Automatic Recognition and Extraction of Web Comics

Ben E. Cline
braininfo *at* benjysbrain.com

Abstract

The ability to data mine comic strips from the web is useful for creating customized daily comics pages; however, automatically selecting the primary comic from a comic strip web page is difficult because banners, advertisements, and decorative images can be of the same image type as the primary comic. Web comics can be color or monochrome, line drawings, graphic novel-style drawings, 3D rendered images, or even photographs with speech balloons. Furthermore, a page might contain a primary comic that is the most recent comic for the strip along with previous comics. Comic page formats also vary between sites, and site formats change over time. A technique was developed for creating data mining software to extract the primary comic on a daily basis from a number of comic strip web pages with a goal of robustness in the environment of changing web site layouts. Image and web page features that are useful in automatically identifying primary comics are described, as is the training of support vector machines (SVM) using these features. The ComicsMiner demonstration software, which uses the SVMs and additional knowledge rules, was developed to test the primary comic extraction technique. It was evaluated on sets of web sites containing traditional newspaper comic sites, editorial comic sites, and web-only comic sites. The sites contain comics with a wide variety of styles, formats, and sizes embedded in a variety of web page formats. Evaluation of the software demonstrates that the system has a high accuracy rate in extracting primary comics.

Introduction

The ability to data mine comic strips from the web is useful for creating customized daily comics pages. The process starts with a user-defined list of comics and the web site URLs for sites containing these comics. The extraction software then runs every morning to extract the primary comic from each site so that the user can be presented with a daily comics page. The extraction system must be able to work reliably and deal with a variety of comic formats, web site changes, comic format changes, and changes to advertisement images without manual intervention.

The primary comic on a comics web site is the most recent comic related to the comic strip. Recognition of the primary comic image on a page crowded with similar image types can be difficult. The difficulties arise from the diversity of comic styles and formats, the placement of primary comics among similar image types, the variety of styles of the containing web sites, and the dynamic nature of the sites.

Comics on the web can be traditional newspaper comics, editorial cartoons, or web-only comics. They vary from traditional line drawings to images created with 3D rendering software. They can be monochromatic or in color. They can even be photographic images with speech balloons and text added. To further complicate the identification of comics, advertising images, banner images, and decorative images can share many of the features of comic images, e.g., a typical page might contain comics and advertisements that are all line drawings.

Adding to the difficulty of identifying the primary comic image on a page, comic web pages are typically dedicated to a single comic but vary greatly in format from one comic page to the next. Some pages contain a single comic along with banners and advertisements, while others also include previous comics. Some comic authors post comics using blog software and sometimes intermingle comic images with those of a personal nature, e.g., a page containing comics and photographs of the attendees of a comics convention.

Comic web sites also vary in the rate at which new comics are posted. Some comics are updated daily, while others are Sunday-only comics. Some editorial cartoons are produced at a variable rate. Web-only comics, sometimes drawn as a second occupation of the author, can be updated sporadically as time allows.

Another impediment to automatically recognizing the primary comic on a web page is that the site layout can be changed at any time to produce a newer, fresher design or to accommodate a new advertisement layout. A data

mining system for comics must be able to handle site changes and still recognize the primary comic on the page.

A technique and demonstration system for automatically extracting primary comic images from web pages containing comics is described in the following sections. The system is robust and handles a wide variety of comic formats and web site styles. It has a high degree of accuracy identifying the primary comic of comics pages.

A number of image and web page features, described in the next two sections, were identified that help in the recognition of primary comic images, and these features were used to train Support Vector Machines (SVM) (Vapnik 1995) that form the heart of the comics extraction software. Images were collected from the sites to be mined and were manually tagged as either the primary comic image on a page or not. This data, along with image and page features, formed the training set for the SVMs. The SVMs are supported by a set of filter rules that increases the accuracy and robustness of the system.

