Missouri S&T Library and Learning Resources

Scholars' Mine

Masters Theses

Student Theses and Dissertations

Fall 2015

A sentence-based image search engine

Weizhi Meng

Follow this and additional works at: https://scholarsmine.mst.edu/masters_theses

 Part of the Computer Engineering Commons

Department:

Recommended Citation

Meng, Weizhi, "A sentence-based image search engine" (2015). *Masters Theses*. 7474.
https://scholarsmine.mst.edu/masters_theses/7474

This thesis is brought to you by Scholars' Mine, a service of the Missouri S&T Library and Learning Resources. This work is protected by U. S. Copyright Law. Unauthorized use including reproduction for redistribution requires the permission of the copyright holder. For more information, please contact scholarsmine@mst.edu.

A SENTENCE-BASED IMAGE SEARCH ENGINE

by

WEIZHI MENG

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN COMPUTER ENGINEERING

2015

Approved by

Dr. Yiyu Shi, Advisor
Dr. Minsu Choi
Dr. Jun Fan

Copyright 2015

WEIZHI MENG

All Rights Reserved

ABSTRACT

Nowadays people are more interested in searching the relevant images directly through search engines like Google, Yahoo or Bing, these image search engines have dedicated extensive research effort to the problem of keyword-based image retrieval. However, the most widely used keyword-based image search engine Google is reported to have a precision of only 39% [1]. And all of these systems have limitation in creating sentence-based queries for images.

This thesis studies a practical image search scenario, where many people feel annoyed by using only keywords to find images for their ideas of speech or presentation through trial and error. This thesis proposes and realizes a sentence-based image search engine (SISE) that offers the option of querying images by sentence. Users can naturally create sentence-based queries simply by inputting one or several sentences to retrieve a list of images that match their ideas well.

The SISE relies on automatic concept detection and tagging techniques to provide support for searching visual content using sentence-based queries. The SISE gathered thousands of input sentences from TED talk, covering many areas like science, economy, politics, education and so on. The comprehensive evaluation of this system was focused on usability (perceived image usefulness) aspect. The final comprehensive precision has been reached 60.7%. The SISE is found to be able to retrieve matching images for a wide variety of topics, across different areas, and provide subjectively more useful results than keyword-based image search engines.

ACKNOWLEDGMENTS

I would like to take this opportunity to thank the following people who have directly or indirectly helped me in academic achievements. Firstly, I would like to thank my Master advisor, Professor Yiyu Shi, for his continuous support and guidance throughout my master's program in computer engineering. I sincerely thank Professor Minsu Choi and Professor Jun Fan for accepting to be a part of my thesis committee and making time for me off their busy schedule.

Also, I would like to thank my family members, especially my wife, who give me lots of support and encouragement. Special thanks to the best gifts in my life, my two daughters, who filled my life up with love and warm. Finally, I would like to thank all my friends in Rolla, they are so friendly and helpful.

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

Table Page

1. INTRODUCTION

Internet has witnessed a great success of social media websites. It increases the number of digital images in the websites. Nowadays people are more interested in searching the relevant images directly through search engines. The most common search engines today offer image search such as Google, Yahoo or Bing. Automatically finding images relevant to a textual query remains a very challenging task. Google image search engine is reported to have a precision of only 39% [1].This thesis proposes and studies a practical scenario, where people do presentation as in Figure 1.1, they always felt troubled by finding images related to their speeches, they hope the screen can show the image related to his speech automatically. The Sentence-based Image

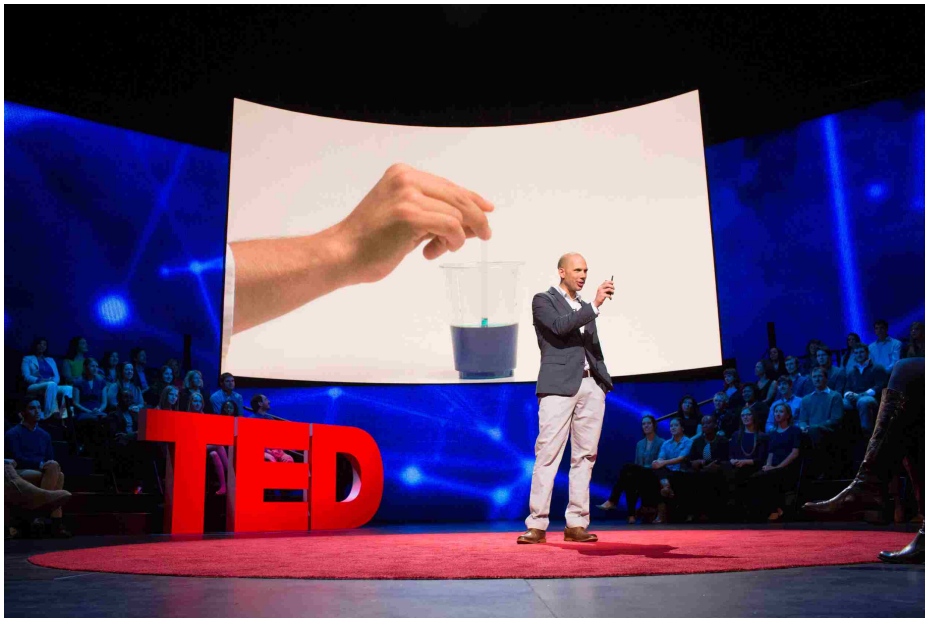

Figure 1.1. A TED speaker was doing presentation

Search Engine (SISE) provided a system to search for such meaningful images that are suitable for sentence. Users can naturally create sentence-based queries by simply typing sentence of the speech to retrieve a list of images that match their ideas well.

The keyword-based search process used by common image search engines, however, can be especially challenging for inexperienced searchers. Studies have shown that keyword-based queries significantly limit the expressiveness of users and, therefore, degrade the effectiveness of search [2]. As a consequence, it may take users a considerable amount of time and effort to discover the right set of keywords through a trial-and error process.

Given the limitation of keyword queries, one way to overcome this is to allow users to use sentence as the target interface to do text query. The SISE system can potentially provide the following benefit: It offers users a faster and more intuitive method to describe an interface by simply input a sentence rather than thinking of many keywords. The SISE system is a practical case where sentence-based search is advantageous for searching full text, Figure 1.2 and Figure 1.3 provide some examples of SISE system practically retrieve an image visually relevant to the sentence user had inputted.

