UNIVERSITY OF KASDI MERBAH OUARGLA

Faculty of New Technologies of Information and Communication

Department of Computer Science and Information Technologies

Thesis

Academic Master

Domain: Science and technology
Sector: Electronics
Technology Specialty: Electronics of embedded systems

Image clustering based on semantic similarity

Presented by
ZINET Ishak, BOUGUERRA Badis

Supervised by:
Dr YOUCEFA Abdelemadjid

Dr BELKEBIR Djalila	President	UKM Ouargla
Dr YOUCEFA Abdelemadjid	Supervised	UKM Ouargla
Dr NASRI Nadjib	Examiner	UKM Ouargla

The academic year 2021/2022

Dedicate

I'd like to dedicate my work:

To my beloved parents who have always been by my side, giving me support, love and cherish.

To my dear brothers and sisters for helping me.

All thanks and appreciation to all those who stood with me, I reached what I dreamed of for a long time, and this would not have happened without the success of Allah.

Acknowledgments

First and foremost, praise and thanks to ALLAH the sacred and the mighty for giving me the ability and helping me in accomplishing this thesis.

My thanks go to all the people who participated in my master's thesis. I would like to thank my supervisor Dr YOUCEFA Abdelemadjid, for his advice, support, and guidance in conducting my research. I would like to thank the jury members Dr BELKEBIR Djalila and Dr NASRI Nadjib for agreeing to judge my work.

Our gratitude to all the professors in the Department of Electronics and Telecommunications, especially those who taught me during my studies. I am also very grateful to all and all of my colleagues.

Table of Contents

Abstract

Image clustering is an interesting field in machine learning and computer vision, in which images are classified into a set of similar groups. Recently, with the explosive growth of the data in the smartphone and the web (Facebook, Instagram…), image clustering has even been a critical field to help the user quickly access the visual information he is looking for. Existing methods of image clustering only used either low-level visual feature, which constitutes a major obstacle to obtaining an accurate set of similar groups. To tackle this problem, we propose a novel algorithm that can cluster images based on the semantic similarity between surrounding texts (concept) of each image. In particular, we group images depending on the semantic similarity of their concepts instead of visual similarity. Conclusively, images are automatically clustered based on the label features. The performance of the cluster was compared based on accuracy. The highest accuracy was obtained by applying the method of Lin with 88.89%.

Keywords: Image clustering, Semantic similarity, Concepts, Ontology.

Résumé

Le regroupement d'images est un domaine intéressant de l'apprentissage automatique et de la vision par ordinateur, dans lequel les images sont classées en un ensemble de groupes similaires. Récemment, avec la croissance explosive des données dans le smartphone et le web (Facebook, Instagram…), le clustering d'images a même été un domaine critique pour aider l'utilisateur à accéder rapidement à l'information visuelle qu'il recherche. Les méthodes existantes de regroupement d'images n'utilisaient que l'une ou l'autre caractéristique visuelle de bas niveau, ce qui constitue un obstacle majeur à l'obtention d'un ensemble précis de groupes similaires. Pour résoudre ce problème, nous proposons un nouvel algorithme qui peut regrouper des images en fonction de la similarité sémantique entre les textes environnants (concept) de chaque image. En particulier, nous regroupons les images en fonction de la similarité sémantique de leurs concepts au lieu de la similarité visuelle. En conclusion, les images sont automatiquement regroupées en fonction des caractéristiques de l'étiquette. Les performances du cluster ont été comparées sur la base de la précision. La précision la plus élevée a été obtenue en appliquant la méthode de Lin avec 88,89 %.

Mots clés : Regroupement d'images, Sémantique similarité, Concepts, Ontologie.

ملخص

يعد تجميع الصور مجالًا مثيرًا للاهتمام في التعلم الآلي ورؤية الكمبيوتر، حيث يتم تصنيف الصور في مجموعة من المجموعات المتشابهة. في الآونة الأخيرة، كان النمو الهائل للبيانات في الهاتف الذكي والويب (فيس بوك وإنستغرام...)، وتجميع الصور مجالًا مهمًا لمساعدة المستخدم على الوصول بسرعة إلى المعلومات المرئية التي يبحث عنها. الأساليب الحالية في تجميع الصور تستخدم فقط ميزة بصرية منخفضة المستوى، مما يشكل عقبة رئيسية أمام الحصول على مجموعة دقيقة من المجموعات المتشابهة. لمعالجة هذه المشكلة، نقترح خوارزمية جديدة يمكنها تجميع الصور بناءً على التشابه الدلالي بين النصوص المحيطة (المفهوم) لكل صورة. على وجه الخصوص، نقوم بتجميع الصور اعتمادًا على التشابه الدلالي لمفاهيمها بدلاً من التشابه البصري. بشكل قاطع، يتم تجميع الصور تلقائيًا بناءً على ميزات التسمية. تم مقارنة أداء الكتلة على أساس الدقة. تم الحصول على أعلى دقة بتطبيق طريقة لين بنسبة 88.89٪.

الكلمات المفتاحية: تجميع الصور والتشابه الدلالي والمفاهيم وعلم الوجود.

List of Figures

List of Tables

Chapter I:

General introduction

I.1 Introduction

In the last few years there has been a growing interest in computer vision. The focus has been on visual features due to the exponential growth of data. Among the most important techniques of data analysis are Clustering and Classification, which help a user to quickly and accurately access it in terms of searching and browsing. In addition, Images clustering is a necessary process this is due to the huge number of images in personal data or in the web. The main objects of cluster images are to regroup related data according to the similarity between images and to preprocess the image data.

In the literature, several theories have been proposed to cluster images based on visual features. Clustering can be done using different techniques like Content-Based Image Retrieval [1, 2], Convolutional Neural networks [3], K-means clustering [4], Mean Shift clustering [5], and DB Scan clustering [6]. However, these above algorithms can consider images to have the same similarity if they share some visual features, for example, an image containing “sheepdog” and another image containing “mop” is in the same category (similar). This depends on the similarity of the visual features shape and also the color.

To overcome this problem, it is necessary to adopt a new algorithm that clusters images by calculating the semantic similarity of words (concept) between images and ignoring low-level visual features.

I.2 Problematic:

The problem can be formulated as follows: Given a set of images, it is required to divide them into multiple groups, such that images in the same group are more similar to each other than images in other groups.

There are many applications of image clustering including image organization and browsing, corpus summarization, and image classification. Among them, which cluster images on the basis of visual features, which causes difficulty in obtaining accurate results from image clustering, the most common problem in clustering is that it is limited to specific features between images such as color or shape, and this leads to the occurrence of groups with different images.

Figure 1. Visually similar images: chihuahuas and blueberry muffins [29].

An image containing a " Chihuahuas " and another picture containing a " blueberry muffins " are considered to have the same meaning. This similarity depends on the shape and color in the clustering.

I.3 Proposed Solution:

The development or improvement of the processes of techniques that help to collect images with high accuracy will help to provide better results, and facilitate the processes of clustering or retrieval or improve the process of clustering applications.

Among the solutions that help to compile better and give more accurate results is the clustering based on the concepts that express in each image, the visual characteristics of the images will be avoided, no matter what, the concept remains the one that expresses the image.

To cluster similar images, the semantic similarity of the images must be calculated. For this, we use the depth of information content between the concepts of images, and the clustering of images that have a great similarity between the concepts, so that similar images in each group differ from the rest of the groups.

I.4 Contributions:

Clustering images have an important role in the presence of the huge amount of data, providing the user with easy access to various data, and among these contributions:
The contributions of this research work are:

- The emergence of more appropriate, accurate, and intense results in searches.
- Linking or clustering images or data in search and organization operations.
- Easily organize big data.
- Users access various data quickly and accurately.

I.5 Thesis structure:

The thesis follow-up is organized as follows. Chapter 1, provides a general introduction to image clustering, and the problems it faces. In chapter 2, we present current techniques that use the visual features of images in clustering and the problem of inaccurate results. In Chapter 3, an overview of Semantic similarity, its concept, methods used to model it, and the difference between them. Then, we will explain the technique used to extract features and measure similarity, and present an algorithm to calculate semantic similarity. Finally, we will draw some conclusions and future works.

Chapter II:

Related work

II.1 Introduction:

There is a huge interest in databases, and in recent times it has become the deployment of large image databases on various applications. Effective access to desired images from large and diverse image databases is now a necessity. Various satellite images, medical and more attract users in various fields, for example, medicine, architecture, geography, and publishing. Although many techniques are used to cluster images on the visual features, they lack accuracy. They focus on a part of the images from their clustering, such as color, shape, image layers, or density.

II.2 Image clustering:

Image clustering is an essential tool for data analysis in machine learning and computer vision. Several applications such as content-based image annotations [7, 8] and image retrieval [9] can be viewed as different instances of image clustering. Clustering images is the process of clustering into groups so that images within the same groups are similar to each other, while those in different groups are different.

Clusters are a difficult concept, as there is evidence that clusters play an important role in organizing information because there are many different clustering algorithms. Different cluster models are used, and different algorithms can be given for each of these cluster models. Clusters found by one clustering algorithm will certainly be different from clusters found by a different algorithm. Clustering can be done using different techniques like:

1. Clustering based on image content.
2. Convolutional Neural networks (CNN).
3. K-means clustering.

II.2.1 Clustering based on image content

Content-based image retrieval (CBIR) is the retrieval or cluster of images based on visual features such as color, size, and shape. Reasons for its development are that in many large image databases, the process extracts the similarities between the images and puts them together based on the shape and content of the images. Content-based image retrieval (CBIR) uses multiple visual features to characterize image content [10].

Figure 2. Sheepdog Image Clustering

We have the above clustering Figure 2 which contains pictures of the "Sheepdog" `and the "mop", the focus in the compilation is based on the shape and color and I consider them to have the same meaning. Clustering based on image content is inaccurate, it is limited to selecting only part of the image or the thing that characterizes the image content.

The role that clustering plays in the retrieval of images on the basis of content, the retrieval algorithm works to identify images that are similar in content, and then retrieve them. This means that the clustering is a selection of the images before they are retrieved.

II.2.2 Convolutional Neural networks (CNN):

Convolutional Neural networks have been very successful for most computer vision tasks such as image recognition, classification, object detection, and segmentation. Even though CNNs are very successful and give superior results as compared to traditional image processing algorithms, the interpretability of their results remains an important issue to be solved. Indeed, lack of interpretability and explain-ability of how CNN works at its various levels[11].

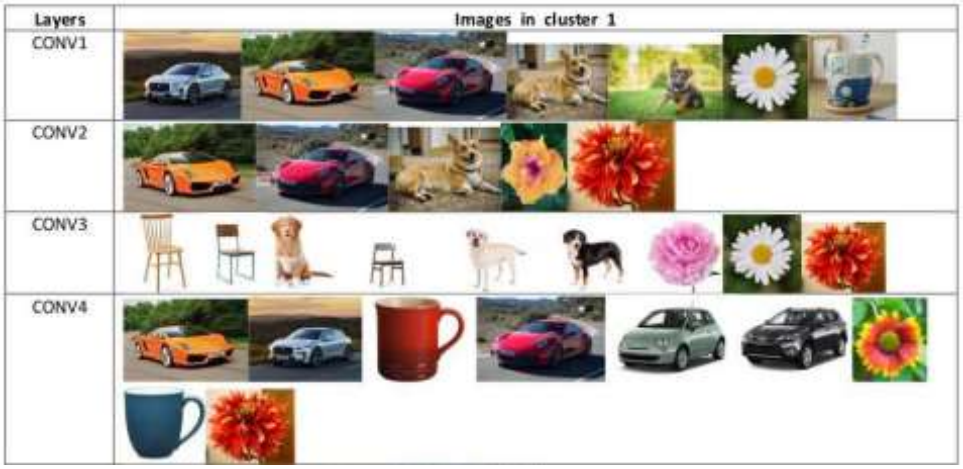


Figure 3. Images clustered based on CNN layer activations

It can be seen from Figure 3 that the clusters in the lower layers (CONV1, CONV2, CONV3, and CONV4) do not reveal class identities, which means that each of these clusters is a mixture of images from different classes. From this, we can infer that the initial convolutional layers encode features that are common to images from all classes which could be basic low-level features such as edges, colors, etc.

