–302, 1996.

[82] S. Staab, R. Studer, H.-P. Schnurr, and Y. Sure, "Knowledge processes and ontologies," *IEEE Intelligent Systems*, vol. 16, no. 1, pp. 26–34, 2001.

[83] N. F. Noy, R. W. Fergerson, and M. A. Musen, "The knowledge model of protégé-2000: Combining interoperability and flexibility," in *Proceedings of the Twelfth International Conference on Knowledge Acquisition, Modeling and Management*, pp. 17–32, 2000.

[84] Y. Sure, M. Erdmann, J. Angele, S. Staab, R. Studer, and D. Wenke, "Ontoedit: Collaborative ontology development for the semantic web," in *Proceedings of the First International Semantic Web Conference*, pp. 221–235, 2002.

[85] M. Richardson and P. Domingos, "Building large knowledge bases by mass collaboration," in *Proceedings of the 2nd International Conference on Knowledge Capture*, pp. 129–137, 2003.

[86] J. Hunter, "Adding multimedia to the semantic web: Building an mpeg-7 ontology," in *Proceedings of the 2001 International Semantic Web Working Symposium*, pp. 261–283, 2001.

[87] S. Liu, L.-T. Chia, and S. Chan, "Ontology for nature-scene image retrieval," in *Proceedings of the International Conference On the Move to Meaningful Internet Systems*, pp. 1050–1061, 2004.

[88] A. Gordon, "Browsing image collections with representations of common-sense activities," *Journal of the American Society for Information Science and Technology*, vol. 52, no. 11, pp. 925–929, 2001.

[89] E. Hyvönen, S. Saarela, A. Styrman, and K. Viljanen, "Ontology-based image retrieval," in *Proceedings of the 2002 Conference on Towards the semantic web and web services*, pp. 15–27, 2003.

[90] S. Radhouani, J. H. Lim, J.-P. Chevallet, and G. Falquet, "Combining textual and visual ontologies to solve medical multimodal queries," in *Proceedings of the 2006 International Conference On Multimedia & Expo*, pp. 1853–1856, 2006.

[91] B. Hu, S. Dasmahapatra, P. H. Lewis, and N. Shadbolt, "Ontology-based medical image annotation with description logics," in *Proceedings of the Fifteenth IEEE International Conference on Tools with Artificial Intelligence*, pp. 77–77, 2003.

[92] A. Popescu, P.-A. Moëllic, and C. Millet, "Semretriev: an ontology driven image retrieval system," in *Proceedings of the Sixth ACM International Conference on Image and Video Retrieval*, pp. 113–116, 2007.

[93] W. Grosky, D. Sreenath, and F. Fotouhi, "Emergent semantics and the multimedia semantic web," *ACM SIGMOD Record*, vol. 31, no. 4, pp. 54–58, 2002.

[94] X. Wang, W. Ma, Q. He, and X. Li, "Grouping web image search result," pp. 436–439, 2004.

[95] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, and W. Ma, "Igroup: web image search results clustering," pp. 377–384, 2006.

[96] J. Smith and S. Chang, "Searching for images and videos on the world-wide web," *IEEE MultiMedia*, 1997.

[97] X.-Y. Wei and C.-W. Ngo, "Ontology-enriched semantic space for video search," in *Proceedings of the Fifteenth International Conference on Multimedia*, pp. 981–990, 2007.

[98] Z. Wu and M. S. Palmer, "Verb semantics and lexical selection," in *Proceedings of the Thirty-Second Annual Meeting on Association for Computational Linguistics*, pp. 133–138, 1994.

[99] R. M. Haralick, "Statistical and structural approaches to texture," *Proceedings of IEEE*, vol. 67, no. 5, pp. 786–804, 1979.

[100] S. ichi Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks*, vol. 12, no. 6, pp. 783–789, 1999.

[101] S. Tong and E. Y. Chang, "Support vector machine active learning for image retrieval," in *ACM Multimedia*, pp. 107–118, 2001.

[102] S. Weng, H. Tsai, S. Liu, and C. Hsu, "Ontology construction for information classification," *Expert Systems with Applications*, vol. 31, no. 1, pp. 1–12, 2006.

[103] L. Khan and F. Luo, "Ontology construction for information selection," in *Proceedings of the Fourteenth IEEE International Conference on Tools with Artificial Intelligence*, pp. 122–127, 2002.