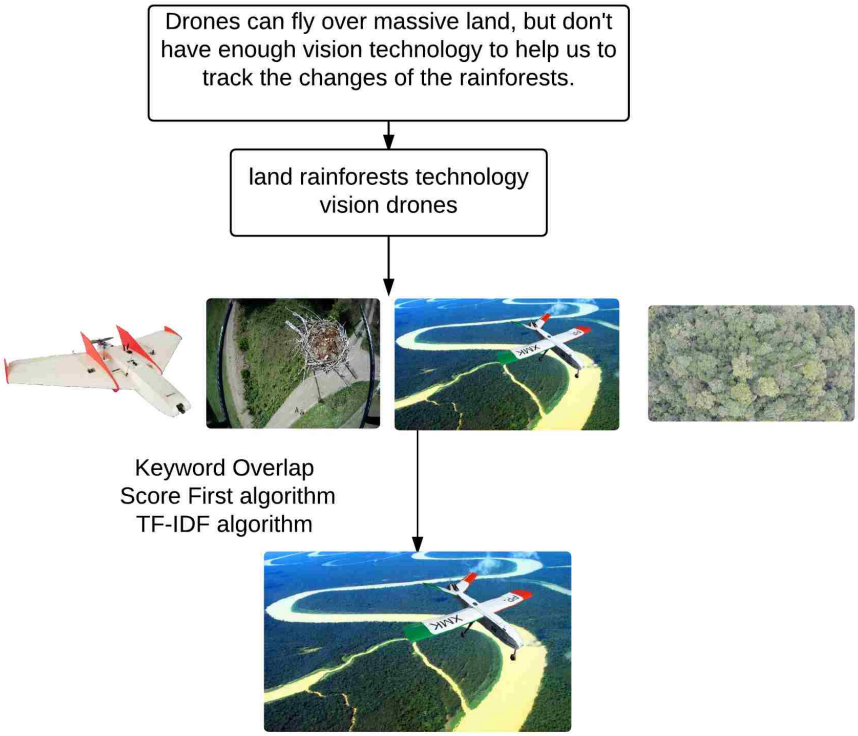

Figure 1.2. The first example of SISE retrieve image relevant to sentence

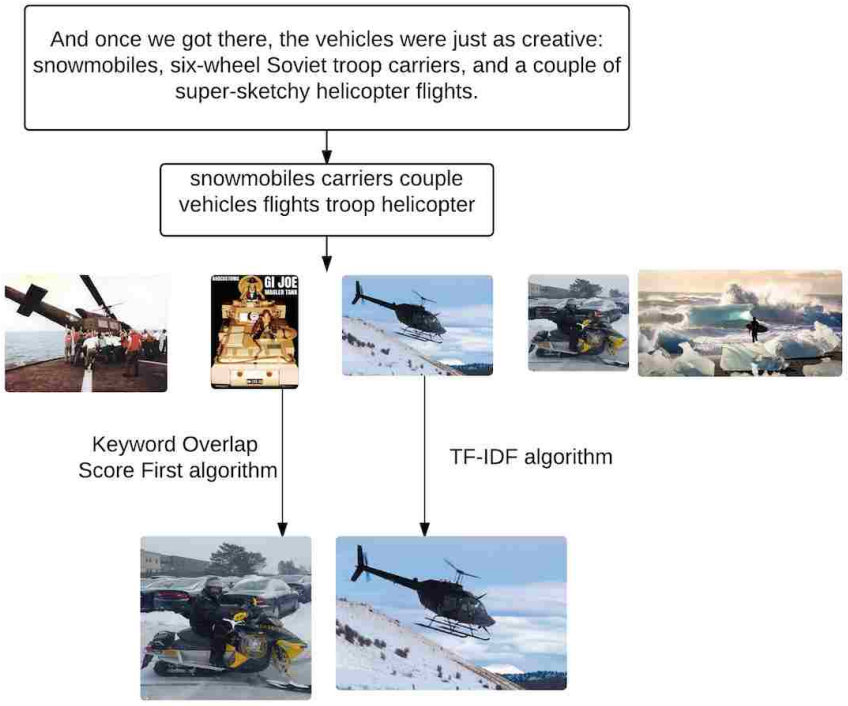

Figure 1.3. The second example of SISE retrieve image relevant to sentence

2. SENTENCE-BASED IMAGE SEARCH ENGINE MODEL

As in Figure 2.1, the SISE concludes four modules: Text Extraction Module, Image Retrieval Module, Tag Retrieval Module and Tag Ranking Module. In this Model, the input is one or several sentences, output is a list of matching images for sentence. In Text Extraction Module, several keywords will be extracted, and then Image Retrieval Module will gather the images based on these keywords from Internet. While the images are downloaded, Tag Retrieval Module retrieves tags for each image at the same time. Finally, several algorithms will be used to choose the best suitable images based on image tags.

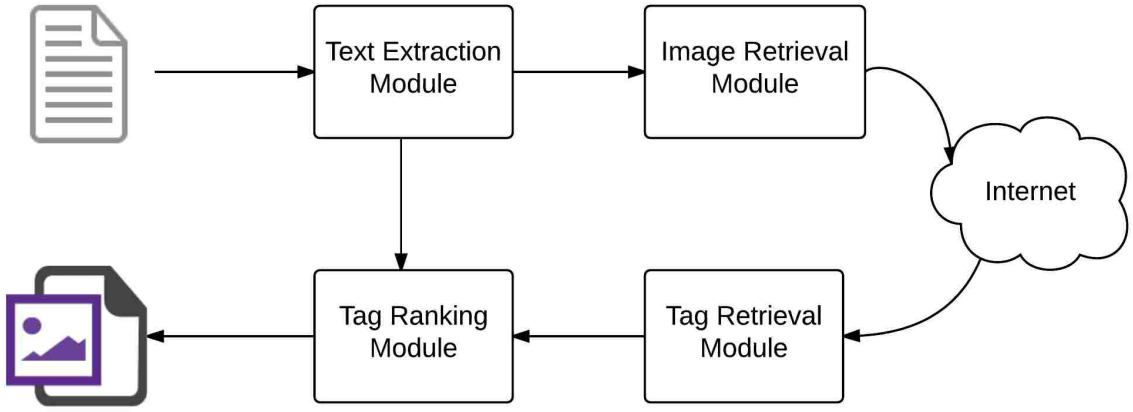

Figure 2.1. Sentence-Based Image Search Engine Architecture

The SISE relies on automatic concept detection and tagging techniques to provide support for searching images using sentence-based queries. Automatic concept detection realized in Text Extraction Module is based on Natural Language Processing (NLP) and WordNet methods, tagging techniques realized in Tag Retrieval Module is based on Computer Vision (CV) methods. In Image Retrieval Module, a web crawler was designed to fit intuitively to Google (keyword-based) image search engine to download images. In Tag Ranking Module, this thesis proposes three algorithms such as Keyword Overlap, Score First and TF-IDF to make comparison and use them to find most suitable images for sentence.

3. SENTENCE-BASED IMAGE SEARCH IMPLEMENTATION

3.1. TEXT EXTRACTION MODULE IMPLEMENTATION

3.1.1. Text Extraction Module Architecture. Nouns often function as verb subjects and objects, as predicative expressions, and as the complements of prepositions. In both Tag-Based Image Search and Content-Based Image Search, nouns are most important part of a sentence to represent the main idea. The images are always classified by nouns as the keywords, when searching on the database, using nouns are efficient to do query and find matching results. In Figure 3.1 the Text Extraction Module extract nouns from sentence using in a series of processes as Natural Language Processing (NLP) Noun Extraction, Stop Words Method and Repeat Words Removal. The simplified nouns will be used to retrieve images in Image Retrieval Module. After retrieving enough images, the selection of images should be

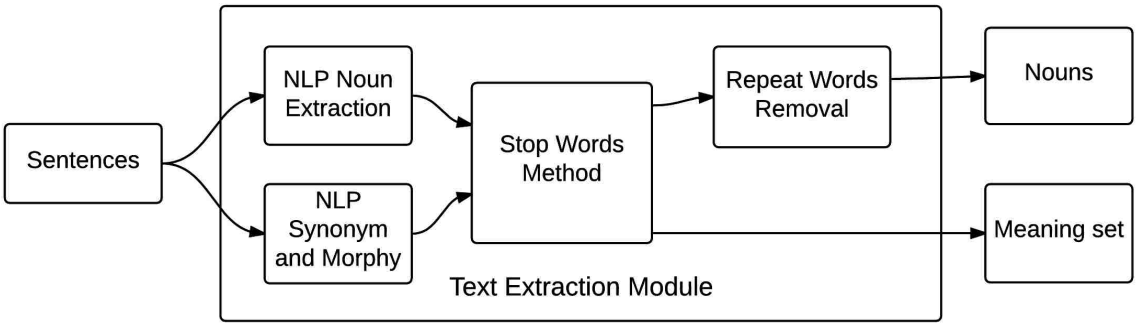

Figure 3.1. Text Extraction Module Architecture

based on good keywords summation of the image, parsing the sentence to get main words of text is a good choice to be part of the meaning set, but it is not enough, it also needs to be processed in a series of NLP Synonym and Morphy Methods and Stop Words Method. Figure 3.2 shows an example of text extraction in SISE, the

meaning set will be used to find the most related image in Tag Ranking Module, it will be discussed later.