II.2.3 K-means clustering:

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

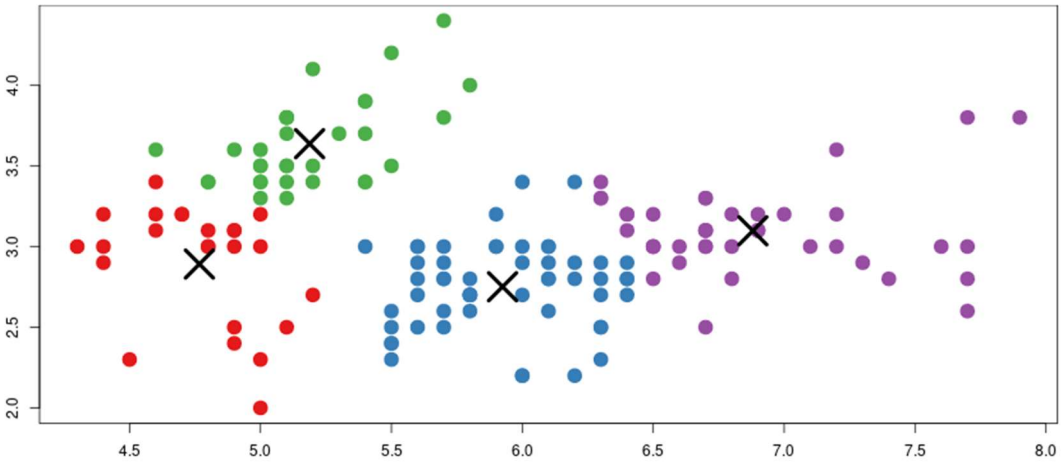


Figure 4. K-means clustering

The images are converted to pixels, represented in a graph, and the number of groups of k is specified. The clustering process is done based on the distance between pixels. Each cluster is created and defined by its centroid (see Figure 3). Each data point is then assigned to its nearest centroid, based on some choice of distance function and distant points are ignored, and this causes incorrect clustering.

II.3 Image clustering application:

We have seen many methodologies and approaches to clustering in machine learning, let's take a quick overview of the implementations of clustering:

- Bioinformatics: Medical imaging.
- Search engines: Search result clustering.
- Sales and marketing: market segmentation.
- Clustering is also used in outlier detection applications such as and detection of credit card fraud.
- Clustering also helps in classifying documents on the web for information discovery.
- In the field of biology, it can be used to derive plant and animal taxonomy, and categorize genes with similar functionalities.

II.4 Conclusion:

In this chapter, we have focused our attention on clustering, some of the techniques that have caused an obstacle to getting better results using visual features, and we've shown the different uses of clustering. In the next chapter, we will focus on semantic similarities and methods of modeling the difference between them.

Chapter III:

Semantic Similarity

III.1 Introduction:

One of the major issues in semantic similarity research is related to natural language processing NLP as it plays an important role in information retrieval, information mining, text mining, web mining, and many other applications such as artificial intelligence and cognitive science as well. The semantic similarity has been used in many scientific assessments and measurements as well as to decipher the complex interface that runs behind the process of sensory perception for a long time.

we provide details about Semantic similarity. Then, we then introduce models to measure the semantic similarity between concepts. Finally, we explain the difference between the methods.

III.2 Semantic similarity (SS)

Semantic similarity is the semantic closeness between two words or the semantic distance between the two words (the two concepts). From the conceptual side, uses of semantic similarity refer to the idea of commonalities in characteristics between any two words or concepts within a language. Although it is a relational property between concepts, it can also be defined as the measurement of conceptual similarity between two or more words.

The similarity between concepts is a quantitative measure of information, which is calculated between concepts according to the properties of concepts and their relationships. Semantic similarity measures have many applications in information extraction (IE) [13], word meaning clarification [14], bioinformatics [15, 16]etc.

All similar concepts may be related but the opposite is not true. Suppose C1 and C2 are concepts that belong to two different nodes N1 and N2 in a given ontology. The similarity between these two concepts is determined by the distance between the nodes N1 and N2. Both N1 and N2 can be thought of as an ontology or classification that contains a set of synonymous terms. The two terms are synonymous if they are in the same node. When we take the issue of SS, our rating system returns a score that lies between 0 and 1, where 0 indicates no similarity and 1 indicates very high similarity.

Computationally, semantic similarity can be estimated by determining topological similarity, using ontology to determine the distance between terms/concepts. We have chosen well-established and widely used measures of semantic similarity at the word level. These are: Resnik similarity [19], Jiang similarity [20], Leacock similarity [21], Lin similarity [22], and

Wu similarity [23]. The methods have been previously used for lexical and textual semantic relatedness.

III.3 Ontology:

The primary use of the word "ontology" is in the discipline of philosophy, where it means "the study or theory of the explanation of Being"; Hence it defines an entity or being and its relationship with and activity in its environment. In other disciplines, such as software engineering and artificial intelligence, it is defined as an “explicit formal specification of a common concept” [17].

Ontology [18], which is used in order to support interoperability and mutual understanding between different parties, is a key component in solving the problem of semantic heterogeneity, enabling semantic interoperability between different web applications and services. Ontology provides a common understanding of the domain that can be communicated between people and heterogeneous and widespread application systems.

The goal of ontology is to achieve knowledge that is common and transferable between people and between application systems. Thus, ontology [27] plays an important role in achieving interoperability across organizations and on the semantic web [28], because it aims to obtain domain knowledge and its role is to create explicit semantics in a general way—the semantics between them.

III.4 WordNet:

WordNet is an on-line lexical reference system developed at Princeton University [24]. WordNet attempts to model the lexical knowledge of a native speaker of English. WordNet can also be seen as ontology for natural language terms. WordNet v.2.0 contains around 100,000 terms, organized into taxonomic hierarchies. Nouns, verbs, adjectives and adverbs are grouped into synonym sets (synsets). The synsets (or concepts) are related to other synsets higher or lower in the hierarchy defined by different types of relationships. Illustrates a fragment of the WordNet Is-A hierarchy.

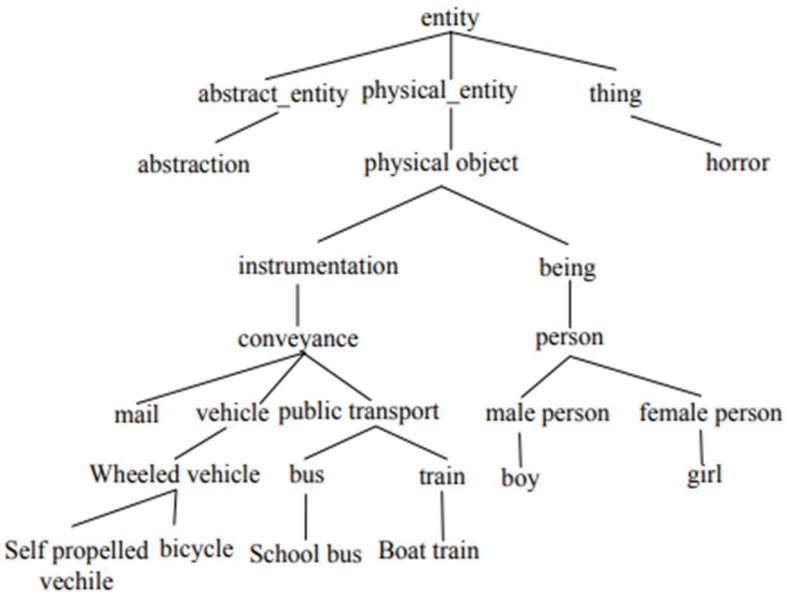


Figure 5. A Fragment of is-a Relation in WordNet

III.5 Word-level Semantic Similarity Measures:

III.5.1 Philip Resnik Similarity (RES):

Resnik's [25] similarity is based on the is-a relationship in the classification of WordNet, where each node represents a unique WordNet concept. According to this scale, two nodes are considered more similar if they share more information or the similarity value is large. This shared information is determined by the information content (IC) of the nodes that comprise these nodes in the classification. Formally, the IC is calculated as:

$$IC = -\log P(C) \qquad (1)$$

Suppose C1 and C2 are concepts in the WordNet classification and the C conceptual node is the lowest common child node of C1 and C2. Moreover, let P(C) be the probability of occurrence of the longest common sub-node C and the probability of the C node is found simply by normalizing the occurrences of concepts with the total number of names in the classification.

$$P(C) = \frac{f(c)}{N} \; and \; f(c) = \sum_{n \in \mathcal{W}} count(n) \qquad (2)$$

Where W (C) is the set of concepts in which the word w occurs and each occurrence of a word is considered a repeat of all concepts containing that word. Resnik similarity is indicated as maximum IC on all concepts to which both words belong. Officially, it is defined as:

$$Sim_{res}(C1, C2) = IC\big(LCS(C1, C2)\big) \qquad (3)$$

Where, LCS is the lowest common subsumer of concept nodes C1 and C2 defined as the common parent of these nodes with minimum nodes distance.

III.5.2 Jiang and Conrath Similarity (JNC):

Jiang et al [20]. use the same concept of information content and take into account the distance between the selected concepts. Regarding this, JNC combines a node-based approach and an edge-based approach. There is no way to discern the semantic similarity between them. However, regarding the semantic similarity between the two concepts, JNC uses the IC values of these concepts along with the IC value of the LCS for these two concepts. Therefore, the similarity will be different because the IC value of the house and the apartment is not the same.

$$distance_{jic}(C1, C2) = IC(C1) + IC(C2) - 2 * IC(LCS) \qquad (4)$$

$$Sim_{jic}(C1, C2) = \frac{1}{distance_{jic}} = \frac{1}{IC(C1) + IC(C2) - 2 * IC(LCS)} \qquad (5)$$

Where, IC stands for information content and LCS is the Lowest Common Subsumer of concepts C1 and C2 defined as the common parent of these with minimum node distance.

III.5.3 Leacock & Chodorow Similarity (LCH):

The similarity of two Lisk [21] concepts is defined as a function of the overlap between the corresponding definitions, as provided by the dictionary. It is based on an algorithm proposed by Lisk (1986) [26] as a solution to demystify the meaning of the word. The Lisk [21] Similarity Scale is not limited to semantic networks but can be used with any dictionary that provides word definitions.

Leacock and Chodorow [21] approach is based on the shortest path length between two nouns in an is-a relationship in WordNet ontology. Basically, the shortest path is that in which there

is a lesser number of intermediate nodes. The value was scaled by depth D where depth is measured as the length of the longest distance from leaf node to root node of the word net hierarchy. The similarity relevance is defined as follows:

$$Sim_{lch}(C1, C2) = -\log \frac{length}{2 * D} \qquad (6)$$

Where min (length (C1, C2)) is the shortest path between two concepts where D is the maximum depth in the WordNet ontology. The approach focuses on nodes rather than links, so synsets (synonym words) are only one unit distance from each other.