As described in the Evalution section, the ComicsMiner demonstration system was evaluated on two sets of comic sites for accuracy, recall, and precision over a two month period as the sites changed. Both sets of sites contained traditional comic sites, editorial cartoon sites, and web-only comic sites along with a small number of news and commerce sites. Sites from the first set of sites were manually selected while the sites from the second set were selected randomly from a master list of comics sites. The first set had more traditional comics while the second set contained more nontraditional web-only comics with a greater variety in style, size, and format. The software extracted primary comics accurately from both sets of sites. Additional experiments were performed to determine how well the system performed in mining comics from sites not used in SVM training. In this case, the SVM was less accurate; however, the overall system still performed well.

Comic Image Features

Web page image features are used to differentiate between the primary comic image and other images on a page. Because of the diversity of comic pages and comic formats, there is no single feature that can be used to select the primary comic from a set of comic pages containing traditional and web comics. Finding a set of features that works in all cases is perhaps impossible. The approach taken here is to find a set of features that can be used to create a classifier that locates the primary comic with a high degree of accuracy and then supplement the classifier with heuristic rules to improve its reliability.

Web page image features belong to two categories: low-level image processing features and context. Low-level image processing features include the image size, area, texture features, and color information. The context features of an image include the image location on the page, the HTML structure in which it is contained, and surrounding text. Because comic images are presented using the HTML IMG tag, the optional alternative text (ALT attribute) sometimes contains context information about the image.

A web page image has a real size and a rendered size. If the IMG tag contains HEIGHT and/or WIDTH attributes, the rendered size of the image might be different than the real size. On some pages, an image is used in different ways with different sizes, e.g., a primary comic at full size might also be used as a link image using the HEIGHT and WIDTH attributes to render it as a thumbnail. An image also has real and rendered areas and aspect ratios.

Image sizes and aspect ratios have been used successfully in web image retrieval systems to filter images (Nakapan, Halin, Bignon, and Wagner 2004, Hu and Bagga 2003). Cline (2008) used image size and aspect ratio to aid in the identification of traditional comics.

Low-level image processing techniques have been used in a number of web image crawlers to determine which images are photographs and which are line drawings. For example Frankel, Swain, and Athitsos (1996) used color features and Hu and Bagga (2003) used both color features and the DCT (Digital Cosine Transform) of image subregions.

Since most traditional comics are line drawings, image features that can identify drawings are useful in identifying traditional comics. Because some web-only comics are not line drawings and because banners and advertisements can be either line drawings or photographic images, differentiating between image types is not a perfect means to identify a comic; however, it does provide useful information in concert with other image features. A number of low-level image processing features were investigated for this study including angular second moment, contrast, correlation, and grayscale histograms (Haralick, Shanmugam, and Dinstein 1973). Figure 1 shows the typical difference between the grayscale histograms of photographs and line drawings. This feature is computationally inexpensive but performed better at differentiating between line drawings and photographs than the other texture features considered.

Some systems use a combination of low-level image processing techniques and information about text near the image to categorize images (La Cascia, Sethi, and Sclaroff 1998, Feng, Shi, and Chua 2004, and Wang and Kan 2006). Text can come from HTML text, image filenames, and the contents of the optional IMG tag attribute ALT. Others have used Optical Character Recognition (OCR) to use text in images as a feature used for categorization, e.g., Bigham, Kaminsky, and Ladner (2006).

Some systems use web page layout to determine content salience. These web page segmentation systems attempt to separate web page regions into components such as navigation menus, advertisements, banners, and main content (Cai, Yu, Wen, and Ma 2003, Yi, Liu, and Li 2003, Hattori, Hoashi, Matsumoto, and Sugaya 2007, and Kohlschutter and Hejdl 2008). Various techniques are used including text density, object placement on a web page, and underlying HTML structure.

Page layout is important for determining the primary comic on a comic page. These comics are typically given a prominent position on the page. A common format has the primary comic close to the upper left corner of the page under a banner; however, other styles are sometimes used.