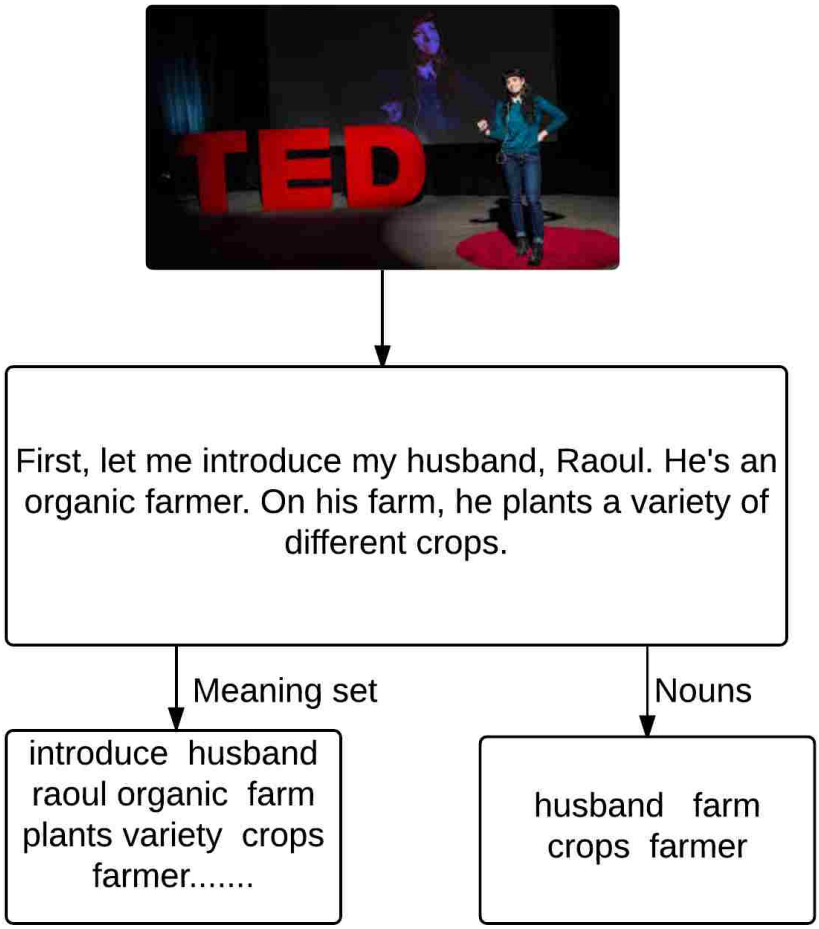

Figure 3.2. An example of text extraction in SISE

3.1.2. NLP Noun Extraction. Natural Language Processing (NLP) is a field at the intersection of computer science, linguistics and artificial intelligence, which aims to make the underlying structure of language available to computer programs for analysis and manipulation. A Part-Of-Speech Tagger (POS Tagger) [3, 4] is a method of NLP that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., as in Figure 3.3, this is an example of parsing sentence using POS Tagger.

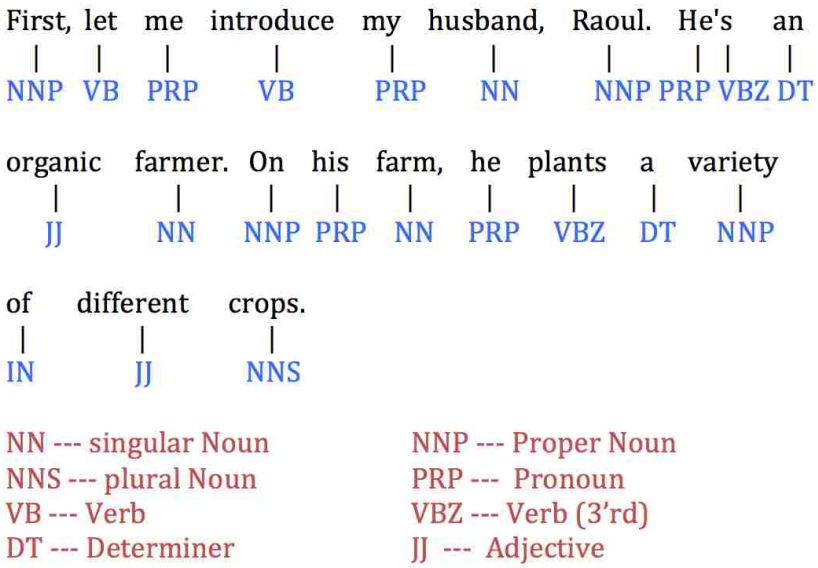

Figure 3.3. An example of parsing sentence using Part-Of-Speech Tagger (POS Tagger)

3.1.3. Noun Extraction Architecture. NLP Noun Extraction part uses Information Extraction System to do nouns extraction, which was built on POS Tagger Method. Figure 3.4 shows the architecture for a simple information extraction system. First, the raw text of the document is split into sentences using a sentence

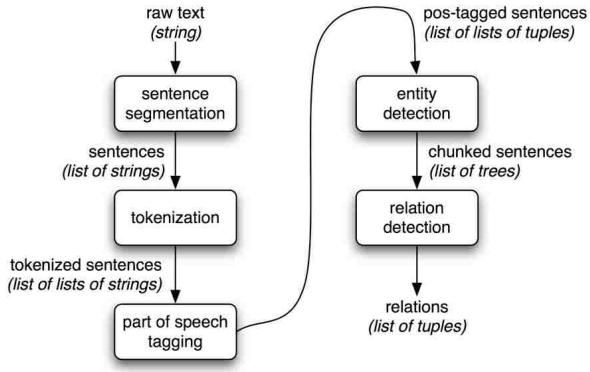

Figure 3.4. Simple Pipeline Architecture for an Information Extraction System

segmenter, and each sentence is further subdivided into words using a tokenizer. Next, each sentence is tagged with POS tags, which will prove very helpful in the

next step, named entity detection. In this step, mentions of potentially interesting entities in each sentence was searched. Finally, the likely relation between different entities in the text was detected using entity detection. The basic technique for entity detection is chunking, which segments and labels multi-token sequences as illustrated in Figure 3.5. The smaller boxes show the word-level tokenization and POS tagging, while the large boxes show higher-level chunking. Each of these larger boxes is called a chunk.

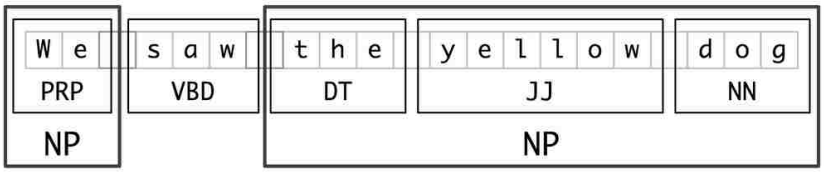

Figure 3.5. Segmentation and Labeling at both the Token and Chunk Levels

One of the most useful sources of information for NP-chunking is POS tag. This is one of the motivations for performing POS tagging in SISE. This approach is demonstrated using an example sentence that has been part-of-speech tagged in Figure 3.6. A chunk grammar is used to create an NP-chunker, it consists of rules that indicate how sentences should be chunked.

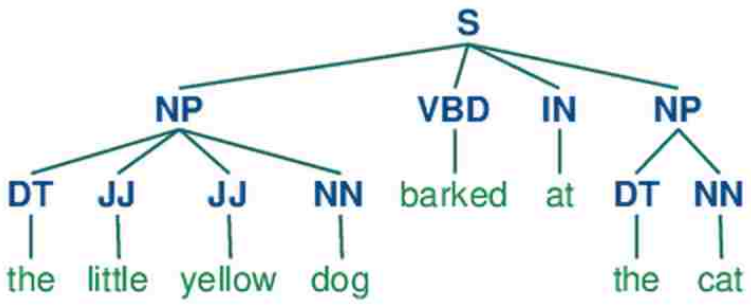

Figure 3.6. Example of a Simple Regular Expression Based NP Chunker

3.1.4. Stop Words Method. Sometimes, some extremely common words appear to be of little value in helping select documents matching a user need, are

excluded from the vocabulary entirely. These words are called stop words. The general strategy for determining stop words is to sort the terms by collection frequency (the total number of times each term appears in the document collection), and then to take the most frequent terms, often hand-filtered for their semantic content relative to the domain of the documents being indexed, as a stop word, the members of which are then discarded during indexing. In computing, stop words are words which are filtered out before or after processing of natural language data (text).[5]

An example of a stop list is shown in Figure 3.7. Using a stop list significantly reduces the number of postings that a system has to store.

a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will	with					

Figure 3.7. A stop list of 25 semantically non-selective words

Though stop words usually refer to the most common words in a language, there is no single universal list of stop words used by all processing of natural language tools, and indeed not all tools even use such a list. Some tools specifically avoid removing these stop words to support phrase search. Stop Words is a main part of Text Extraction Module, removing stop words from sentences will largely reduce the number of keywords to search images. The SISE gathers stop words from multiple stop words database as in Figure 3.8 showed, after gathering from English stop words list, Long stop word List, MySQL stop words and Google History, the stop words database in SISE concludes almost 1000 stop words now, and it will grow in the future.