III.5.4 Wu and Palmer Similarity (WUP) [23]:

The similarity metric measures the depth of two given concepts in the WordNet taxonomy, and the depth of the least common subsumer (LCS), and combines these figures into a similarity score:

$$Sim_{wup}(C1, C2) = \frac{2 * depth(LCS)}{depth(C1) + depth(C2)} \qquad (7)$$

III.5.5 Lin Similarity

The similarity between the two terms should be measured as the ratio between the amount of information needed to explain their commonalities and the information needed to fully describe them. Lin [22], Jiang & Conrath [20] extended the Resnick scale of IC by incorporating IC of individual concepts. Lane determined the similarity between the two concepts by taking the quotient between the double IC of the LCS concept and the sum of the IC of the two concepts as shown in the equation. This is similar to the procedure suggested by Wu & Palmer; The difference in the use of IC rather than the depth of concepts

.

$$Sim_{Lin}(C1, C2) = \frac{2 * IC(LCS)}{IC(C1) + IC(C2)} \qquad (8)$$

III.6 Comparison of Different Semantic Similarity Measures [30]:

Category	Principle	Measure	Features	Advantages	Disadvantages
Path based	Function of path length linking the concepts and the position of the concepts in the taxonomy	W&P	Path length to subsumer, scaled by subsumer path to root	Simple	Two pairs with the same lso and equal lengths of shortest path will have the same similarity
		L&C	Count of edges between and log smoothing	Simple	Two pairs with equal lengths of shortest path will have the same similarity
IC based	The more common information two concepts share, the more similar the concepts are.	Resnik	IC of lso	Simple	Two pairs with the same lso will have the same similarity
		Lin	IC of lso and the compared concepts	Take the IC of compared concepts into considerate	Two pairs with the same summation of IC(c1) and IC(c2) will have the same similarity
		Jiang	IC of lso and the compared concepts	Take the IC of compared concepts into considerate	Two pairs with the same summation of IC(c1) and IC(c2) will have the same similarity

Table 1. Different Semantic Similarity Measures.

III.7 Conclusion:

In this chapter, we have discussed in general terms the semantic similarity and the words in WordNet. In addition to that, we have explained the different methods of calculating semantic similarity and the difference between them. In the next chapter, we will introduce our algorithm for image clustering and calculating semantic similarity between concepts using an ontology.

Chapter IV:

Applying Semantic Similarity in cluster images

IV.1 Introduction:

In this chapter, we will study the practical side, how semantic similarity works, and its steps for clustering images. Then we provide the details of our experimental setup scheme. Finally, we will report and discuss the results obtained.

IV.2 Semantic Similarity in Cluster images:

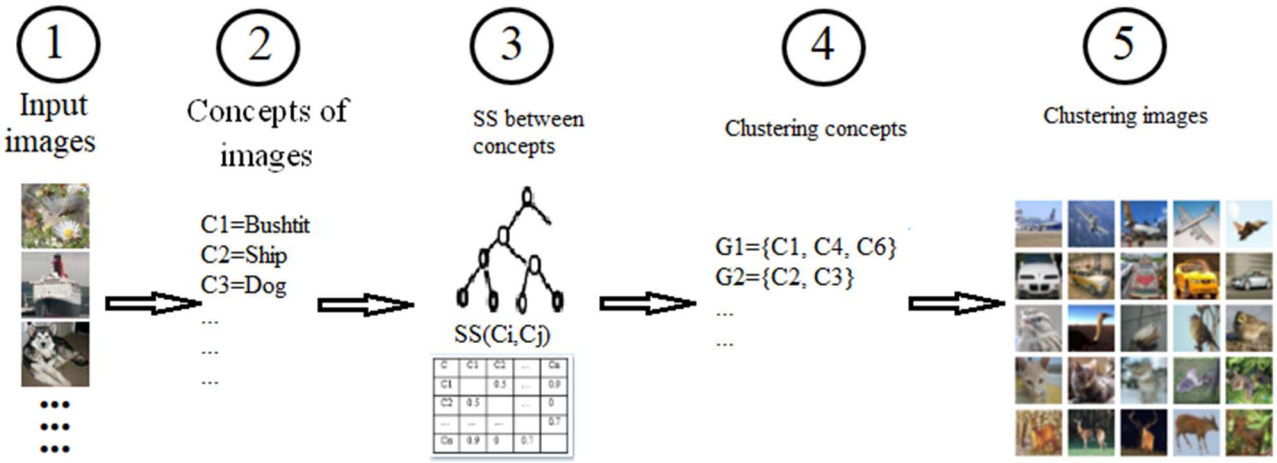


Figure 6. Illustration of the steps for clustering images in our algorithm.

To cluster images in our algorithm is divided into 5 steps as illustrated in figure 5

1. Input images:

To accomplish the practical work, we used in our thesis a database of 320 image elements. From these images we extract the concepts for which we calculate the semantic similarity between them, the number of concepts is not equal to the number of images. Concepts are represented in an ontology.

2. Concepts of images:

We present the concepts that were used in our study to cluster the images, which consist of 45 concepts. which is next:

1. Animal: 20 concepts

A. Predator: 7 concepts
   - Wild: Lion, Tiger, Crocodile, Snake.
   - Marin: Shark, Lion fish, Fangtooth.

B. Pets: 13 concepts
   - Wild: Dog, Horse, Sheep, Chicken, Cat, Shepherd dog.
   - Aerial: Sparrow, Pigeon, Parrot.
   - Aquatic: Dolphin, Salmon, Tuna, Alaska pollock.

2. Transportation: 10 concepts
   - Wild: Car, Bus, Truck, Bike.
   - Marin: Submarine, Ship, Boats.
   - Aerial: Space shuttle, Civil plane, Glider
3. Plants: 5 concepts
   - Wild: Trees, Flowers, Herbs.
   - Aquatic: Water lilies, Nelumbo nucifera.
4. Buildings: 5 concepts
   - Hotel, Government, Warehouse, Cottage, House.
5. Cleaning: 5 concepts
   - Mop, Broom, Squeegee, Sponge, Hose.

3. Our ontology:

In computer science and information science, an ontology encompasses a representation, formal naming, and definition of the categories, properties, and relations between the concepts, data, and entities that substantiate one, many, or all domains of discourse. More simply, ontology is a way of showing the properties of a subject area and how they are related, by defining a set of concepts and categories that represent the subject.

Representation of our Ontology:

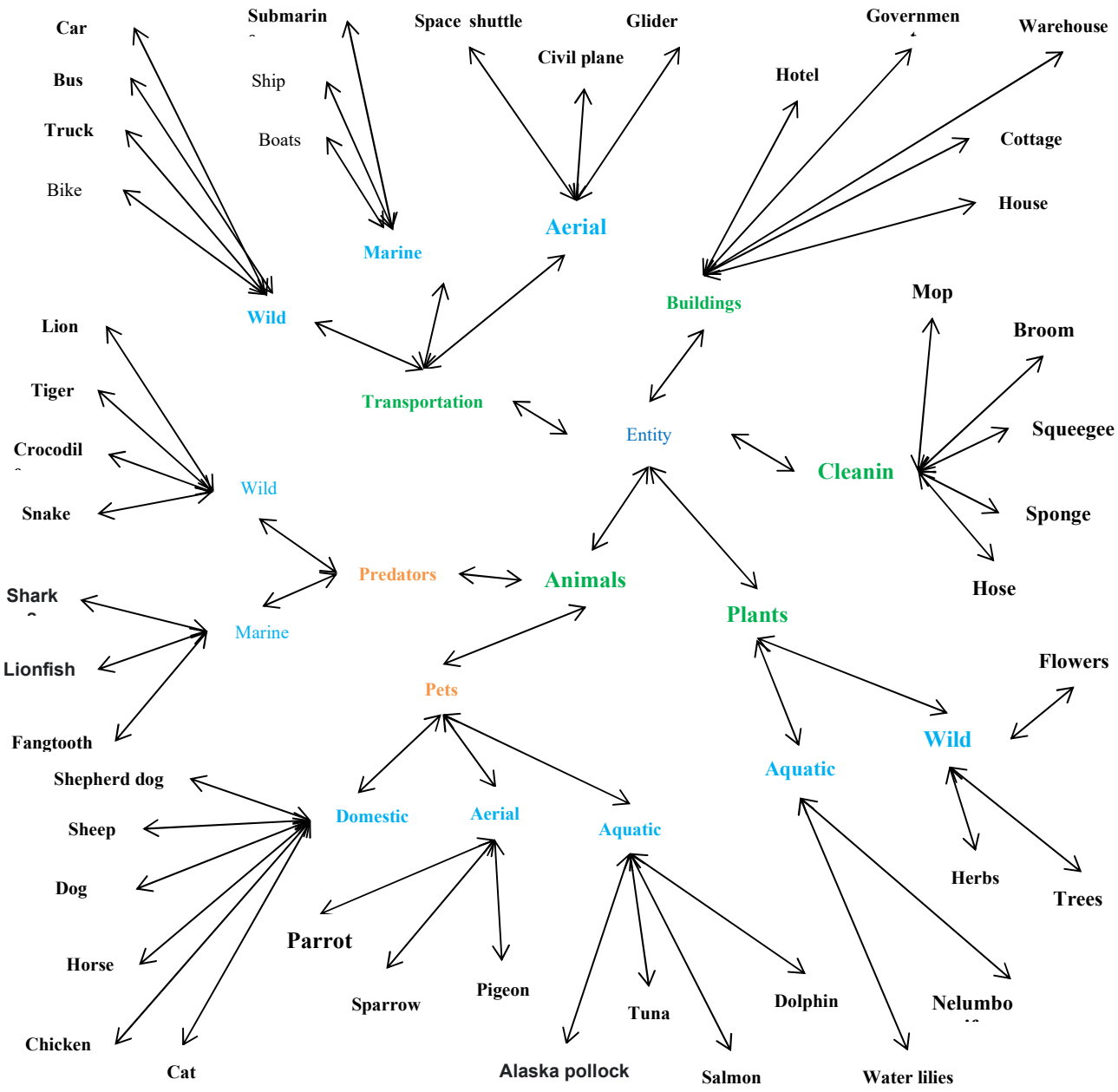


Figure 7. Ontology of Concepts 45

4. Clustering concepts:

Apply both Wu & Palmer and Lin methods to calculate semantic similarity between concepts.

4.1 Wu and Palmer Similarity:

In order to apply Wu and Palmer method, we will calculate the depth for C1 and C2, the depth is the distance or the number of relationships between the root and leaf node (concept), then calculate the depth of LCS.

Appendix A

$$Annex\ 1: C1 = Lion,\ \ C2 = Sheep\ \Rightarrow\ Sim_{wup}(C1, C2) = 0.5$$

$$Annex\ 2: C1 = Cat,\ \ C = dog\ \Rightarrow\ Sim_{wup}(C1, C2) = 0.75$$

$$Annex\ 3: C1 = Car,\ \ C2 = Cat\ \Rightarrow\ Sim_{wup}(C1, C2) = 0$$

Calculation of SS between concepts:

C	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	…
C1	-	0	0	0	0	0	0	0	0	0.33	0.33	0.67	0	0	0	0	0	0	0	…
C2	0	-	0.5	0.25	0.25	0.25	0	0	0	0	0	0	0.75	0.5	0.25	0.25	0.25	0	0	…
C3	0	0.5	-	0.25	0.25	0.25	0	0	0	0	0	0	0.5	0.75	0.25	0.25	0.25	0	0	…
C4	0	0.25	0.25	-	0.5	0.5	0	0	0	0	0	0	0.25	0.25	0.75	0.5	0.5	0	0	…
C5	0	0.25	0.25	0.5	-	0.5	0	0	0	0	0	0	0.25	0.25	0.5	0.75	0.5	0	0	…
C6	0	0.25	0.25	0.5	0.5	-	0	0	0	0	0	0	0.25	0.25	0.5	0.5	0.75	0	0	…
C7	0	0	0	0	0	0	-	0.33	0	0	0	0	0	0	0	0	0	0.67	0.33	…
C8	0	0	0	0	0	0	0.33	-	0	0	0	0	0	0	0	0	0	0.33	0.67	…
C9	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	…
C10	0.33	0	0	0	0	0	0	0	0	-	0.33	0.33	0	0	0	0	0	0	0	…
C11	0.33	0	0	0	0	0	0	0	0	0.33	-	0.33	0	0	0	0	0	0	0	…
C12	0.67	0	0	0	0	0	0	0	0	0.33	0.33	-	0	0	0	0	0	0	0	…
C13	0	0.75	0.5	0.25	0.25	0.25	0	0	0	0	0	0	-	0.5	0.25	0.25	0.25	0	0	…
C14	0	0.5	0.75	0.25	0.25	0.25	0	0	0	0	0	0	0.5	-	0.25	0.25	0.25	0	0	…
C15	0	0.25	0.25	0.75	0.5	0.5	0	0	0	0	0	0	0.25	0.25	-	0.5	0.5	0	0	…
C16	0	0.25	0.25	0.5	0.75	0.5	0	0	0	0	0	0	0.25	0.25	0.5	-	0.5	0	0	…
C17	0	0.25	0.25	0.5	0.5	0.75	0	0	0	0	0	0	0.25	0.25	0.5	0.5	-	0	0	…
C18	0	0	0	0	0	0	0.76	0.33	0	0	0	0	0	0	0	0	0	-	0.33	…
C19	0	0	0	0	0	0	0.33	0.67	0	0	0	0	0	0	0	0	0	0.33	-	…
C20	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	…
C21	0.33	0	0	0	0	0	0	0	0	0.67	0.33	0.33	0	0	0	0	0	0	0	…
C22	0.33	0	0	0	0	0	0	0	0	0.33	0.67	0.33	0	0	0	0	0	0	0	…
C23	0.67	0	0	0	0	0	0	0	0	0.33	0.33	0.67	0	0	0	0	0	0	0	…
C24	0	0.75	0.5	0.25	0.25	0.25	0	0	0	0	0	0	0.75	0.5	0.25	0.25	0.25	0	0	…
C25	0	0.5	0.75	0.25	0.25	0.25	0	0	0	0	0	0	0.5	0.75	0.25	0.25	0.25	0	0	…
C26	0	0.25	0.25	0.75	0.5	0.5	0	0	0	0	0	0	0.25	0.25	0.75	0.5	0.5	0	0	…
C27	0	0.25	0.25	0.5	0.5	0.75	0	0	0	0	0	0	0.25	0.25	0.5	0.5	0.75	0	0	…
C28	0	0	0	0	0	0	0.33	0.67	0	0	0	0	0	0	0	0	0	0.33	0.67	…
C29	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	…
C30	0.33	0	0	0	0	0	0	0	0	0.67	0.33	0.33	0	0	0	0	0	0	0	…
C31	0.33	0	0	0	0	0	0	0	0	0.33	0.67	0.33	0	0	0	0	0	0	0	…
C32	0.67	0	0	0	0	0	0	0	0	0.33	0.33	0.67	0	0	0	0	0	0	0	…
C33	0	0.75	0.5	0.25	0.25	0.25	0	0	0	0	0	0	0.75	0.5	0.25	0.25	0.25	0	0	…
C34	0	0.25	0.25	0.75	0.5	0.5	0	0	0	0	0	0	0.25	0.25	0.75	0.5	0.5	0	0	…
C35	0	0.25	0.25	0.5	0.5	0.75	0	0	0	0	0	0	0.25	0.25	0.5	0.75	0.5	0	0	…
C36	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	…
C37	0	0.25	0.25	0.75	0.5	0.5	0	0	0	0	0	0	0.25	0.25	0.75	0.5	0.5	0	0	…
C38	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	…
C39	0	0.25	0.25	0.75	0.5	0.5	0	0	0	0	0	0	0.25	0.25	0.75	0.5	0.5	0	0	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Table 2. Matrix word to word (WUP)

4.2 Lin Similarity:

To apply Lin method, we will calculate the probability for each concept, all concepts have the same probability and the concept is equal to the sum of the concepts. Then we calculate the LCS probability which is the probability of occurrence of repetition in the number of concepts.

Appendix B

$$Annex\ 1: C1 = Lion,\ \ C2 = Sheep\ \Rightarrow\ Sim_{Lin}(C1, C2) = 0.37$$

$$Annex\ 2: C1 = Cat,\ \ C2 = Dog\ \Rightarrow\ Sim_{Lin}(C1, C2) = 0.49$$

Calculation of SS between concepts:

C	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	…
C1	-	0.43	0.46	0.37	0.51	0.43	0.51	0.46	0.40	0.46	0.46	0.62	0.43	0.46	0.37	0.51	0.43	0.51	0.46	…
C2	0.43	-	0.46	0.37	0.51	0.43	0.51	0.46	0.40	0.46	0.46	0.43	0.62	0.46	0.37	0.51	0.43	0.51	0.46	…
C3	0.46	0.46	-	0.40	0.56	0.46	0.56	0.51	0.43	0.51	0.51	0.46	0.46	0.70	0.40	0.56	0.46	0.56	0.51	…
C4	0.37	0.37	0.40	-	0.43	0.37	0.43	0.40	0.34	0.40	0.40	0.37	0.37	0.40	0.51	0.43	0.37	0.43	0.40	…
C5	0.51	0.51	0.56	0.43	-	0.51	0.62	0.56	0.46	0.56	0.56	0.51	0.51	0.56	0.43	0.81	0.51	0.62	0.56	…
C6	0.43	0.43	0.46	0.37	0.51	-	0.51	0.46	0.40	0.46	0.46	0.43	0.43	0.46	0.37	0.51	0.62	0.51	0.46	…
C7	0.51	0.51	0.56	0.43	0.62	0.51	-	0.56	0.46	0.56	0.56	0.51	0.51	0.56	0.43	0.62	0.51	0.81	0.56	…
C8	0.46	0.46	0.51	0.40	0.56	0.46	0.56	-	0.43	0.51	0.51	0.46	0.46	0.51	0.40	0.56	0.46	0.56	0.7	…
C9	0.40	0.40	0.43	0.34	0.46	0.40	0.46	0.43	-	0.43	0.43	0.40	0.40	0.43	0.34	0.46	0.40	0.46	0.43	…
C10	0.46	0.46	0.51	0.40	0.56	0.46	0.56	0.51	0.43	-	0.51	0.46	0.46	0.51	0.40	0.56	0.46	0.56	0.51	…
C11	0.46	0.46	0.51	0.40	0.56	0.46	0.56	0.51	0.43	0.51	-	0.46	0.46	0.51	0.40	0.56	0.46	0.56	0.51	…
C12	0.62	0.43	0.46	0.37	0.51	0.43	0.51	0.46	0.40	0.46	0.46	-	0.43	0.46	0.37	0.51	0.43	0.51	0.46	…
C13	0.43	0.62	0.46	0.37	0.51	0.43	0.51	0.46	0.40	0.46	0.46	0.43	-	0.46	0.37	0.51	0.43	0.51	0.46	…
C14	0.46	0.46	0.70	0.40	0.56	0.46	0.56	0.51	0.43	0.51	0.51	0.46	0.46	-	0.40	0.56	0.46	0.56	0.51	…
C15	0.37	0.37	0.40	0.51	0.43	0.37	0.43	0.40	0.34	0.40	0.40	0.37	0.37	0.40	-	0.43	0.37	0.43	0.40	…
C16	0.51	0.51	0.56	0.43	0.81	0.51	0.62	0.56	0.46	0.56	0.56	0.51	0.51	0.56	0.43	-	0.51	0.62	0.56	…
C17	0.43	0.43	0.46	0.37	0.51	0.62	0.51	0.46	0.40	0.46	0.46	0.43	0.43	0.46	0.37	0.51	-	0.51	0.46	…
C18	0.51	0.51	0.56	0.43	0.62	0.51	0.81	0.56	0.46	0.56	0.56	0.51	0.51	0.56	0.43	0.62	0.51	-	0.56	…
C19	0.51	0.51	0.56	0.43	0.62	0.51	0.81	0.56	0.46	0.56	0.56	0.51	0.51	0.56	0.43	0.62	0.51	0.81	-	…
C20	0.40	0.40	0.43	0.34	0.46	0.40	0.46	0.43	0.56	0.43	0.43	0.40	0.40	0.43	0.34	0.46	0.40	0.46	0.43	…
C21	0.46	0.46	0.51	0.40	0.56	0.46	0.56	0.51	0.43	0.7	0.51	0.46	0.46	0.51	0.40	0.56	0.46	0.56	0.51	…
C22	0.46	0.46	0.51	0.40	0.56	0.46	0.56	0.51	0.43	0.51	0.7	0.46	0.46	0.51	0.40	0.56	0.46	0.56	0.51	…
C23	0.62	0.43	0.46	0.37	0.51	0.43	0.51	0.46	0.40	0.46	0.46	0.62	0.43	0.46	0.37	0.51	0.43	0.51	0.46	…
C24	0.43	0.62	0.46	0.37	0.51	0.43	0.51	0.46	0.40	0.46	0.46	0.43	0.62	0.46	0.37	0.51	0.43	0.51	0.46	…
C25	0.46	0.46	0.70	0.40	0.56	0.46	0.56	0.51	0.43	0.51	0.51	0.46	0.46	0.70	0.40	0.56	0.46	0.56	0.51	…
C26	0.37	0.37	0.40	0.51	0.43	0.37	0.43	0.40	0.34	0.40	0.40	0.37	0.37	0.40	0.51	0.43	0.37	0.43	0.40	…
C27	0.43	0.43	0.46	0.37	0.51	0.62	0.51	0.46	0.40	0.46	0.46	0.43	0.43	0.46	0.37	0.51	0.62	0.51	0.46	…
C28	0.51	0.51	0.56	0.43	0.62	0.51	0.81	0.56	0.46	0.56	0.56	0.51	0.51	0.56	0.43	0.62	0.51	0.81	0.56	…
C29	0.40	0.40	0.43	0.34	0.46	0.40	0.46	0.43	0.56	0.43	0.43	0.40	0.40	0.43	0.34	0.46	0.40	0.46	0.43	…
C30	0.46	0.46	0.51	0.40	0.56	0.46	0.56	0.51	0.43	0.7	0.51	0.46	0.46	0.51	0.40	0.56	0.46	0.56	0.51	…
C31	0.46	0.46	0.51	0.40	0.56	0.46	0.56	0.51	0.43	0.51	0.7	0.46	0.46	0.51	0.40	0.56	0.46	0.56	0.51	…
C32	0.62	0.43	0.46	0.37	0.51	0.43	0.51	0.46	0.40	0.46	0.46	0.62	0.43	0.46	0.37	0.51	0.43	0.51	0.46	…
C33	0.43	0.62	0.46	0.37	0.51	0.43	0.51	0.46	0.40	0.46	0.46	0.43	0.62	0.46	0.37	0.51	0.43	0.51	0.46	…
C34	0.37	0.37	0.40	0.51	0.43	0.37	0.43	0.40	0.34	0.40	0.40	0.37	0.37	0.40	0.51	0.43	0.37	0.43	0.40	…
C35	0.43	0.43	0.46	0.37	0.51	0.62	0.51	0.46	0.40	0.46	0.46	0.43	0.43	0.46	0.37	0.51	0.62	0.51	0.46	…
C36	0.40	0.40	0.43	0.34	0.46	0.40	0.46	0.43	0.56	0.43	0.43	0.40	0.40	0.43	0.34	0.46	0.40	0.46	0.43	…
C37	0.37	0.37	0.40	0.51	0.43	0.37	0.43	0.40	0.34	0.40	0.40	0.37	0.37	0.40	0.51	0.43	0.37	0.43	0.40	…
C38	0.40	0.40	0.43	0.34	0.46	0.40	0.46	0.43	0.56	0.43	0.43	0.40	0.40	0.43	0.34	0.46	0.40	0.46	0.43	…
C39	0.37	0.37	0.40	0.51	0.43	0.37	0.43	0.40	0.34	0.40	0.40	0.37	0.37	0.40	0.51	0.43	0.37	0.43	0.40	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Table 3. Matrix word to word (LIN)

5. Concepts clustering:

Groups that have similar concepts for both methods Wu & Palmer and Lin. so that they are different from the rest of the groups.

