A SEARCH ENGINE FOR WEB IMAGES USING DOCUMENT TEXT STEMMING

Ryan Hardt, Ethan V. Munson
Department of EECS, University of Wisconsin-Milwaukee, 2200 E. Kenwood Blvd., Milwaukee, WI, U.S.A.

Hien Nguyen
Department of Mathematical and Computer Sciences, University of Wisconsin-Whitewater, Milwaukee, WI, U.S.A.

Keywords: Stemming, image retrieval, text-based image retrieval, information retrieval.

Abstract: A Web image search application was built using a previously-developed image relevance model for retrieval of images via text-based image retrieval. The application includes a text stemmer that converts a word to a canonical form, making it possible to match text in the face of changes in tense or plurality that have little effect on semantics. The usefulness of stemming in Web image retrieval was evaluated via a test on ten queries that were submitted both with and without stemming. Relevance of retrieved images was determined via ratings by three trained individuals. With stemming, the average unique relevance recall (a measure of the proportion of relevant images returned by one algorithm and not another) was 27.7%, while without stemming, it was only 0.5%. These results may more accurately apply to queries containing at least one plural noun, present tense verb, present participle verb, or past tense verb.

1 INTRODUCTION

Google Image Search (Google, 2007), and Yahoo Search - Image Search (Yahoo! Inc., 2007) are image retrieval search engines (Zhang, 2005) that are actually text-based. They take advantage of the fact that images on Web pages are inherently annotated by text within HTML documents. Text describing images present on a Web page is often contained within the document. It has been shown that certain locations in an HTML document are particularly likely to contain these annotations, such as the image filename, page title, or text surrounding the image (Thao, 2003). This approach to text-based image retrieval does not require a formal image annotation or content analysis process. Because of this, existing Web content can be used without requiring manual annotation of images that are already present on the Web.

When Web pages are searched for text matching a query, each word in the query is compared to each word in the Web page in an attempt to find a match. However, if the user enters the word "crying" in the query, this will not match other forms of this word like "cry," "cries," or "cried." Additionally, a common convention in naming multiple files of similar subject matter is to append an underscore followed by a sequence number. An example of this for an image filename may be "crying_02.jpg." A straight forward word-for-word comparison would not match this image filename to the input query of "crying." These short-comings could potentially cause desirable images to be left out of the result set.

Our research expands upon work that has shown that the combination of the structure of HTML documents and image locations within those documents can be used to determine the relevance of an image based on a given query (Thao, 2003). Through analysis of HTML features using a collection of 2,400 Web pages and 5,806 images gathered by queries returned from a Google Web page search (Google, 2007), a statistical model combining the most useful features was developed. This paper discusses the use of this statistical model in building a search engine for images on the Web. This search engine makes use of a technique called "stemming," in which any word in the query or Web page is reduced to its stem. A stem is a canonical form that all other variations of a given word will be

Hardt R., V. Munson E. and Nguyen H. (2008).
A SEARCH ENGINE FOR WEB IMAGES USING DOCUMENT TEXT STEMMING.
In *Proceedings of the Fourth International Conference on Web Information Systems and Technologies*, pages 223-230
DOI: 10.5220/0001526502230230

reduced to if processed by the same stemming algorithm. By stemming both the query and returned Web pages, relevant images will not be omitted from the result set simply because of variations in word form between a Web page and the query.

Because Web pages contain non-semantic text, stemming techniques used on traditional text documents are not necessarily directly applicable. While some research has shown that stemming is successful when searching for images using queries of certain lengths on some datasets (Hull, 1995) or when substantial search techniques are applied in addition to traditional text-based searching (Kraaij, 1996), other research claims that stemming has little effect (Harman, 1991). We aim to test our hypothesis that stemming is beneficial for text-based image retrieval using queries containing specific parts of speech.

As part of this research, we implemented a set of software tools in order to:

- Allow the user to submit a one or two word query
- Stem the query and the text in a Web page
- Obtain an image relevance rating
- Provide an interface to view results

2 BACKGROUND AND RELATED WORK

An image retrieval system is one in which a query is made to a system in an attempt to find images that correspond to that query. Image content must be translated in some way to allow for comparison to the search query. It is in identifying both the relevant image features as well as the information source that makes this translation challenging.