3.1.5. NLP Synonym and Morphy. As in Figure 3.9, if the keyword in sentence is “motor vehicle”, the meaning set should concludes keywords like “motorcar”, “truck” and etc.., SISE used NLP Synonym method from WordNet to do this work. Besides, if the keyword is like “gas guzzler”, the meaning set should have

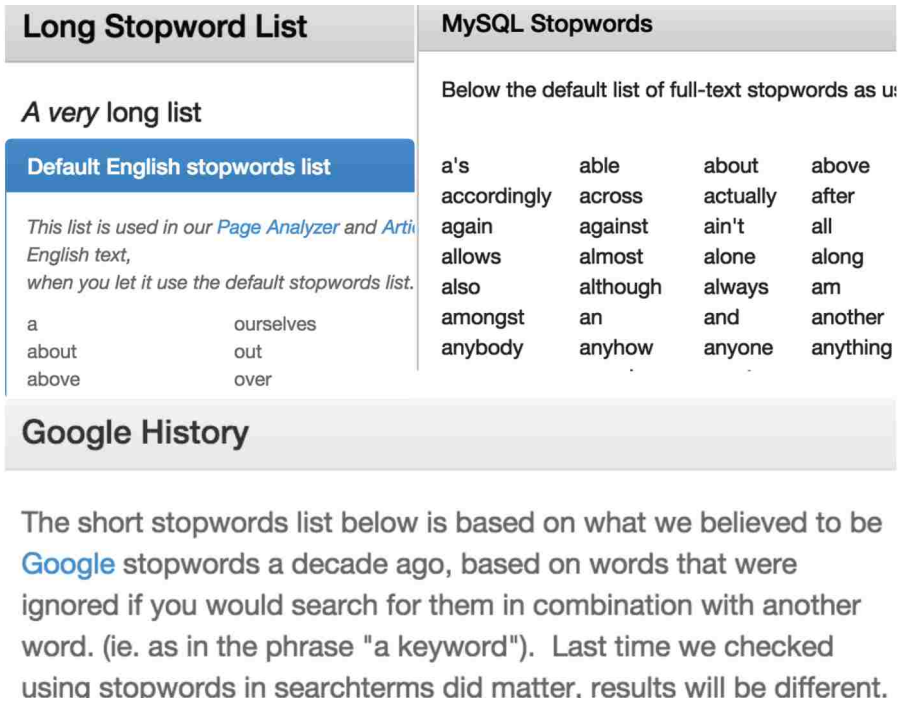

Figure 3.8. Gathering stop words from multiple database

keyword like “motorcar” to avoid missing some other useful image tags, SISE uses NLP Morphy method from WordNet to achieve this goal.

WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.

Synonymy is one of the lexical semantic relations (LSRs), which are the relations between meanings of words. By definition, synonyms are one of two or more words or expressions of the same language that have the same or nearly the same meaning in some or all senses. For an image, the meaning may be expressed in a different way, in information extraction, it is useful to know if two word have the same or very similar semantic content. Words that denote the same concept and are interchangeable in many contexts–are grouped into unordered sets (synsets). The main relation among words in WordNet is synonymy, the majority of the WordNet’s relations connect words from the same part of speech (POS). Thus, WordNet really

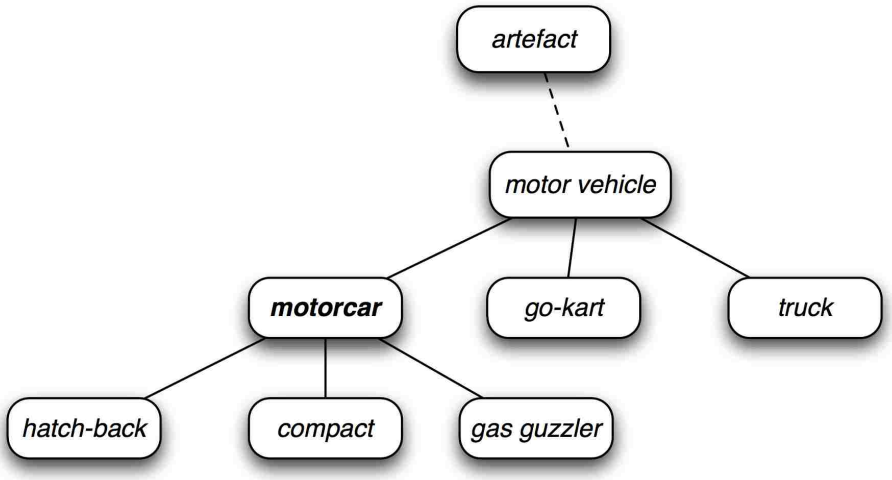

Figure 3.9. An example of synonym and morphy of "motor vehicle"

consists of four sub-nets, one each for nouns, verbs, adjectives and adverbs, with few cross-POS pointers.

Morphy is a morphological processor native to WordNet. The WordNet interfaces invoke Morphy to lemmatize a word as part of the lookup process (e.g. you query "enlightened", it returns the results for both "enlightened" and, via Morphy, "enlighten"). "nltk morphy" is a lemmatizer (a stemmer with principles). It enables you to reduce words to their root form in English, using the Morphy algorithm that is built into WordNet, together with NLTK's POS.

3.1.6. Repeat Words Removal. In Text Extraction Module, nouns are used as group of keywords to search images, so the repeat keywords should be removed. However, meaning set is used as representation of the meaning of query sentences, the more times a word occurred, the importance of that word increased. Counting the number of word is used in Tag Ranking Module to calculate the term frequency (TF) of keyword. This will be discussed later.

3.2. IMAGE RETRIEVAL MODULE IMPLEMENTATION

In keyword-based image search engine, images are richly illustrated by tags. Image queries in the form of sentences ensure the visual relevance to the target interface, whereas queries in the form of keywords ensure the textual relevance to the pertinent computing tasks. So Image Retrieval Module in SISE is based on modern image search engine that was designed to help to find images on the Internet. As in Figure 3.10, SISE is based on Google image search engine, which is one of the most powerful and popular image search engines now.

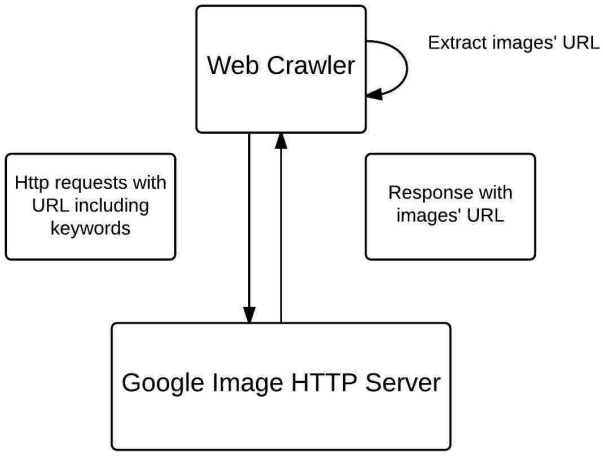

Figure 3.10. Web crawler for Image Retrieval Module

3.2.1. Google Images. Google Images is a search service owned by Google and introduced in July 2001. The keywords for the image search are based on the filename of the image, the link text pointing to the image, and text adjacent to the image. When searching for an image, a thumbnail of each matching image is displayed. When the user clicks on a thumbnail, the image is displayed in a box over the website that it came from. The user can then close the box and browse the website, or view the full-sized image. This section describes the methods for downloading the initial pool of images (together with associated meta-data) from the Internet, and the initial filtering that is applied.