1. Wu & Palmer groups

The groups are formed according to the closeness of the semantic similarity between the concepts and the apparent results in which there is a clear discrepancy in defining the areas of the groups.

$$Cluster_1\{C_{10}, C_{22}, C_{33}\};\ Cluster_2\{C_4, C_{16}, C_{27}\};\ Cluster_3\{C_9, C_{20}, C_{38}\}$$

2. Lin groups

The results shown from the semantic similarity calculation require a field to be defined so that the concepts are combined into one group:

$$Sim \geq 0.7 \Rightarrow Cluster_1\{C_3, C_{14}\};\ Cluster_2\{C_5, C_{16}\};\ Cluster_3\{C_{10}, C_{11}, C_{21}\}$$

$$0.5 > Sim > 0.7 \Rightarrow Cluster_1\{C_4, C_{37}, C_{39}\};\ Cluster_2\{C_{13}, C_{33}\}$$

6. Image clustering:

After the process of clustering the concepts into groups, the role comes to the pictures. The group of grouped pictures is the result of clustering concepts into groups. We can also determine the number and type of group that appears to us, and the result appears as follows:

Figure 8. Image clustering

IV.3 Clustering images Algorithm

Summarize the steps of the clustering process in the algorithm

Algorithm: Clustering images with SS

01: Begin
02: **INPUT**: I = {I1, I2, … In}
03: Extraction: C = {C1, C2, … Cm}
04: Measure the SS between concepts
05: Using Eq. (7) or Eq. (8)
06: Compute Sim (Ci, Cj) i and j to n, i=1 with $i \neq j$
07: **if** Sim (Ci, Cj) $\geq \alpha$ (Threshold)
08: Then Select Ci and Cj are in the same cluster
09: **else**
10: Then Select Ci and Cj aren't in the same cluster
11: **end if**
12: **OUTPUT**: Cluster images { Cluster 1, Cluster 2, … Cluster y}.
13: End

The algorithm focuses on depth to calculate the semantic similarity between concepts. At first, we will define the database to cluster the images that have a semantic affinity, then extract the concepts, then determine the depth of the two concepts and calculate the semantic similarity according to Eq (7) or Eq (8). To cluster which has affinity and similarity value equal to α, we compare the semantic similarity between concepts, identify the concepts that have semantic similarity greater or equal to α, and then group the images for these concepts.

IV.4 Experimentation and Validation:

We will display the database used. Determining the cluster number of human judgments in order to calculate the aggregation accuracy of both Wu & Palmer Lin and the best method between them. Then we display the results by the selected clusters.

IV.4.1 Datasets:

To demonstrate the effectiveness of the proposed approach, we performed a database procedure consisting of 320 elements of various images, the images, the images on which we are experimenting are about the concepts presented in the thesis, and all the image concepts are represented on the ontology.

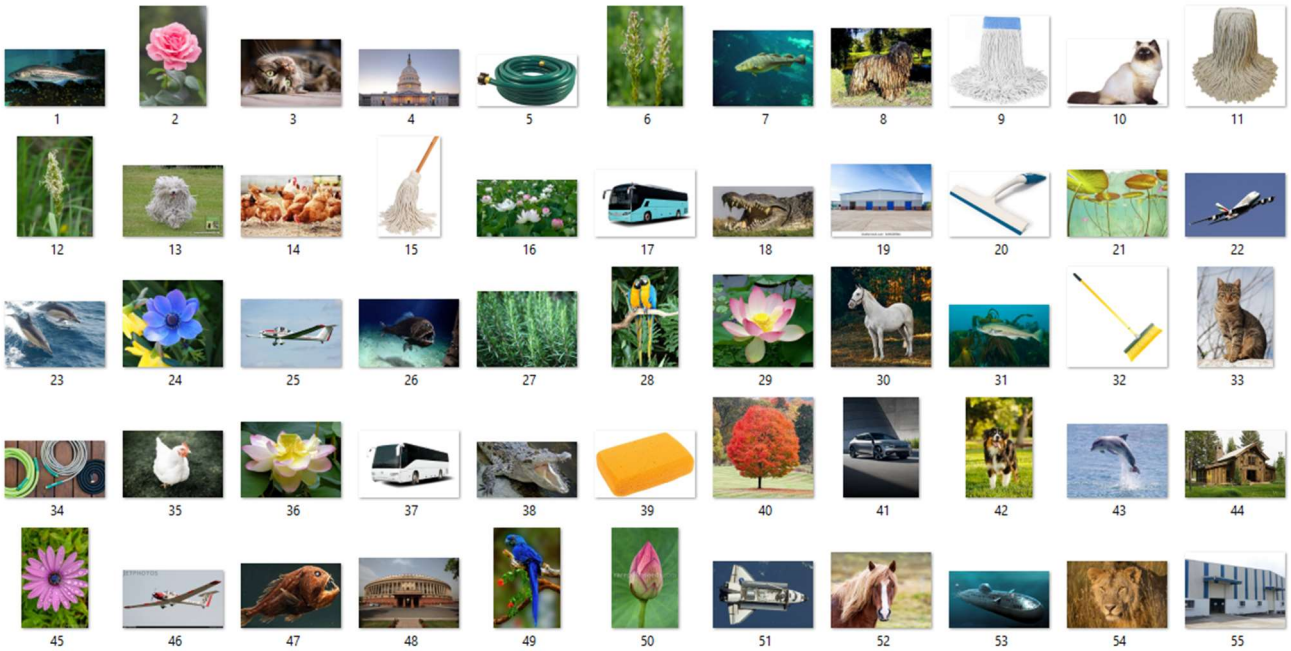


Figure 9. Capture our Datasets

IV.4.2 Experimental results:

IV.4.2.1 Human judgments:

To verify the accuracy of the effectiveness of the approach proposed in the thesis, we have transferred the experiment to reality to see human judgment in the clustering of images.

Person 1:

Cluster 1: 45, 50, 79 and 113 (Flowers)

Cluster 2: 143, 150, 152 and 153 (Ship)

Cluster 3: 4, 78 (Building)

Person 2:

Cluster 1: 13, 42 and 107 (Dog)

Cluster 2: 61, 92, 115 and 128 (Shark)

Cluster 3: 44, 103 and 120 (House)

Person 3:

Cluster 1: 83, 99, 103 and 145 (House)

Cluster 2: 21, 56 and 168 (Plants)

Cluster 3: 10, 33 and 64 (Cat)

Person 4:

Cluster 1: 18, 131, 204, 42 and 92 (Predators Animals).

Cluster 2: 24, 45, 50, 89, 116 and 199 (Wild Planets).

Cluster 3: 99 and 135 (Hotels).

Person 5:

Cluster 1: 5, 11, 15, 39 and 104 (Cleaning).

Cluster 2: 61, 26, 88, 188 and 144 (Marine Predators Animals).

Cluster 3: 25, 73, 155, 181 and 216(Civil plane).

Person 6:

Cluster 1: 35, 66, 98 and 187 (Chicken).

Cluster 2: 17, 68, 110, 126 and 217 (Transportation).

Cluster 3: 75, 83, 103 and 145 (House).

Person 7:

Cluster 1: 9, 11, 15, 20 and 32 (Broom).

Cluster 2: 152, 182, 195 and 217 (Ship).

Cluster 3: 4, 19, 48, 58 and 165 (Building).

Person 8:

Cluster 1: 2, 24, 45, 76 and 100 (Wild Plants _Flowers).

Cluster 2: 152, 182, 195 and 217 (Ship).

Cluster 3: 4, 19, 48, 58 and 165 (Building).

Person 9:

Cluster 1: 13, 30, 66, 90 and 173 (Pets Animals).

Cluster 2: 44, 55, 123, 137 and 146 (Warehouse).

Cluster 3: 70, 105, 158 and 199 (Trees).

Person 10:

Cluster 1: 136, 227, 229 and 299 (Building).

Cluster 2: 282, 289, 316, 317 and 146 (Domestic).

Cluster 3: 242, 285, 294, 308 and 318 (Transportation).

Person 11:

Cluster 1: 245, 299, 310 and 312 (Transportation).

Cluster 2: 202, 254, 259, 276 and 284 (Sparrow).

Cluster 3: 227, 235, 297 and 306 (Government).

Person 12:

Cluster 1: 148, 180, 228 and 288 (Transportation).

Cluster 2: 131, 174, 176 and 212 (Predators Animals).

Cluster 3: 274, 271, 282 and 286 (Ship).

Person 13:

Cluster 1: 96, 293, 307 and316 (Cat).

Cluster 2: 9, 20, 32, 39 and 104 (Cleaning).

Cluster 3: 54, 122, 133, 176, 221 and 222(Wild Predators Animals).

Person 14:

Cluster 1: 3, 13, 42, 64 and 107 (Domestic).

Cluster 2: 126, 140, 157, 196 and 205 (Bick).

Cluster 3: 87, 272, 296 and 300 (Truck).

Person 15:

Cluster 1: 49, 90, 117, 132 and 161 (Aerial Pets Animals).

Cluster 2: 44, 48, 75, 9,9 112 and 136 (Building).

Cluster 3: 17, 37, 68, 141 and 310 (Bus).

Person 16:

Cluster 1: 1, 7, 23, 43, 69 and 92 (Aquatic Animals).

Cluster 2: 178, 214, 214, 271 and 286 (Flowers).

Cluster 3: 247, 271 and 286 (Ship).

Cluster 4: 110, 114, 155, 194 and 196 (Civil plane).

Person 17:

Cluster 1: 3, 33, 186, 283 and 316 (Cat).

Cluster 2: 19, 44, 99, 229 and 314 (Building).

Cluster 3: 179, 198, 263 and 305 (Boat).

Person 18:

Cluster 1: 17. 73. 252. 167 and 268 (Wild Transportation).

Cluster 2: 18, 172, 174, 234 and 292 (Predators Animals).

Cluster 3: 133, 201, 222, 265 and 266 (Tiger).

Person 19:

Cluster 1: 21, 63, 147, 199 and 270 (Plants).

Cluster 2: 153, 195, 225, 285 and 315 (Ship)

Cluster 3: 49, 59, 161, 259 and 289 (Aerial Animal).

Person 20:

Cluster 1: 99, 135, 304 and 314 (Hotel).

Cluster 2: 15, 32, 39 and 104 (Cleaning).

Cluster 3: 22, 110, 151, 249 and 301 (Aerial Transportation).

Based on different human judgments and clustering for each person, we can suggest a comprehensive number of combinations for them is K = 9.

1. Wild Predator Animal.
2. Marin Predator Animal.
3. Aquatic Pets Animal.
4. Aerial Pets Animal.
5. Domestic Pets Animal.
6. Transportation.
7. Plants.
8. Building.
9. Cleaning.