In this section, we will discuss research in which an image relevance model for an image retrieval system was developed as well as the concept behind stemming.

2.1 Previous Project Work

This paper continues research in which a relevance model for image search on the Web was developed. Query text was searched for among 53 different HTML formatting features, such as image filename, page title, and image ALT attribute. Using stepwise logistic regression, a relevance model containing 13 independent variables was developed.

Table 2.1: Twenty-four queries and their categories used to generate data.

ID	Query	Category
1	Bill Gates	Famous People
2	George Bush	Famous People
3	Britney Spears	Famous People
4	John White	Less Famous People
5	Michael Brown	Less Famous People
6	William Black	Less Famous People
7	New York	Famous Places
8	Michigan Lake	Famous Places
9	Yellowstone Park	Famous Places
10	Spokane	Less Famous Places
11	Burlington Vermont	Less Famous Places
12	Haw River	Less Famous Places
13	New Year	Holiday
14	Thanks Giving	Holiday
15	Halloween	Holiday
16	Happy Child	Concept
17	Sad Woman	Concept
18	Burning House	Concept
19	Raining	Phenomenon
20	Volcano Erupt	Phenomenon
21	Bomb Explode	Phenomenon
22	Eiffel Tower	Landmarks
23	Vietnam Memorial	Landmarks
24	Statue Liberty	Landmarks

2.1.1 Query Selection

To develop this model, a set of 24 queries belonging to one of eight different categories was chosen as shown in Table 2.1. The variety of query categories was chosen in an effort to evaluate possible differences in effectiveness among different types of image content.

2.1.2 Image Downloading and Result Storage

With over 11.5 billion Web pages present on the Web (Gulli, 2005), the task of building a system to collect a representative sample of the Web would be very difficult. Because of this, the Google Web page search engine was used as an information source. A command line interface was used to extract up to 1000 URLs returned by Google and randomly select the URLs of 100 pages to download.

Each of these links was visited and its contents downloaded. Both the HTML document and all of its images were stored locally. A database was used

to store information about the downloaded web pages, images, and the image ratings.

2.1.3 Document Analysis

After all HTML documents and images were downloaded, they were examined to discover which HTML features are useful for determining image relevance. A total of 53 HTML features were identified as possibly useful in a relevance model. After the images were rated for relevance by human raters, each of the 53 features was analyzed using frequency-based statistics and logistic regression, which will be discussed in the next section. Based on this analysis, 13 of the original features were determined to be useful in determining image relevance. These features are shown in Table 2.2.

Text in the search query was compared to text in each of these features in a given Web page returned from Google. All features except for image filename, image path, page filename, and page path were allowed to return a match with the presence of a maximum of 20 characters between any two query terms. This allows adjectives or adverbs to appear between query terms in the Web page without preventing a match. These terms must, however, appear in the same order as they appear in the query. The 4 previously mentioned features must contain at least one matching term to return a match. For every image, each feature was assigned either a 1 if the feature was present in relation to the image and contained text matching the query text, or a 0 otherwise. These values were computed for all features for all images and recorded in the database.

2.1.4 Data Analysis

Each of the 5,806 images was rated by three human raters who were instructed on how to rate the images. Images received one of three ratings: relevant, somewhat relevant, or irrelevant. For the purpose of data analysis, an image was considered relevant if it was rated relevant by at least two of the three human raters.

Each of the 53 HTML features was examined independently using frequency-based statistics. The frequency at which the HTML features both occurred in the documents and contained matching query text or did not occur in the documents was recorded. This information was cross-tabulated with the relevance ratings of the three human raters to identify 19 HTML features that appeared to be useful for the construction of a data model.

Table 2.2: Thirteen useful HTML features for determining image relevance.

ID	Feature
1	Image Filename
2	Page Title
3	Page Filename
4	ALT Attribute
5	Image Path
6	Page Path
7	Cell Below
8	Meta Description
9	Cell Above
10	Other Body Text
11	Anchor Text
12	Cell Right
13	Cell Left

Logistic regression identifies the linear combination of a set of independent variables to determine the probability of an event occurring. For this research, the set of independent variables was the HTML features and the event is the successful returning of an image based on a given query. Through a process called *forward stepwise regression*, each feature was analyzed and added to an equation one at a time according to its improvement to the model. The resulting formula used 13 of the remaining 19 HTML features. This process resulted in a data model for determining image relevance.