3.2.2. Resent Image Search Method. In recent years there has been considerable interest in learning from the images and associated text that can be found on the web. Some authors have focused on images and their associated tags on photo sharing websites like Flickr, see e.g. [6, 7], while others have focused on general web images gathered using existing text or image search engines [8, 9, 10, 11]. Most of these methods rely on visual consistency to identify images that are relevant to the query terms, among a set of several hundreds to thousands of images obtained using the search engine.

Generative approaches learn a model on the images obtained from the search engine and then rank them by the likelihood of the images under the model. Images may be indexed or categorized based on visual features, terms and key-terms, assigned subjects, or image types [12]. The text gathered may be the image file name, captions, web page titles, and other text near the image tags. Annotating images for indexing is quite demanding. An alternative is to use image properties that are less likely to require intervention.

3.2.3. How Image Search Engine Work. A common misunderstanding when it comes to image search is that the technology is based on detecting information in the image itself. But most Image Search Engines work like this, the metadata of the image is indexed and stored in a large database and when a search query is performed, the image search engine looks up the index, and queries are matched with the stored information. The results are presented in order of relevancy. The usefulness of an image search engine depends on the relevance of the results it returns, and the ranking algorithms are one of the keys to becoming a big player.

3.2.4. Image Processing Technique. The search engines use the image processing techniques for finding the images from the World Wide Web. Image processing is any form of signal processing for which the input is an image, such as a photograph, the output may be either an image or a set of characteristics or

parameters related to the image. The purpose of image processing is visualization, image sharpening and restoration, image retrieval, measurement of pattern and image recognition. Image processing is classified into analog and digital image processing. Analog image processing is conducted on two-dimensional signals by means of analog input and output. For this type, the analyst must apply a combination of personal knowledge and collateral data to image processing. Digital image processing is the use of computer algorithms to perform image processing on digital images. As a subcategory or field of digital signal processing, digital image processing has many advantages over analog image processing. It allows a much wider range of algorithms to be applied to the input data. Feature is an interesting part of an image, and features are used as a starting point for many computer vision algorithm. Feature detection is a low-level image processing operation.

Feature extraction is a special form of dimensionality reduction and transforming the input data into the set of features. Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data. Feature learning or representation learning is a set of techniques in machine learning that learn a transformation of "raw" inputs to a representation that can be effectively exploited in a supervised learning task such as classification. An image retrieval system is a computer system for browsing, searching and retrieving images from a large database of digital images. Most traditional and common methods of image retrieval utilize some method of adding metadata such as captioning, keywords, or descriptions to the images so that retrieval can be performed over the annotation words. Manual image annotation is time consuming, laborious and expensive; to address this, there has been a large amount of research done on automatic image annotation. Additionally, the increase in social web applications and the semantic web have inspired the development of several web-based image annotation tools. Automatic image annotation (also known as automatic image tagging or linguistic indexing) is the process by

which a computer system automatically assigns metadata in the form of captioning or keywords to a digital image.

3.2.5. Automatic Image Annotation. Numerous algorithms have been proposed for automatic image annotation [13]. They can roughly be grouped into two major categories, depending on the type of image representations used. The first group of approaches are based upon global image features [14], such as color moment, texture histogram, etc. The second group of approaches adopts the local visual features. [15, 16] segment image into multiple regions, and represent each region by a vector of visual features. Approaches [17, 18] extend the bag-of-features or bag-of-words representation, which was originally developed for object recognition, for automatic image annotation. More recent work [19] improves the performance of automatic image annotation by taking into account the spatial dependence among visual features. Other than predicting annotated keywords for the entire image, several algorithms [20] have been developed to predict annotations for individual regions within an image. Despite these developments, the performance of automatic image annotation is far from being satisfactory. The text-based approaches use the associate text to derive the content of image. Image file names, anchor texts, surrounding paragraphs, even the whole text of the hosting web page are examples of textual content that is often used in such systems.

3.2.6. An Example of Image Search Engine Architecture. The general architecture of the system is depicted in Figure 3.11. The system consists of 3 main parts: the segmentation module (Part I), the clustering (Part II) and the keyword extraction module (Part III).

The image search engines can automatically identify a limited range of visual content, e.g. faces, trees, sky, buildings, flowers, colors etc. This can be used alone, as in content-based image retrieval, or to augment metadata in an image search. Besides the Visual Segmentation Module, Images and Textual Blocks can also gather some

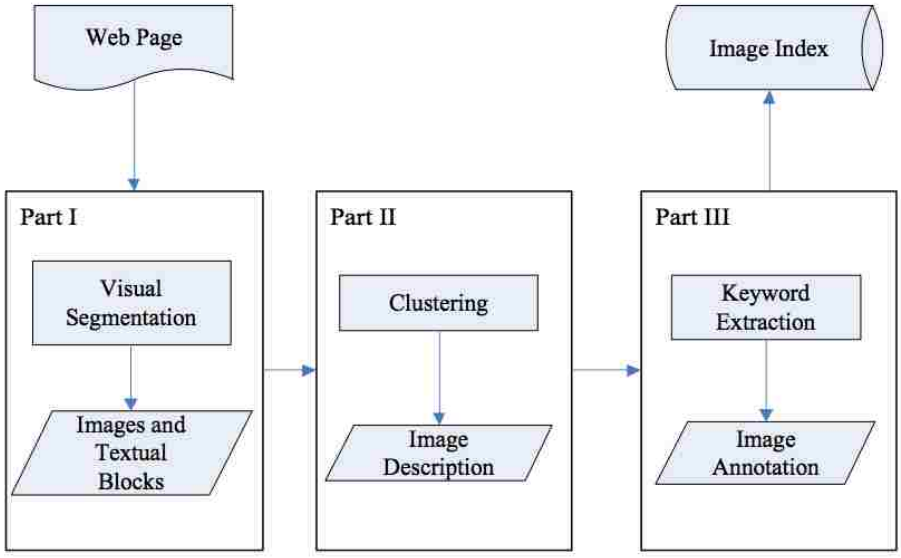

Figure 3.11. An example of Image Search Engine system architecture

labels for images, each image has a ground-truth relevance label, indicating whether or not it is relevant to the query.

1) The content extraction of each web image is based on textual information that exists in the same web document and refers to this image. Initially both image and text blocks must be identified. In order to obtain the set of visual segments that form a web page, the Visual Based Page Segmentation (VIPS) algorithm [21] is widely used. The VIPS algorithm extracts the semantic structure of a web page based on its visual representation. It attempts to make full use of the page layout structure by extracting blocks from the DOM tree structure of the web page and locating separators among these blocks. Therefore, a web page is represented as a set of blocks that bare similar Degree of Coherence (DOC). With the permitted DOC (pDOC) set to its maximum value, it obtains a set of visual blocks that consist of visually indivisible contents.

2) For each visual block, obtained in the previous step, the VIPS algorithm returns the two-dimensional Cartesian coordinates of its location in the web page.

The HTML source code that corresponds to each one of these blocks is used in order to classify them into two categories: (i) image blocks, and (ii) text blocks. The objective of the second module of the proposed system is to assign each text block to an image block.

3) When performing a search the user receives a set of thumbnail images, sorted by relevancy. Each thumbnail is a link back to the original web site where that image is located. Using an advanced search option the user can typically adjust the search criteria to fit their own needs, choosing to search only images or animations, color or black and white, and setting preferences on image size.

3.3. TAG RETRIEVAL MODULE IMPLEMENTATION

Tag Retrieval Module is an important component of image search engine. However, in databases such as Flickr or Facebook, large fraction (over 50% in Flickr) of images have no tags at all and are hence never retrieved for text queries.

3.3.1. Tag Retrieval Method. Retrieving tags from images is a difficult machine learning task, different type of objects require different image descriptors, Convolutional neural networks are often used in image recognition systems, Figure 3.12 has shown some examples of retrieving tags from images .