Datasets	Images	Cluster	Average concepts in each Cluster
Our datasets	320	9	5

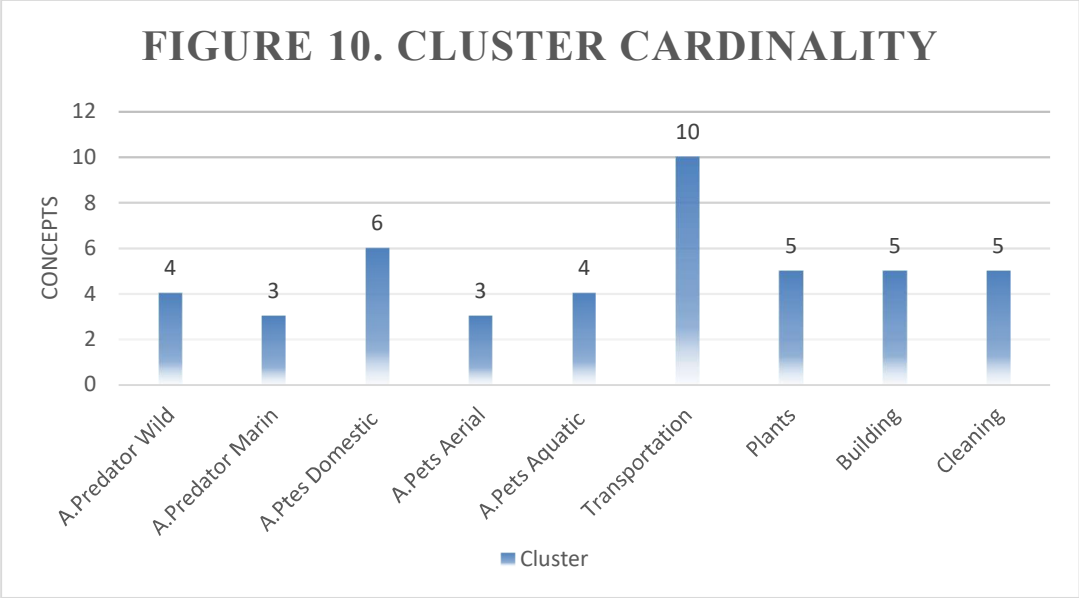
Table 4. Image datasets using in our experiments

Calculate the average (Ave) of the concepts in each cluster

$$Ave = \frac{Total\ number\ of\ concepts}{Number\ cluster} = \frac{45}{9} = 5$$

IV.4.2.2 Cluster cardinality:

Cluster cardinality is the number of examples per cluster:



To determine the accuracy of image clustering for both Wu & Palmer and Lin, we compute the accuracy at different thresholds:

Wu & Palmer [23]	Threshold	0.5	0.7
	Number cluster	7	5
	Accuracy	77.78%	55.56%

Table 5. Clustering accuracy (WUP)

$$Acc_{WUP} = \frac{Number\ cluster}{Total\ number\ of\ cluster} = \frac{7}{9} = 77.78\%$$

The results of calculating the accuracy concerning Wu & Palmer, give good results at the threshold value of 0.5 by 77.78%, even with the increase in the threshold value, but the accuracy remains good, and the accuracy of clustering when calculating the semantic similarity using the depth for each of the two thresholds.

Lin [22]	Threshold	0.5	0.7
	Number cluster	8	2
	Accuracy	88.89%	22.22%

Table 6. Clustering accuracy (LIN)

As for the results for Lin, it gave excellent results at the threshold value of 0.5 by 88.89, but the increase in the threshold value reduces the accuracy to become weak.

Methods of SS	Original cluster size	Correctly clustered groups	Accuracy
Wu & Palmer [23]	9	7	77.78%
Lin [22]	9	8	88.89%

Table 7. Clustering accuracy comparison

The accuracy results were very close, but Lin, who relies on information content for image clustering, is 88.89% better than Wu & Palmer, who relies on depth versus 77.78%.

The increase in the threshold value makes Wu & Palmer better than Lin in clustering.

IV.4.3.3 Clustering results:

We will perform some results of Lin method that showed better accuracy, and show the various clustering that have been identified from human judgments.

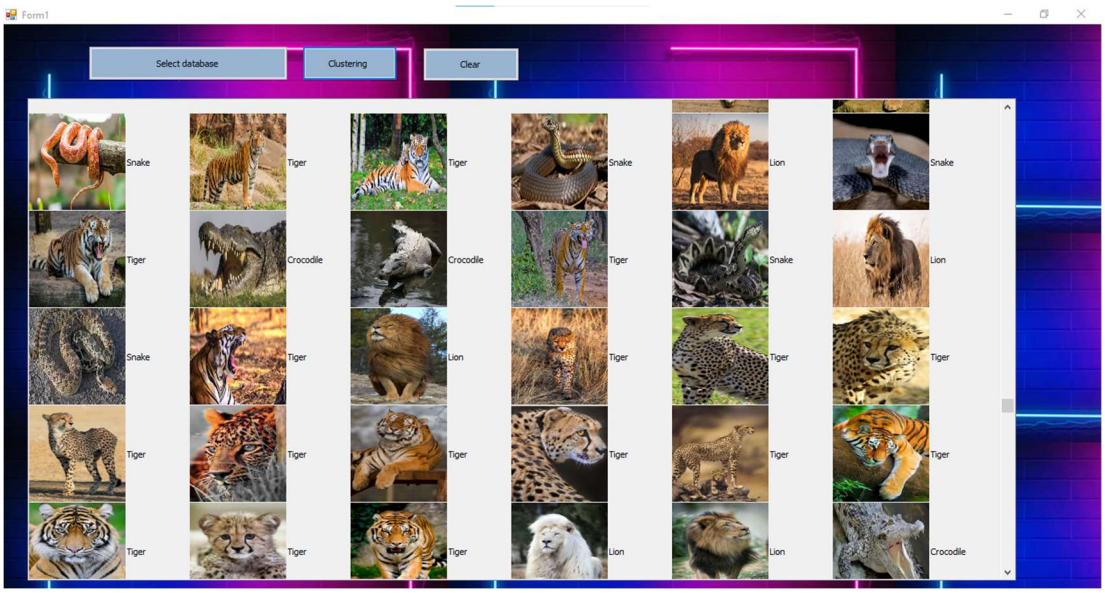


Figure 11. Wild Predator Animals images clustering

Figure 11 shows the results of a clustering that contains all images of wild predatory animals.

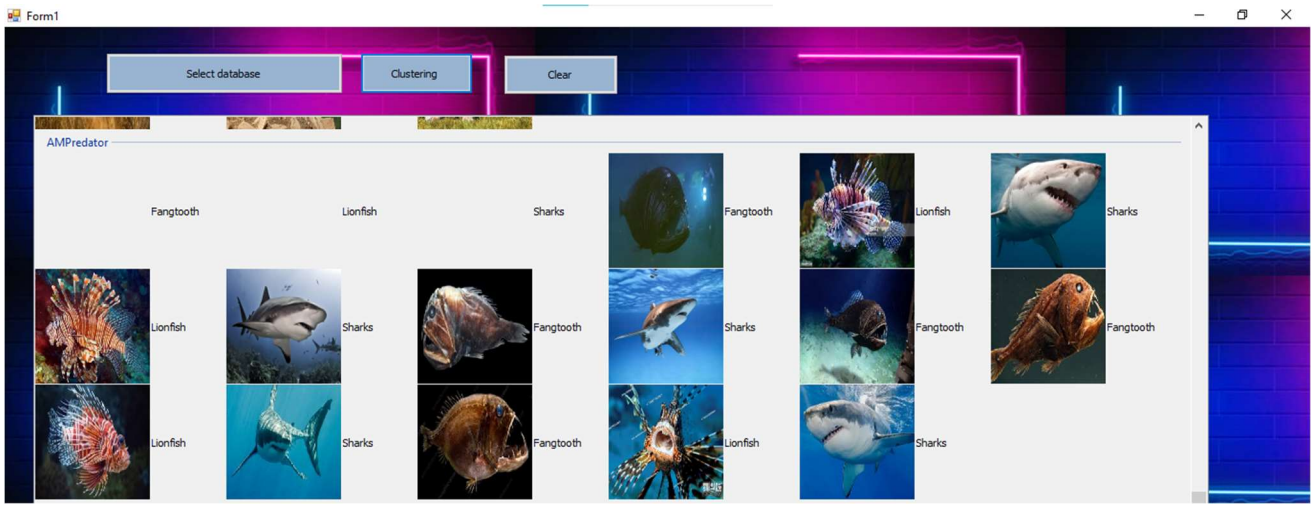


Figure 12. Marin Predator Animals images clustering

Figure 12 shows the results of a clustering that contains all images of Marin predatory animals.

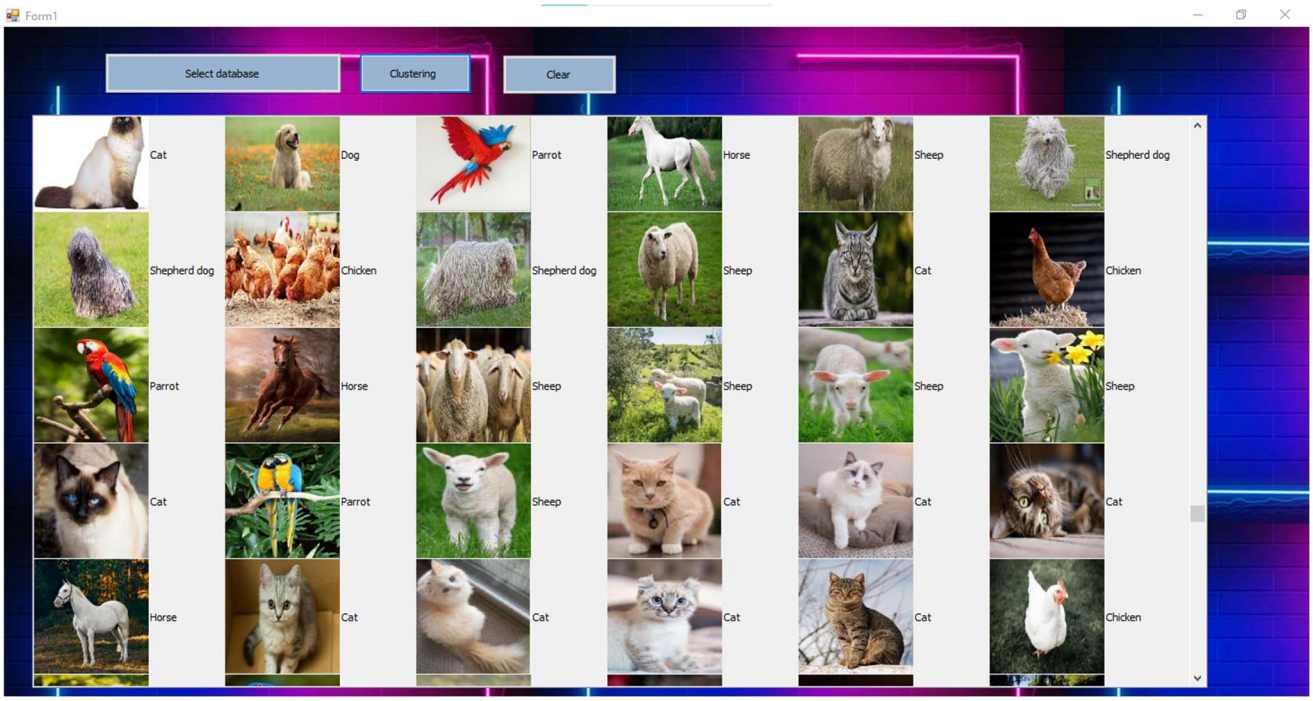


Figure 13. Domestic Pets Animals images clustering

Figure 13 shows the results of a clustering that contains all images of Domestic Pets animals.

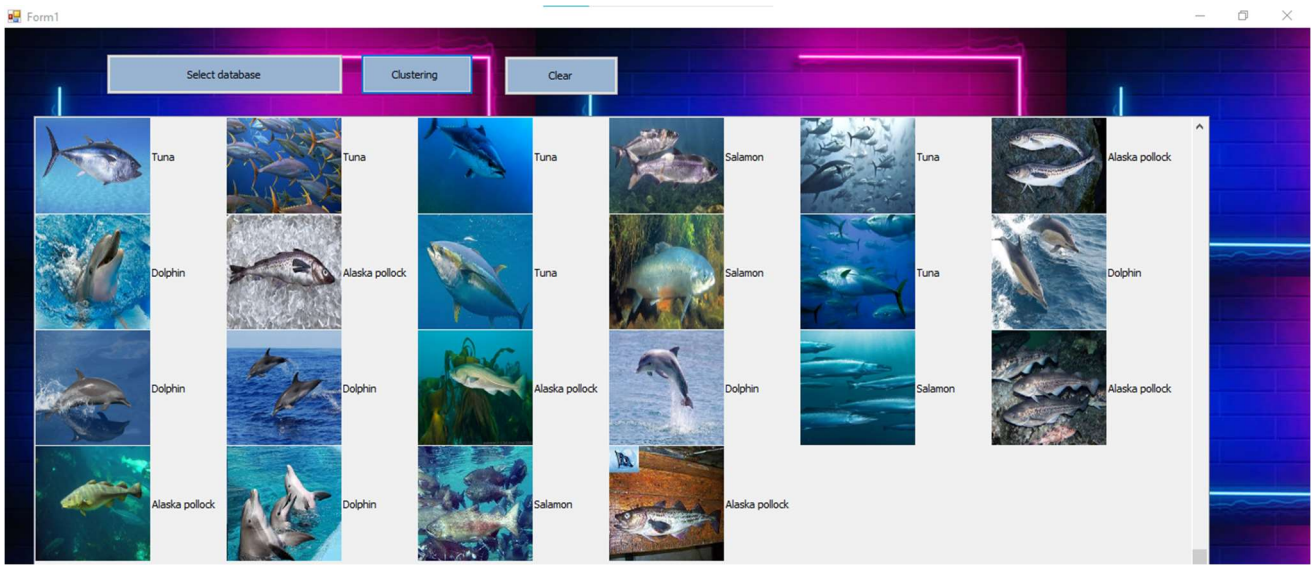


Figure 14. Aquatic Pets Animals images clustering

Figure 14 shows the results of a clustering that contains all images of Aquatic Pets animals.