2.2 Stemming

Stemming is the process of reducing a word to its stem. A typical English word contains a *stem*, which refers to the central meaning of the word. In addition to a stem, a word also consists of affixes, or prefixes and suffixes. These affixes are appended to allow slightly different meanings or usages of the stem. All forms of a given word will ideally result in the same term when stemmed using the same stemming algorithm. This stem does not have to be a word itself. A collection of related words and their stems can be seen in Table 2.3.

Table 2.3: Various stems and the words that generated them.

Stem	Words
cri	cry, cries, crying, cried
bike	bike, bikes, biking, biked
run	run, runs, running
dog	dog, dogs

Words are typically composed of a prefix, root, and suffix. Prefixes are considered to have more

influence over the meaning of a word. Thus, their removal from a word can drastically alter the meaning of a word. On the other hand, suffixes generally do not provide a substantial amount of meaning to a word. Therefore, the removal of suffixes is the focus of most stemmers. English has two main types of suffixes, inflectional and derivational (Tars, 1976). Common endings for both types of suffixes can be seen in Table 2.4 and Table 2.5. While it should be noted that some specific domains have suffixes that provide important information about the word, suffix removal is generally acceptable in stemming.

Table 2.4: Common inflectional suffixes.

Plural Nouns	-s, -es, -ies
Present Tense Verbs	-s, -es, -ies
Present Participle Verbs	-ing
Past Tense Verbs	-ed, -ied

Table 2.5: Common derivational suffixes.

Nominalizations	-ing, -ion, -ment, -ist, -ness, -ship
Adjectivals	-al, -ic, -ful, -ous
Adverbials	-ly
Verbals	-ize

There are two types of errors subject to stemming: *understemming* and *overstemming* (Paice, 1994). Understemming occurs when a stemming algorithm does not reduce words that refer to the same concept to the same stem. Stemmers that are prone to this type of error are referred to as "*light stemmers*." Overstemming occurs when a stemming algorithm reduces words that refer to different concepts to the same stem. Stemmers that are prone to this type of error are referred to as "*heavy stemmers*." For this application, we used a light stemmer to reduce the probability of obtaining undesirable matches.

3 METHODOLOGY

The goal of this project was to build a Web application that allowed users to search for images both on the Web and on a local image collection and to explore the usefulness of stemming in the area of text-based image retrieval. To accomplish this, it was necessary to: implement a collection of tools to search and parse Web documents, create an interface to search a locally stored image collection, create a Web interface, and implement a stemming algorithm. This section will address these issues.

3.1 Accessing Content from the Web

When a query is submitted, a request is sent to Google for a Web page search. The format of the returned page was examined to discover the HTML structure in which the returned links were specified. This HTML document is then parsed and the returned links extracted and stored in a list. Only the first 100 URLs returned are stored due to the time required to parse each page.

3.2 Accessing Locally Stored Web Pages

In addition to online Web image retrieval, this application also provides an interface to an existing collection of 2,400 Web documents and 5,806 images generated and stored by Thao (Thao, 2003).

Our application allows queries to be sent to this database. Only the original 24 queries can be submitted for a search against the Web pages stored in the database. Instead of obtaining a list of URLs from Google, the user receives a list of local file paths from the database that point to the downloaded Web documents.

3.3 Stemming Software

To stem text, we used the Snowball stemming algorithm (Porter, 2007), which is based on the Porter algorithm (Van Rijsbergen, 1980). This algorithm is regarded as the de-facto standard in English stemming for information retrieval systems.

Even if stemming is desired, the query remains unstemmed when sent to Google. Also, for some words, the stemmed form returned by our software is not a standard English word, which would make a standard text search substantially less accurate. It is important that we are examining the same dataset when comparing image retrieval with and without stemming. If the query were stemmed before being sent to Google, the pages returned would be different from a dataset generated by an unstemmed query, and the pages would likely be of varying content.