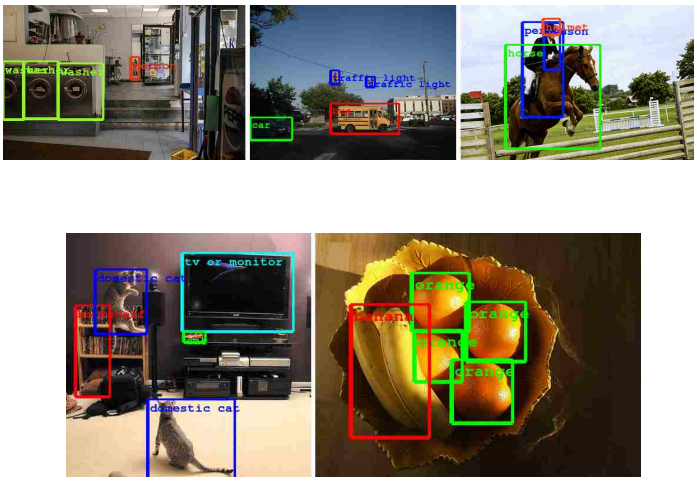

Figure 3.12. Tag Retrieval example

They have achieved an error rate of 0.23 percent on the MNIST database, which as of February 2012 is the lowest achieved on the database. [22] Another paper on using CNN for image classification reported that the learning process was fast.[23]

3.3.2. Convolutional Neural Network (CNN). Convolutional Neural Networks (CNN) are very similar to ordinary Neural Networks(NN): They are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. The whole network still express a single differentiable score function: From the raw image pixels on one end to class scores at the other. And they still have a loss function (e.g. SVM) on the last (fully-connected) layer and all the tips/tricks we developed for learning regular Neural Networks still apply. The difference between CNN and NN is that the inputs are images, which allows us to encode certain properties into the network. These then make the forward function more efficient to implement and vastly reduces the amount of parameters in the network.

As in Figure 3.13, in regular NN architecture, it receives an input (a single vector), and transforms it through a series of hidden layers. Each hidden layer is

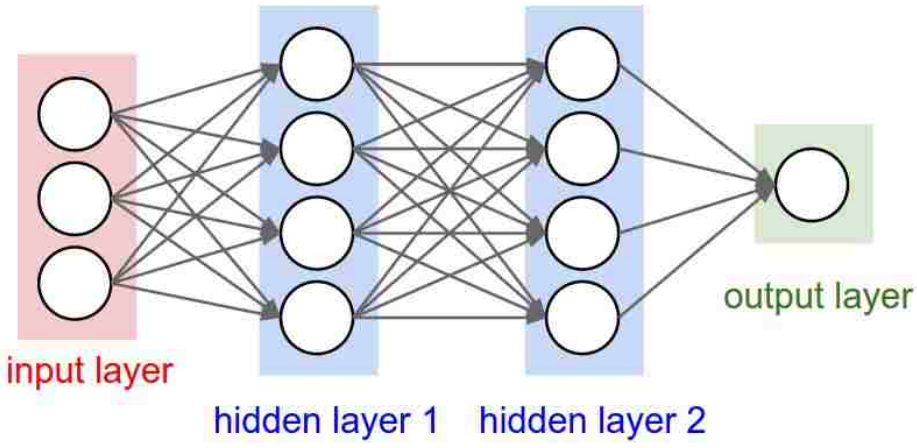

Figure 3.13. Regular Neural Network(NN) architecture

made up of a set of neurons, where each neuron is fully connected to all neurons in the previous layer, and where neurons in a single layer function completely independently

and do not share any connections. The last fully connected layer is called the "output layer" and in classification settings it represents the class scores.

In CNN architecture, the layers of a CNN have neurons arranged in 3 dimensions: width, height and depth. Every layer of a CNN transforms the 3D input volume to a 3D output volume of neuron activations. In Figure 3.14, the red input layer holds the image, so its width and height would be the dimensions of the image, and the depth would be 3 (Red, Green, Blue channels).

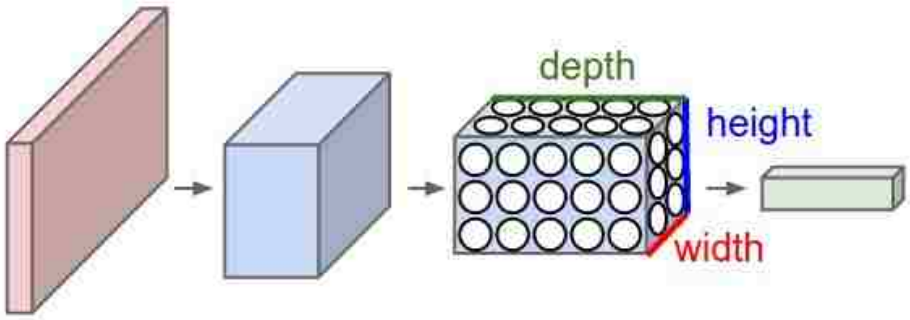

Figure 3.14. Convolutional Neural Network(CNN) architecture

3.3.3. Convolutional Layer. The Convolutional layer is the core building block of a CNN, and its output volume can be interpreted as holding neurons arranged in a 3D volume. As in Figure 3.15, is an example of layers in CNN, the initial volume stores the raw image pixels and the last volume stores the class scores. Each volume of activations along the processing path is shown as a column. Since it's difficult to visualize 3D volumes, we lay out each volume's slices in rows. The last layer volume holds the scores for each class, but here we only visualize the sorted top 5 scores, and print the labels of each one.

3.3.4. GPU. With the rise of efficient GPU computing, it has become possible to train larger networks. Several improvements provided more efficient ways to train convolutional neural networks with more layers.

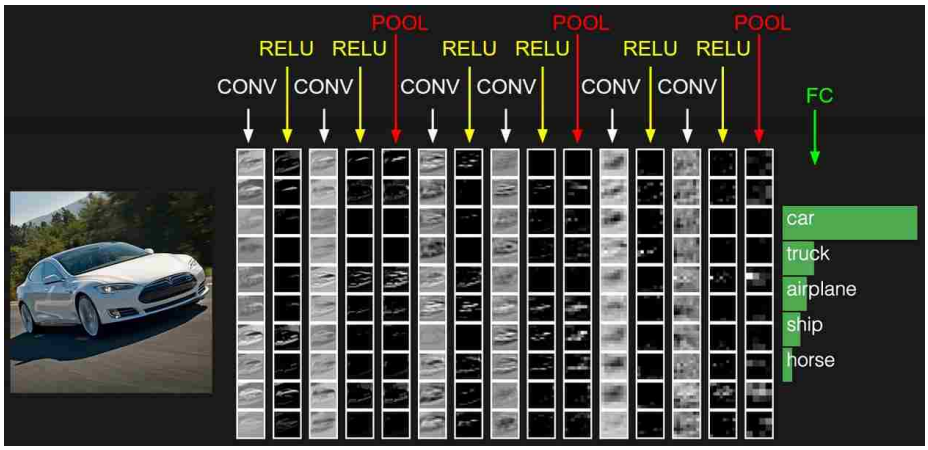

Figure 3.15. An example of layers in CNN

3.4. TAG RANKING MODULE IMPLEMENTATION

To choose the most related images for sentence, the information in tags retrieved by Tag Retrieving Module is required. The data in tags is a group of keywords and scores, as a part of sentence, the group of keywords retrieved by Text Extraction Module can represent the meaning of sentence well. Based on keywords from sentence and tags, the first algorithm SISE used is Keyword Overlap algorithm, which finds the maximum overlapping keywords. The second algorithm ranks images using scores with keywords. The third algorithm is TF-IDF, which will be discussed latter.

3.4.1. Keywords Overlap Algorithm. The Keyword Overlap algorithm is a very basic retrieval algorithm. The algorithm simply returns the image that has most keywords overlapped, the keywords overlapped represent the intersection of the set of words in the tags and the set of words in the query. For example, if the query is "cat dog horse" and the keywords in tags of two images are "cat horse tram carriage" and "cat tram carriage", it would return the first image with tag "cat horse tram carriage", because the number of keywords overlapped is 2.

3.4.2. Score First Algorithm. Score First Algorithm is similar to Keywords Overlap Algorithm, besides count the number of keywords in the intersection

of query and tags, it focus more on the percentage of similarity. For the example above, if the query is "cat dog horse" and the keywords in tags of two images are "cat horse tram carriage" and "cat tram carriage", but the score of "cat" in second image is 90 while the sum of scores of "cat" and "horse" is only 80, it would return the second image with tag "cat tram carriage", because the total score is higher.