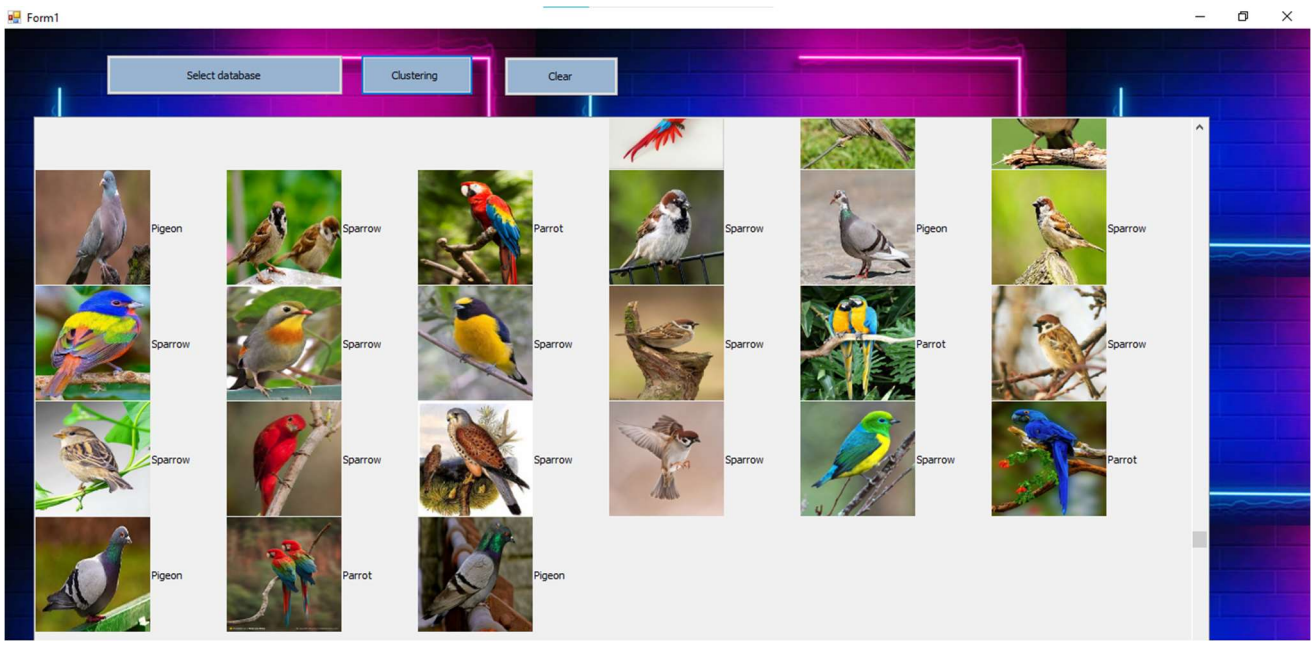


Figure 15. Aerial Pets Animals images clustering

Figure 15 shows the results of a clustering that contains all images of Aerial Pets animals.

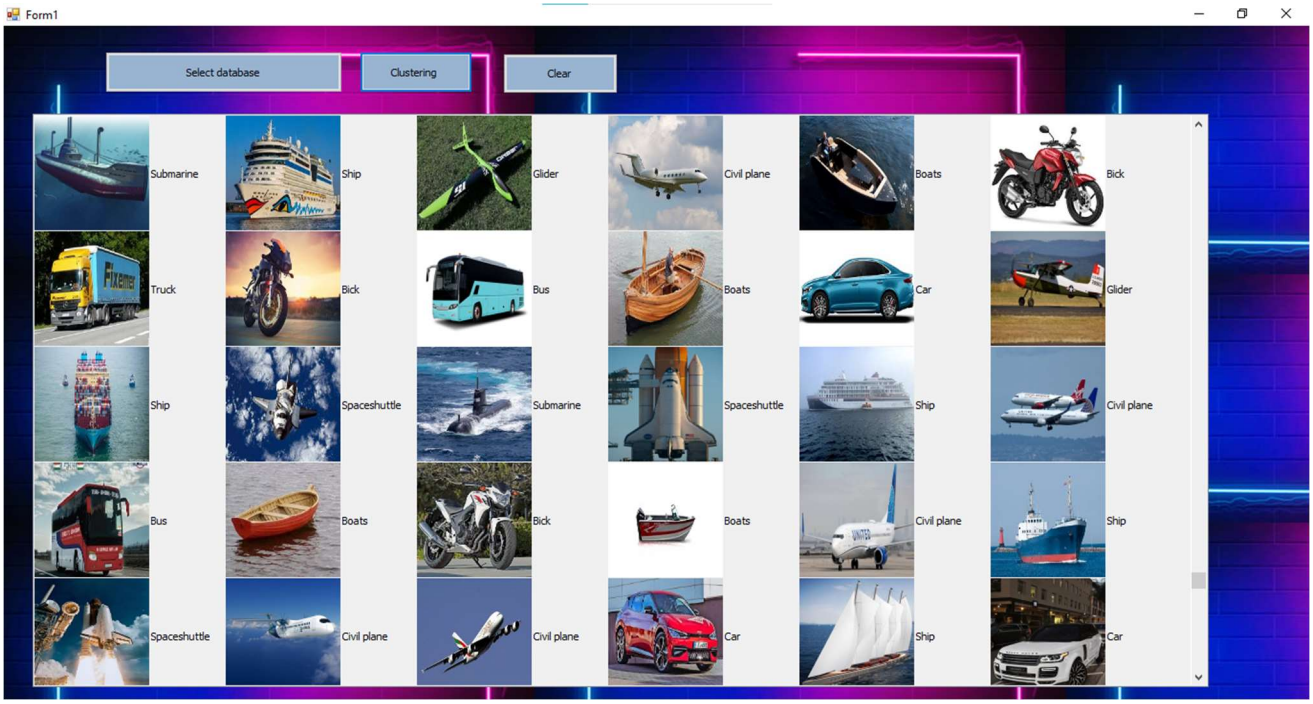


Figure 16. Transportation images clustering

Figure 16 shows the results of a clustering that contains all images of Transportation.

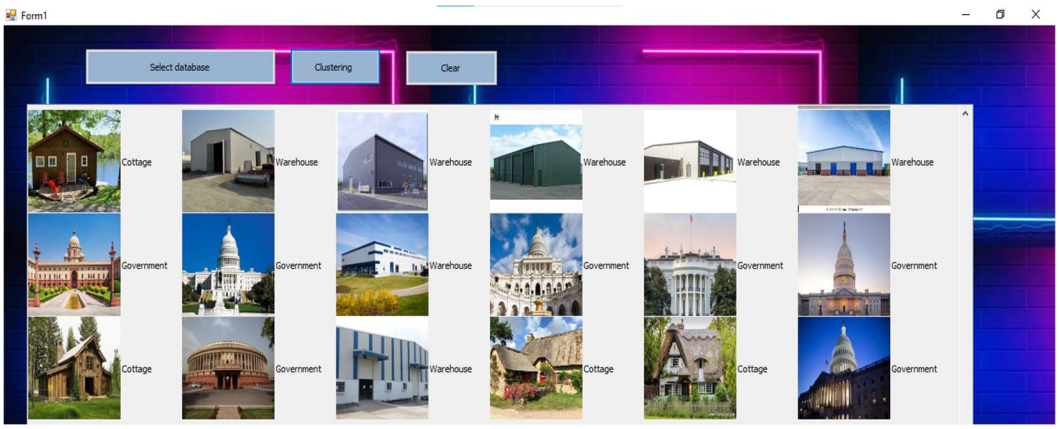


Figure 17. Building images clustering

Figure 17 shows the results of a clustering that contains all images of Building.

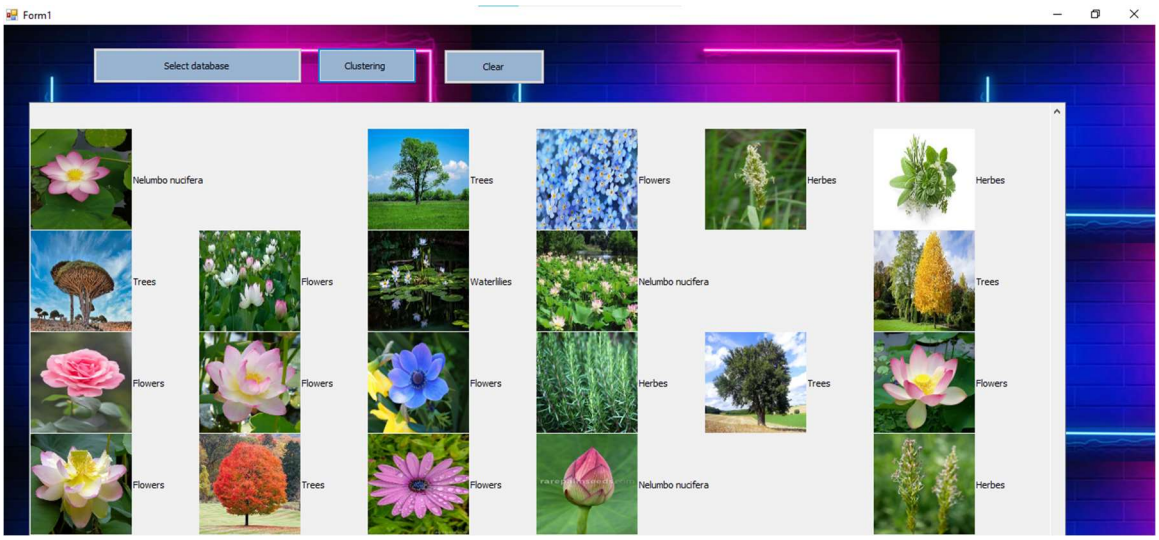


Figure 18. Plants images clustering

Figure 18 shows the results of a clustering that contains all images of Plants.

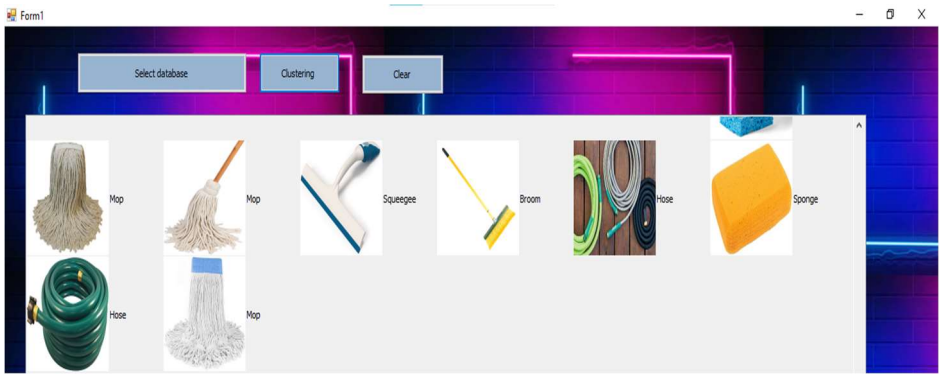


Figure 19. Cleaning images clustering

Figure 19 shows the results of a clustering that contains all images of Plants.