Stemming is implemented on each of the 13 evaluated HTML features individually. When a feature is being explored for the presence of query text, it first has all punctuation removed, since the algorithm will not appropriately stem any word containing punctuation. It is important that each

HTML feature is extracted from the Web document and stored appropriately in a Web page object before the document is stemmed. If the document was stemmed before the HTML feature extraction, the HTML tags would no longer be present as expected, and the features would be difficult to extract. If a punctuation mark is found in one of the HTML features, it is replaced with a space so that the stemmer will see two separate words. The HTML feature being examined is then passed through the stemmer in its entirety and then compared to the stemmed search query. The comparison made between feature and query text is explained in the next section.

3.4 Determining Relevant Images

When comparing query text to text contained in an HTML feature, one of two relevant situations exists.

- The HTML feature exists in relation to the image and contains text matching the query text.
- The HTML feature does not exist in relation to the image or does not contain any query text.

A value is assigned to both situations: 1 and 0 respectively. These values are used in the following equation along with the coefficients seen in Table 3.1 to generate a relevance probability using an equation developed by forward stepwise logistical regression:

$$P_{event} = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} \qquad (3.1)$$

where $Z = B_0 + B_1X_1 + B_2X_2 + ... + B_kX_k$

If the equation produces a probability that meets a specified threshold, the image will be added to a list of other relevant images. Thresholds of 0.7 and 0.5 were used for searches on the Web and on the local database respectively. The threshold was set lower for a local database search due to a relatively small number of images present. These thresholds were chosen arbitrarily through trial and error. It should be noted that images that have a height or width less than 100 pixels or have an aspect ratio greater than 2 are determined irrelevant. These characteristics are common for decorative images or advertisements.

3.5 Image Presentation

When all of the Web pages have been processed, the list of image URLs is sent back to the search page. Images are sorted and displayed according to decrea-

Table 3.1: Relevance model coefficients computed by logistical regression.

Feature	B
Constant	-2.317
Image Filename	1.886
Page Title	1.092
Page Filename	.867
ALT Attribute	1.076
Image Path	1.060
Page Path	-.787
Cell Below	.709
Meta Description	1.092
Cell Above	.664
Other Body Text	-.222
Anchor Text	2.023
Cell Right	.331
Cell Left	.370

-sing relevance. Images are displayed in 3 rows of 5 images each. Images are scaled according to a set width to create a uniform table width. Each image is a link to the original image. When more than 15 images are returned, links to navigate image pages are displayed below the image table.

4 RESULTS

Using the application described in the previous section, data was gathered to examine the effectiveness of stemming in text-based image retrieval for images on the Web. The 24 queries used to produce the local dataset were queried against the dataset with the addition of stemming. Additionally, images produced by 10 queries using pages gathered by Google were examined with and without stemming. When stemming was implemented in a search, it was performed on a given HTML feature only when no match was found for that HTML feature without stemming.

This data was analyzed using two techniques: precision and unique relevance recall. Both of these metrics will be discussed in this section. The results for both queries on the local dataset and queries on the Web will be presented, and an analysis of issues relating to stemming will be given.

4.1 Analysis Models

To evaluate our application, we used two metrics called precision and unique relevance recall. Precision is a measure of the accuracy of a search. It measures the percentage of relevant results among all results provided. Unique relevance recall (URR)

is a metric used to compare two or more search algorithms (Kowalski, 1997). It measures the number of unique relevant results produced by one algorithm and not produced by other algorithms. Because we are exploring the effectiveness of stemming, we compare the results produced by a search without stemming with those produced by a search with stemming. The equations for precision and unique relevance recall are:

$$\text{Precision} = \frac{Number\_Relevant\_Retrieved}{Total\_Number\_Retrieved} \quad (4.1)$$

$$\text{URR} = \frac{Number\_Unique\_Relevant}{Number\_Relevant} \quad (4.2)$$

Image relevance was determined by three trained image raters. If at least two of the three raters deemed an image relevant, the image was determined to be relevant. The goal of this analysis is simply to test our hypothesis that stemming is a useful addition to a text-based image retrieval system for use on the Web.