3.4.3. TF-IDF Algorithm. Terms Frequency and Inverse Document Frequency (TF-IDF) value increases proportionally to the number of times a keyword appears in the tag of image, but is offset by frequency of the keyword in other tags of images, which helps to control the fact that some keywords are generally more common than others to show outstanding feature.

TF-IDF is used for text matching [24]. It is frequently used as a weighting factor in information retrieval and text mining.

TF-IDF stands for term frequency-inverse document frequency, and TF-IDF weight is often used in information retrieval and text mining. The weight is a statistical measure used to evaluate importance of word to document in a collection or corpus. Frequency of a word appears in document as offset in corpus. TF-IDF implementation is incorporated to improve keywords filtering for screening high-level categories. TF-IDF can be successfully used for text filtering in categories subject to keywords that does text summarization and classification. In Figure 3.16, we have shown formulas that we have used.

According to the Keyword Overlap algorithm, in image tag t_k, the frequency $f(k_i)$ represent the times keyword k_i has been occurred in query sentence q_j, and the image set totally has n images. Then, the average length of image tags would be

$$avg = \frac{\sum_{1 \leq h \leq n} |t_h|}{n} \tag{1}$$

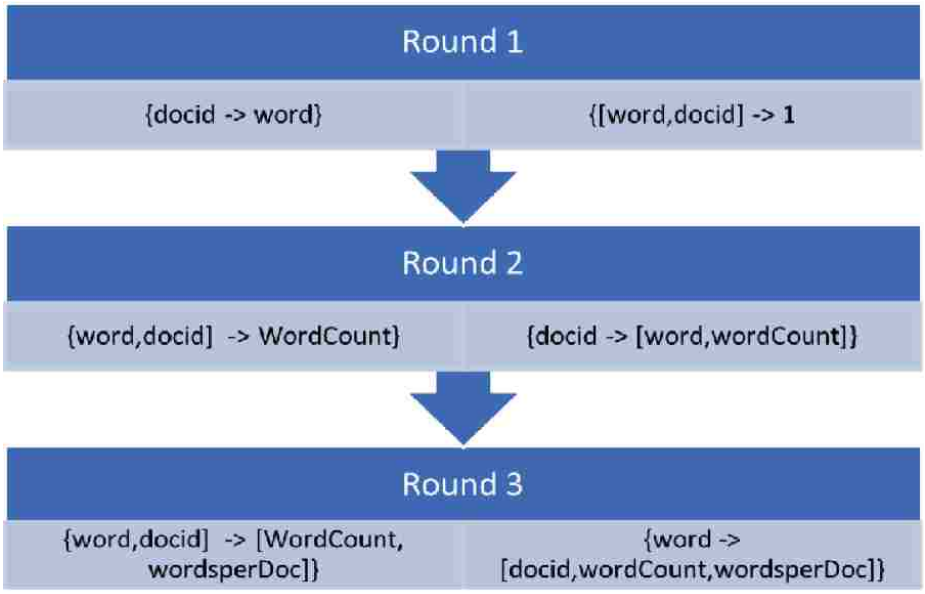

Figure 3.16. TF-IDF algorithm flow chart

The ratio for image tag t_k compared to average length of image tags would be

$$ratio = \frac{|t_k|}{avg} \tag{2}$$

Term frequency of keyword k_i in query sentence q_j would be

$$tf = \frac{f(i,j)}{f(i,j) + 2 * ratio}. \tag{3}$$

The inverse frequency of keyword k_i in the whole image tags set ϕ would be

$$idf = \log_{10} \frac{n}{\sum_{1 \leq h \leq n} f(i, \phi(h))}. \tag{4}$$

Finally, the term weight of image tag t_k would be

$$w(t_k) = \sum_{1 \leq h \leq |t_k|} tf * idf \tag{5}$$

4. EXPERIMENT RESULTS AND EVALUATION

4.1. EXPERIMENT DATA

TED is a nonprofit devoted to spreading ideas, usually in the form of short, powerful talks (18 minutes or less). TED covers almost all topics from science to business to global issues. It provided mass sentences related to talks and speech from different areas. The SISE gathered thousands of input sentences from TED talk, covering many areas like science, economy, politics, education and so on. Google open API offered good web crawler framework to retrieve mass related image resource, SISE retrieves thousands of images using Google API to match the concept of sentence.

4.2. EXPERIMENT STEPS

4.2.1. Searches on Noun Phrases. As in Figure 4.1, a group of nouns got by Text Extraction Module is the basic form to do image search using Google image search engine. There are only a few of retrieved images that match the meaning of sentence well, sometimes none.The accuracy depends on NLP Noun Extraction part in Text Extraction Module. The Goal of this experiment is to choose images that match the concept of sentence well.

4.2.2. Comprehensive Evaluation. In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the query:

$$precision = \frac{|\{relevant\ documents\} \bigcap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \tag{6}$$

For example for a text search on a set of documents precision is the number of correct results divided by the number of all returned results.

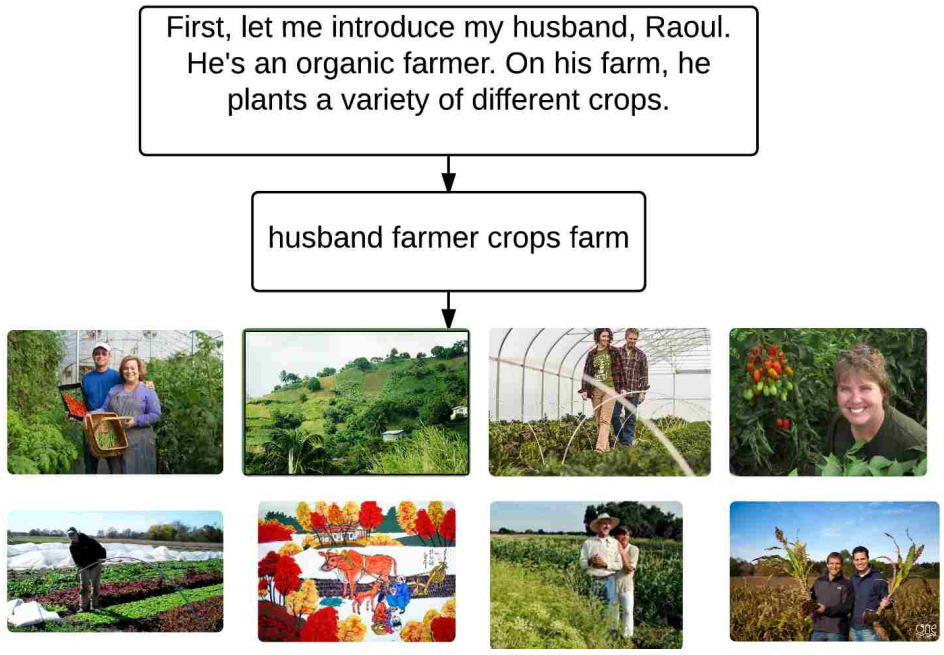

Figure 4.1. An example of SISE gathers nouns and images

Precision is also used with recall, the percent of all relevant documents that is returned by the search. The two measures are sometimes used together in the F1 Score (or f-measure) to provide a single measurement for a system.
Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$recall = \frac{|\{relevant\ \ documents\} \bigcap \{retrieved\ \ documents\}|}{|\{relevant\ \ documents\}|} \tag{7}$$

For example for text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned
In binary classification, recall is called sensitivity. So it can be looked at as the probability that a relevant document is retrieved by the query.

The comprehensive evaluation of this system was focused on usability (perceived image usefulness) aspect. As in Figure 4.2, the image chosen by TF-IDF algorithm is most relevant to the sentence.