General conclusion

Image clustering is a challenging technique of clustering images from a large data image. In this thesis, we have focused on clustering images using WordNet's semantic similarity metrics based on the is-a relationship, according to information content or depth between concepts.

Experiments conducted in this thesis, concerning words in WordNet that were supported by human judgments in clustering images, showed better results. Lin's method, which relies on the information content IC of the images, outperformed Wu & Palmer, in the accuracy of the clustering with 10 %. Which makes it superior to the visual features, in the clustering of images.

A proposal for future work is that the approach that depends on the visual features (visual similarity) is developed to be an addition to the proposed approach (semantic similarity).

Reference:

1. Pattanaik, S. and D. Bhalke, *Beginners to content-based image retrieval.* International Journal of Science, Engineering and Technology Research, 2012. **1**: p. 40-44.
2. Singha, M. and K. Hemachandran, *Content based image retrieval using color and texture.* Signal & Image Processing, 2012. **3**(1): p. 39.
3. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems, 2012. **25**.
4. Hartigan, J.A. and M.A. Wong, *Algorithm AS 136: A k-means clustering algorithm.* Journal of the royal statistical society. series c (applied statistics), 1979. **28**(1): p. 100-108.
5. Cheng, Y., *Mean shift, mode seeking, and clustering.* IEEE transactions on pattern analysis and machine intelligence, 1995. **17**(8): p. 790-799.
6. Ester, M., et al. *A density-based algorithm for discovering clusters in large spatial databases with noise*. in *kdd*. 1996.
7. Russell, B.C., et al., *LabelMe: a database and web-based tool for image.* Int. J. of Computer Vision, 2005. **77**(1).
8. Sclaroff, S., et al., *Unifying textual and visual cues for content-based image retrieval on the world wide web.* Computer Vision and Image Understanding, 1999. **75**(1-2): p. 86-98.
9. Datta, R., et al., *Image Retrieval: Ideas, Influences, and Trends of the New Age-Addendum.* Proceedings MIR'05 (ACM), 2005.
10. da Silva Torres, R. and A.X. Falcao, *Content-based image retrieval: theory and applications.* RITA, 2006. **13**(2): p. 161-185.
11. Dai, B., Y. Zhang, and D. Lin. *Detecting visual relationships with deep relational networks*. in *Proceedings of the IEEE conference on computer vision and Pattern recognition*. 2017.
12. Budanitsky, A. and G. Hirst, *Evaluating wordnet-based measures of lexical semantic relatedness.* Computational linguistics, 2006. **32**(1): p. 13-47.
13. Sim, K.M. and P.T. Wong, *Toward agency and ontology for web-based information retrieval.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2004. **34**(3): p. 257-269.
14. Patwardhan, S., *Incorporating dictionary and corpus information into a context vector measure of semantic relatedness*. 2003, University of Minnesota, Duluth.
15. Lord, P.W., et al., *Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.* Bioinformatics, 2003. **19**(10): p. 1275-1283.
16. Al-Mubaid, H. and H.A. Nguyen. *A cluster-based approach for semantic similarity in the biomedical domain*. in *2006 international conference of the IEEE engineering in medicine and biology society*. 2006. IEEE.

17. M. Ehrig and J. Euzenat, "State of the Art on Ontology Alignment", Knowledge Web Deliverable D2.2.3, University of Karlsruhe, 2004.

18. H. Mihoubi, A. Simonet, and M. Simonet, "An Ontology Driven Approach to Ontology Translation", In Proceedings of DEXA, 2000, pp.573-582

19. Resnik, P., *Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language.* Journal of artificial intelligence research, 1999. **11**: p. 95-130.

20. Jiang, J.J. and D.W. Conrath, *Semantic similarity based on corpus statistics and lexical taxonomy.* arXiv preprint cmp-lg/9709008, 1997.

21. Leacock, C., *Filling in a sparse training space for word sense identification.* Ph. D. thesis, Macquarie University, 1994.

22. Lin, D. *Principle-based parsing without overgeneration*. in *31st annual meeting of the association for computational linguistics*. 1993.

23. Wu, Z. and M. Palmer, *Verb semantics and lexical selection.* arXiv preprint cmp-lg/9406033, 1994.

24. Fellbaum, C., *A semantic network of English verbs.* WordNet: An electronic lexical database, 1998. **3**: p. 153-178.

25. Resnik, P., *Using information content to evaluate semantic similarity in a taxonomy.* arXiv preprint cmp-lg/9511007, 1995.

26. Lesk, M. *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone*. in *Proceedings of the 5th annual international conference on Systems documentation*. 1986.

27. D. Fensel, "Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce", Springer, 2001

28. D. Tidwell, "Web Services-The Web's Next Revolution", IBM Web Service Tutorial, 29 Nov. 2000, http://www-106.ibm.com/developerworks/edu/ws-dwwsbasics-i.html.

29. Chihuahua or muffin, search for the best computer vision API https://www.freecodecamp.org/news/chihuahua-or-muffin-my-search-for-the-best-computer-vision-api-cbda4d6b425d/

30. A. Budanitsky and G. Hirst, "Evaluating WordNet-based measures of lexical semantic relatedness", Computational Linguistics, vol. 32, no. 1, (2006).

Appendix A

Wu and Palmer Similarity:

$$sin_{wup}(C1, C2) = \frac{2 * depth(LCS)}{depth(C1) + depth(C2)}$$

$depth(LCS) = N;\ depth(C1) = N1;\ depth(C2) = N2$

$Annex\ 1$:

$C1 = Lion,\ \ C2 = Sheep\ \Rightarrow\ N = 1;\ N1 = 4;\ N2 = 4$

$$Sim_{wup}(C1, C2) = \frac{2 * N}{N1 + N2} = \frac{2(1)}{4 + 4} = \frac{2}{8} = 0.5$$

$Annex\ 2$:

$C1 = Cat,\ \ C = dog\ \Rightarrow\ N = 3;\ N1 = 4;\ N2 = 4$

$$Sim_{wup}(C1, C2) = \frac{2 * N}{N1 + N2} = \frac{2(3)}{4 + 4} = \frac{6}{8} = 0.75$$

$Annex\ 3$:

$C1 = Car,\ \ C2 = Cat\ \Rightarrow\ N = 0;\ N1 = 3;\ N2 = 4$

$$Sim_{wup}(C1, C2) = \frac{2 * N}{N1 + N2} = \frac{2(0)}{3 + 4} = \frac{0}{7} = 0$$

Appendix B

Lin Similarity

$$Sim_{Lin}(C1, C2) = \frac{2 * IC\big(LCS(C1, C2)\big)}{IC(C1) + IC(C2)}$$

$$IC = -\log P(C)\ \Rightarrow\ P(C) = \frac{f(c)}{N}\ and\ f(c) = \sum_{n \in \mathcal{W}} count(n)$$

$Annex\ 1$:

$C1 = Lion,\ \ C2 = Sheep$

$$f(C1, C2) = \sum_{n \in \mathcal{W}} count(n) = 11, \qquad P(C) = \frac{f(c)}{N} = \frac{11}{45}$$

$$IC = -\log P(C1, C2) = -\log P\left(\frac{11}{45}\right);\ IC(C1) = IC(C2) = -\log\left(\frac{1}{45}\right)$$

$$Sim_{Lin}(C1, C2) = \frac{-2 * \log\left(\frac{11}{39}\right)}{-\log\left(\frac{1}{45}\right) - \log\left(\frac{1}{45}\right)} = 0.37$$

Annex 2:

$C1 = Cat,\ \ C2 = Dog$

$$f(C1, C2) = \sum_{n \in \mathcal{W}} count(n) = 7, \qquad P(C) = \frac{f(c)}{N} = \frac{7}{45}$$

$$IC = -\log P(C1, C2) = -\log P\left(\frac{7}{45}\right);\ IC(C1) = IC(C2) = -\log\left(\frac{1}{45}\right)$$

$$Sim_{Lin}(C1, C2) = \frac{-2 * \log\left(\frac{7}{39}\right)}{-\log\left(\frac{1}{45}\right) - \log\left(\frac{1}{45}\right)} = 0.49$$

Abstract

Image clustering is an interesting field in machine learning and computer vision, in which images are classified into a set of similar groups. Recently, with the explosive growth of the data in the smartphone and the web (Facebook, Instagram…), image clustering has even been a critical field to help the user quickly access the visual information he is looking for. Existing methods of image clustering only used either low-level visual feature, which constitutes a major obstacle to obtaining an accurate set of similar groups. To tackle this problem, we propose a novel algorithm that can cluster images based on the semantic similarity between surrounding texts (concept) of each image. In particular, we group images depending on the semantic similarity of their concepts instead of visual similarity. Conclusively, images are automatically clustered based on the label features. The performance of the cluster was compared based on accuracy. The highest accuracy was obtained by applying the method of Lin with 88.89%.

Keywords: Image clustering, Semantic similarity, Concepts, Ontology.

Résumé

Le regroupement d'images est un domaine intéressant de l'apprentissage automatique et de la vision par ordinateur, dans lequel les images sont classées en un ensemble de groupes similaires. Récemment, avec la croissance explosive des données dans le smartphone et le web (Facebook, Instagram…), le clustering d'images a même été un domaine critique pour aider l'utilisateur à accéder rapidement à l'information visuelle qu'il recherche. Les méthodes existantes de regroupement d'images n'utilisaient que l'une ou l'autre caractéristique visuelle de bas niveau, ce qui constitue un obstacle majeur à l'obtention d'un ensemble précis de groupes similaires. Pour résoudre ce problème, nous proposons un nouvel algorithme qui peut regrouper des images en fonction de la similarité sémantique entre les textes environnants (concept) de chaque image. En particulier, nous regroupons les images en fonction de la similarité sémantique de leurs concepts au lieu de la similarité visuelle. En conclusion, les images sont automatiquement regroupées en fonction des caractéristiques de l'étiquette. Les performances du cluster ont été comparées sur la base de la précision. La précision la plus élevée a été obtenue en appliquant la méthode de Lin avec 88,89 %.

Mots clés : Regroupement d'images, Sémantique similarité, Concepts, Ontologie.

ملخص

يعد تجميع الصور مجالًا مثيرًا للاهتمام في التعلم الآلي ورؤية الكمبيوتر، حيث يتم تصنيف الصور في مجموعة من المجموعات المتشابهة. في الآونة الأخيرة، كان النمو الهائل للبيانات في الهاتف الذكي والويب (فيس بوك وإنستغرام...)، وتجميع الصور مجالًا مهمًا لمساعدة المستخدم على الوصول بسرعة إلى المعلومات المرئية التي يبحث عنها. الأساليب الحالية في تجميع الصور تستخدم فقط ميزة بصرية منخفضة المستوى، مما يشكل عقبة رئيسية أمام الحصول على مجموعة دقيقة من المجموعات المتشابهة. لمعالجة هذه المشكلة، نقترح خوارزمية جديدة يمكنها تجميع الصور بناءً على التشابه الدلالي بين النصوص المحيطة (المفهوم) لكل صورة. على وجه الخصوص، نقوم بتجميع الصور اعتمادًا على التشابه الدلالي لمفاهيمها بدلاً من التشابه البصري. بشكل قاطع، يتم تجميع الصور تلقائيًا بناءً على ميزات التسمية. تم مقارنة أداء الكتلة على أساس الدقة. تم الحصول على أعلى دقة بتطبيق طريقة لين بنسبة 88.89٪.

الكلمات المفتاحية: تجميع الصور والتشابه الدلالي والمفاهيم وعلم الوجود.