4.2 Local Image Search Results

Each of the 24 queries used to generate the local dataset were rerun on the local dataset both with and without stemming. In both cases the relevant images were noted and search precision calculated. The images returned using both algorithms were compared, and uniquely relevant images were identified. In 21 of the 24 queries, the results were identical between the two searches. In the remaining 3 queries, stemming produced all of the images returned by the search without stemming as well as a few additional images. Unlike the other 21 queries, these 3 queries contain terms that lend themselves well to stemming. In 2 of these 3 queries, the additional images returned by stemming were relevant to the search query. The results from the local data search can be seen in Table 4.1.

4.3 Web Image Search Results

To examine the effectiveness of stemming on text-based image retrieval on the Web, 10 queries were used. These queries were selected for their perceived appropriateness for stemming. They are not meant to be a representative set of queries for the application of Web image retrieval. These 10 queries were submitted both with and without stemming. The resulting relevant images were identified and the search precisions recorded. In 8 of the 10 queries, the result sets where stemming was implemented contained all of the images produced without stemming along with additional relevant images. In the remaining 2 queries, stemming allowed a match in a feature determined to lower the relevance according to the image relevance equation. The average precision among the 10 queries was 82.5% without stemming and 84.5% with stemming. The average unique relevance recall with stemming was 0.5% and 27.7% with stemming. The results from the Web image search can be seen in Table 4.2.

4.4 Discussion

This research shows that stemming is useful to a certain extent in text-based image retrieval for obtaining additional relevant results. It also shows that stemming a given HTML feature only when a match is not found without stemming allows stemming to be implemented in a manner that is unlikely to exclude results that would have been returned had stemming not been implemented.

On the queries ran against the local data set, a small improvement was seen with the addition of stemming. Of the 24 queries, only 3 produced additional images with stemming. Of these 3 queries, 2 returned additional relevant images. This is likely due to the nature of the queries themselves. A total of 15 of the queries were proper names of people, places, or monuments. Because stemming is applied in an effort to create matches among multiple word forms, it follows that queries of proper names would not benefit from stemming. For the same reason, the 3 queries of holidays, "new year", "thanks giving", and "halloween" are unlikely to benefit from stemming. Of the remaining 6 queries, 2 contained the result set that was improved by stemming. These queries were "burning house" and "raining". Both of these queries contain a present participle verb ending in "ing". Stemming in these situations not only returned additional relevant images, thus raising the URR, but also improved the precision for the search. The one query that produced additional results, none of which were relevant, was the query of "thanks giving". Due to the word "giving" in this query, it is not surprising that stemming produced additional results. However, "thanks giving" is typically written as a single word "thanksgivng", which may have some effect on the results for this query. These results illustrate the fact that stemming is not necessarily beneficial for all types of search queries.

For the result set produced by image search on the Web, stemming proved to be useful. The average URR with stemming was 27.7% while increasing the precision by 2.0% from 82.5% to

Table 4.1: Local data search results.

ID	Query	Precision w/o Stemming	Precision w/Stemming	URR w/o Stemming	URR w/Stemming
1	Britney Spears	0.903	0.903	0	0
2	George Bush	0.623	0.623	0	0
3	Bomb Explode	0.5	0.5	0	0
4	Eiffel Tower	0.703	0.703	0	0
5	Statue Liberty	0.836	0.836	0	0
6	Vietnam Memorial	0.511	0.511	0	0
7	Volcano Erupt	0.583	0.583	0	0
8	Raining	0.286	0.333	0	0.333
9	Bill Gates	0.698	0.698	0	0
10	Happy Child	0.133	0.133	0	0
11	Burning House	0.750	0.800	0	0.25
12	Sad Woman	1.00	1.00	0	0
13	New Year	0.367	0.367	0	0
14	Halloween	0.773	0.773	0	0
15	Thanks Giving	0.200	0.182	0	0
16	John White	0.911	0.911	0	0
17	Michael Brown	0.800	0.800	0	0
18	William Black	0.750	0.750	0	0
19	New York	0.722	0.722	0	0
20	Michigan Lake	0.510	0.510	0	0
21	Yellowstone Park	0.836	0.836	0	0
22	Burlington Vermont	0.640	0.640	0	0
23	Spokane	0.700	0.700	0	0
24	Haw River	0.833	0.833	0	0

Table 4.2: Web image search results.