[104] M. Marszalek and C. Schmid, "Semantic hierarchies for visual object recognition," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–7, 2007.

[105] T. Zesch and I. Gurevych, "Analysis of the Wikipedia Category Graph for NLP Applications," in *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*, pp. 1–8, 2007.

[106] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," in *Proceedings of the 2006 International Conference on Computer Vision and Pattern Recognition*, pp. 13–13, 2006.

[107] R. Srikant and R. Agrawal, "Mining generalized association rules," in *Proceedings of Twenty-First International Conference on Very Large Data Bases*, pp. 407–419, 1995.

[108] V. Haarslev and R. Moller, "Racer system description," in *Proceedings of the International Joint Conference on Automated Reasoning*, pp. 701–705, 2001.

[109] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Wordnet: An online lexical database.," *International Journal of Lexicography*, pp. 235–244, 1990.

[110] I. Horrocks, "Fact and ifact," in *Proceedings of the International Workshop on Description Logics*, pp. 133–135, 1999.

[111] F. Baader and D. Calvanese, *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2003.

[112] F. Badder and U. Sattler, "An overview of tableau algorithm for description logics," *Studia Logica*, vol. 69, no. 1, pp. 5–40, 2001.

[113] M. Paolucci, T. Kawmura, T. Payne, and K. Sycara, "Semantic matching of web services capabilities," in *Proceedings of the First International Semantic Web Conference*, pp. 333–347, 2002.

[114] L. Li and I. Horrocks, "A software framework for matchmaking based on semantic web technology," *International Journal of Electronic Commerce*, vol. 8, no. 4, pp. 39–60, 2004.

[115] C. Zhou, L.-T. Chia, and B.-S. Lee, "Web services discovery with daml-qos ontology," *International Journal of Web Services Research*, vol. 2, pp. 44–76, 2005.

[116] E. Bozask, M. Ehrig, and etl., "Kaon - towards a large scale semantic web," in *Proceedings of the Third international conference on E-Commerce and Web Technologies*, pp. 304–313, 2002.

[117] Racer System GmbH & Co. KG, *RacerPro User's Guide Version 1.9*, 2005.

[118] K. Yanai and K. Barnard, "Probabilistic web image gathering," in *Proceedings of the Seventh ACM SIGMM international workshop on Multimedia information retrieval*, pp. 57–64, 2005.

[119] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features.," in *Proceedings of the Tenth IEEE International Conference on Computer Vision*, pp. 1458–1465, 2005.

[120] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories.," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2169–2178, 2006.

[121] K. Grauman and T. Darrell, "Unsupervised learning of categories from sets of partially matching image features.," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 19–25, 2006.

[122] J. F. Sowa, *Principles of Semantic Networks: Explorations of Representation of Knowledge*. Morgan Kaufmann Publishers, 1991.

[123] X. Jiang and A.-H. Tan, "Ontosearch: A full-text search engine for the semantic web," in *Proceedings of The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, pp. 1325–1330, 2006.

[124] R. J. Anderson, "A spreading activation theory of memory," *Journal of Verbal Learning and Verbal Behavior*, vol. 22, pp. 261–295, 1983.

[125] J. Contreras, V. R. Benjamins, M. Blázquez, S. Losada, R. Salla, J. Sevilla, D. Navarro, J. Casillas, A. Mompó, D. Patón, Ó. Corcho, P. Tena, and I. Martos, "A semantic portal for the international affairs sector.," in *Proceedings of the Fourteenth International Conference on Engineering Knowledge in the Age of the Semantic Web*, pp. 203–215, 2004.

[126] P. Wang, J. Hu, H.-J. Zeng, L. C. 0002, and Z. Chen, "Improving text classification by using encyclopedia knowledge," in *Proceedings of the Seventh IEEE International Conference on Data Mining*, pp. 332–341, 2007.

[127] J. Hu, L. Fang, Y. Cao, H.-J. Zeng, H. Li, Q. Yang, and Z. Chen, "Enhancing text clustering by leveraging wikipedia semantics," in *Proceedings of the Thirty-first Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 179–186, 2008.

[128] T. Gevers and A. W. M. Smeulders, "Color-based object recognition," *Pattern Recognition*, vol. 32, no. 3, pp. 453–464, 1999.

[129] J. van de Weijer and C. Schmid, "Coloring local feature extraction," in *Proceedings of the Ninth European Conference on Computer Vision*, pp. 334–348, 2006.