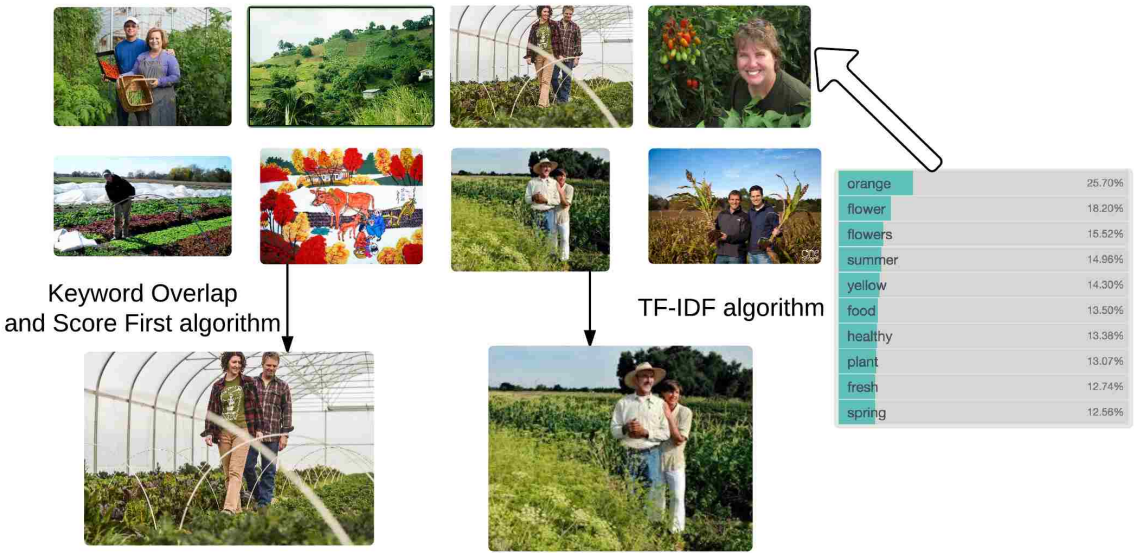

Figure 4.2. An example of images chosen by different algorithms

In the first experiment, the SISE uses 300 groups of sentences, and downloads 10 images for each group of sentences, the number of relevant image groups is 257, the result is shown in Table 4.1.

Table 4.1. Using 300 groups of sentences and 10 images for each group

	Accurate,	Precision	Recall
Keyword Overlap	175	58.3%	68.1%
Score First	168	56%	65.4%
TF-IDF	189	63%	73.5%
Average	177	59.1%	69%

In the second experiment, the SISE uses 300 groups of sentences, and download 20 images for each group of sentences, the number of relevant image groups is 269, the result is shown in Table 4.2.

Table 4.2. Using 300 groups of sentences and 20 images for each group

	Accurate,	Precision	Recall
Keyword Overlap	181	60.3%	67.2%
Score First	173	57.7%	64.3%
TF-IDF	192	64%	71.4%
Average	182	60.7%	67.6%

Figure 4.3 has shown the precision result of three algorithms based on different number of inputted groups of sentences.

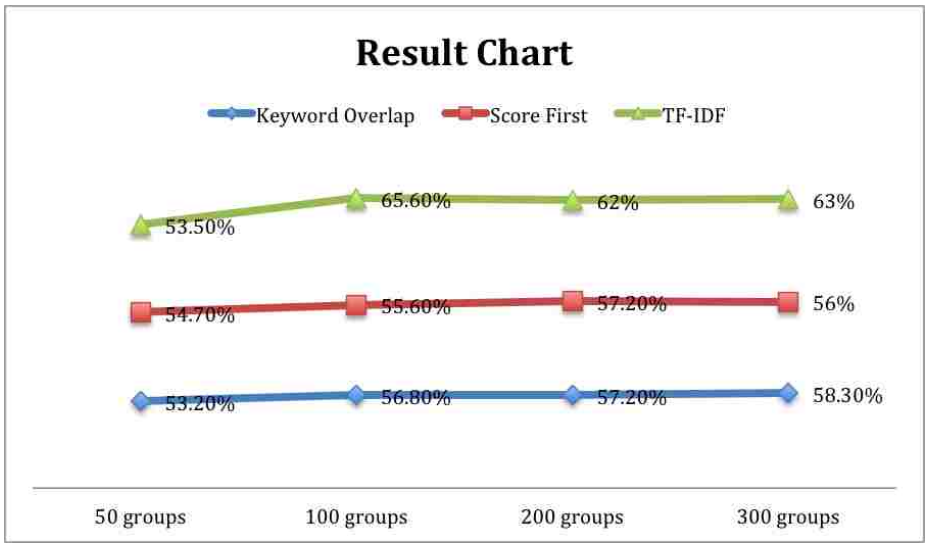

Figure 4.3. Precision based on different groups of sentences

5. CONCLUSION AND FUTURE WORK

SISE is found to be able to retrieve matching images for a wide variety of topics, across different areas, and provide subjectively more useful results than keyword-based image search engines.

The SISE can be worked on improving accuracy by increasing image number in database to be searched for, in this thesis, the trade off will be the time to download many images from Internet, however, if there is a big database can be built to store these images and tags, then SISE will be trained in the database to run faster and get more accurate results. Besides, with the advancement of Natural Language Processing (NLP) techniques, automatic conception detection will be more accurate to retrieve keywords from sentence, and with the advancement of Computer Vision (CV) techniques, SISE can retrieve more related tags from images, the precision can be increased at the same time. So SISE is promising in the future with the advancement of NLP and CV techniques.

BIBLIOGRAPHY

[1] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *ICCV*, 2007.

[2] V. Murdock, D. Kelly, W. B. Croft, N. J. Belkin, and X. Yuan. Identifying and improving retrieval for procedural questions. In *Information Processing and Management*, volume 43, pages 181– 203, January 2007.

[3] K. Toutanova and C. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, pages 63–70, 2000.

[4] K.Toutanova, D. Klein, C. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*, pages 252–259, 2003.

[5] A. Rajaraman and J. D. Ullman. Data mining. In *Mining of Massive Datasets*, pages 1–17, 2011.

[6] X. Li, C. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. In *IEEE Transactions on Multimedia, 11(7)*, November 2009.

[7] K. Wnuk and S. Soatto. Filtering internet image search results towards keyword based category recognition. In *CVPR*, 2008.

[8] T. Berg and D. Forsyth. Animals on the web. In *CVPR*, 2006.

[9] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from googles image search. In *ICCV*, 2005.

[10] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *ECCV*, 2004.

[11] L. J. Li, G. Wang, and L. Fei-Fei. Optimol: automatic object picture collection via incremental model learning. In *CVPR*, 2007.

[12] J. Smith and R. Chang. Searching for images and videos on the world-wide web. In *Center for Telecommunication Research Technical Report*. Columbia University, 1996.

[13] E. Akbas and F. Vural. Automatic image annotation by ensemble of visual descriptors. In *IEEE CVPR*, pages 1–8, 2007.

[14] K. S. Goh, E. Y. Chang, and B. Li. Using one-class and twoclass svms for multiclass image annotation. In *IEEE TKDE*, pages 1333–1346, 2005.

[15] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. In *Journal of Machine Learning Research*, 2003.

[16] P. Duygulu, K. Barnard, J. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *IEEE ECCV*, pages 97–112, 2002.

[17] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, IEEE ECCV*, pages 1–22, 2004.

[18] Y. Jiang, C. Ngo, and J. Yang. Towards optimal bag-offeatures for object categorization and semantic video retrieval. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, pages 494–501, 2007.

[19] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. In *IEEE TPAMI*, pages 985–1002, 2008.

[20] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. In *International Journal of Computer Vision*, pages 157–173, 2008.

[21] D. Cai, S. Yu, J. R. Wen, and W. Y. Ma. Vips: a visionbased page segmentation algorithm. In *Microsoft Research*, 2003.

[22] C. Dan, M. Ueli, and S. Jrgen. Multi-column deep neural networks for image classification. In *CVPR*, 2012.

[23] C. Dan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber. Flexible, high performance convolutional neural networks for image classication. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*, 2011.

[24] A. Mikulik, O. Chum, and J. Matas. Image retrieval for online browsing in large image collections. In *6th International Conference, SISAP*, 2013.

VITA

Weizhi Meng was born in 1988 in Guangxi, China. He received the B.S. degree in computer science from National University of Defense Technology, Changsha, China in July, 2011. Since January 2014, he has been a Master student in Missouri University of Science and Technology in Professor Yiyu Shi's group.