ID	Query	Precision w/o Stemming	Precision w/Stemming	URR w/o Stemming	URR w/Stemming
1	Downhill Skiing	0.828	0.861	0.034	0.222
2	Crying Babies	0.933	0.882	0	0.118
3	Whitewater Rafting	0.841	0.877	0	0.228
4	Sculptures	0.943	0.933	0.014	0.387
5	Mountain Climbing	0.857	0.868	0	0.079
6	Puppies	0.941	0.918	0	0.265
7	Kittens	1.000	1.000	0	0.412
8	Candied Apples	0.667	0.714	0	0.429
9	French Fries	0.742	0.765	0	0.088
10	Runners	0.500	0.636	0	0.545

84.5%. The queries chosen were selected because of their perceived appropriateness for stemming. Each of the 10 queries used in the Web searches contained at least one plural noun, present tense verb, present participle verb, or past tense verb. Therefore, the results produced from stemming may more accurately apply to queries containing one or more of these word forms. Based on the results produced from the local Web page dataset, stemming is not likely to improve queries without these word forms.

Another important observation is that stemming, implemented only when a match was not found using traditional matching alone, excluded very few images returned from a search using only traditional

matching. When implemented on all features regardless of a match without it, stemming excluded many more images produced by the search without stemming. A possible reason for this is due to a common file naming convention for images on the Web. Images tend to have names like "babycrying.jpg" or "puppies02.gif" where either multiple words are combined or a sequence number is appended to the text describing the image. Because the image filename is one of the greatest contributors to image relevance according to our relevance model, a lack of a match in this situation could make the difference between an image being in the result set or being excluded. While a traditional comparison between query text and image filename may find a match in a similar situation, stemming may not find a match. This issue may also arise in the page filename, image path, page path, and ALT attribute, as it is not uncommon for the ALT attribute to be given a value equal to the image filename.

5 CONCLUSIONS

With the availability of nearly any desirable image on the Web, effective tools for searching for images are needed. Because of the incredibly large size of the Web as well as its constantly changing nature, new techniques need to be explored and implemented. While some commercial image search engines exist, the techniques used in these search engines are mostly unknown.

Stemming was most beneficial for queries that contained at least one plural noun, present tense verb, present participle verb, or past tense verb. Image searches with stemming on queries containing at least one of these word forms returned additional relevant images compared to searches without stemming while also slightly increasing precision. Additionally, by stemming an HTML feature only when a query text match was not found without stemming, we were able to obtain nearly all images returned from a search without stemming.

REFERENCES

Google, 2007, "Google." *http://www.google.com/*.

Google, 2007, "Google Image Search."*http:// images.google.com*

Gulli, A., Signorini, A., 2005, "The Indexable Web is More Than 11.5 Billion Pages," *World Wide Web Conference 2005*.

Harman, D., 1991, "How effective is suffixing?," *Journal of the American Society for Information Science*, Vol. 42(1), pp. 7-15.

Hull, D. A., 1996, "Stemming Algorithms: A Case Study for Detailed Evaluation," *Journal of the American Society for Informational Science*, Vol.47, No.1, pp.70-84.

Kowalski, G., 1997, *Information Retrieval Systems - Theory and Implementation*, Springer, pp. 223-233.

Kraaij, W., Pohlmann, R., 1996, "Viewing stemming as recall enhancement," *ACM Special Interest Group on Information Retrieval '96*.

Paice, C. D., 1994, "An Evaluation Method for Stemming Algorithms," *Proceedings of the 17th annual international ACM Special Interest Group on Information Retrieval*.

Porter, M., 2007, "Snowball." *http://snowball.tartarus.org*

Tars, A., 1976, "Stemming as a System Design Consideration," *5th Annual Ada Semantics Iinterface Specification Conference*.

Thao, C., Munson, E., 2005, "A Relevance Model for Web Image Search," *Workshop on Web Document Analysis 2003*.

Van Rijsbergen, C., Robertson, S., Porter, M., 1980, New Models in Probabilistic Information Retrieval, *British Library Research and Development Report*.

Yahoo! Inc., 2007, "Yahoo Image Search." *http:// images.search.yahoo.com/*.

Zhang, C., Chai, J. Y., Jin R., 2005, "User Term Feedback in Interactive Text based Image Retrieval," *ACM Special Interest Group on Information Retrieval '05*.