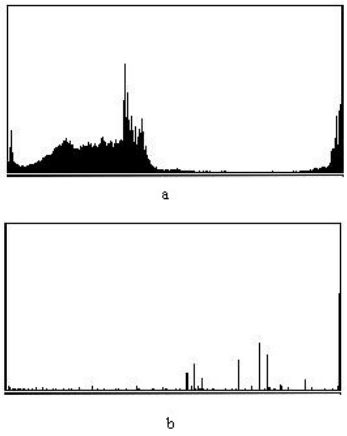

Figure 1- Grayscale histograms of a photo (a) and a traditional comic (b)

Comic Extraction

Both Artificial Neural Network (ANN) and SVM classifiers were constructed for this project. The SVM classifiers performed better than the ANN classifiers and were used for the evaluation tests described in the next section. Because of the non-linear nature of the data, a radial basis function (RBF) kernel was used for the SVMs. The LIBSVM (Chang and Lin 2011) software was used for SVM training and to implement the classifier.

The comic extraction process uses an SVM to help determine the primary comic on a page. The SVM is constructed using the following process:

1. Select a list of image and page features to be used in the categorization process.
2. Define the list of comic sites to be mined.
3. Build a library of images, collected over a period of time, from the sites to be mined.
4. Manually rate the images as being the primary comic or not.
5. Do a course grid search to determine a suitable RBF γ parameter and suitable SVM error penalty (Chang and Lin 2011).
6. Train an SVM using the image and page features in step 1 and the parameters determined in step 5.

The SVM classifier uses the inputs given in Table 1. Height and width are based on the true image size, while area is based on the rendered size of the image. The (x, y) position of the image is the position of the upper left of the image relative to the top left corner of a 1024 x 768 browser window. The avg-delta parameter is the average grayscale difference between adjacent grayscale value counts, and asm-avg is the average angular second moment calculated from the 0°, 45°, 90°, and 135° angular second moments. The parent-tag parameter is based on the parent tag containing the IMG tag indicating the image. Based on the parent tag, this value is given a numeric value between 0 and 1 and represents the complexity of the structure surrounding the IMG tag. A 0.0 value is given to invisible images while 1.0 is given to images inside a complicated HTML structure. Images with a BODY or DIV tag parent have scores in between.

Parameter	Parameter Description
h, w	Real height and width of the image
area	Area of the rendered image
aspect ratio	Image aspect ratio (h/w)
(x, y)	Image position from upper left in pixels
avg-delta	Average grayscale difference
asm-avg	Average angular second moment
parent-tag	Encoded value based on IMG tag parent
alt-text	Encoded value based on ALT text

Table 1 - SVM Input Parameters

The ALT text parameter also varies between 0 and 1 based on the contents of the ALT attribute of the IMG tag. When the text has a keyword that definitely indicates an advertisement, the parameter value is 0.0. A text string that indicates the comic of the day is given a value of 1.0, while missing or indefinite text is assigned a value of 0.5. A table of fifteen keywords, along with the comic strip name and current date, is used to determine this value.

Figure 2 shows the complete architecture of the ComicsMiner demonstration system that uses the SVM classifier and some additional rules. Images collected for training and for production comics extraction are first pre-filtered to remove small images and images from servers whose host name begins with “ads”. The minimum image size considered is 62,000 square pixels, which was determined empirically. The pre-filtering improves the

efficiency of the software, reducing the number of images processed by a factor of eight or more and improves the accuracy of determining the primary comic by removing small decorative images that might otherwise be considered to be comics.

The remaining images from the page under consideration are then processed by the SVM. The images are processed in either one or two passes. On the first pass, only images with an area of 79,000 or more square pixels are processed. If the SVM does not find a primary comic, a second SVM pass is executed with the area filter removed so that the smaller images are processed by the SVM. The two-pass version of the SVM is only slightly more accurate than the single pass version as the majority of comics pages have a primary comic that is larger than 79,000 square pixels; however, this configuration allows the system to capture the rare smaller primary comic while maintaining good accuracy for the typical comics site.

A simple learning system based on user feedback improves the SVM filter. When the user views a daily page of comics created by the system, any false positives such as an advertisement displayed as a primary comic, are flagged by the user. Known advertisement URLs are kept in a table, and the corresponding images are discarded by the SVM filter stage on subsequent runs. This simple technique prevents the system from annoying the user with the same set of advertisements each day and helps to improve the accuracy of the system by removing images from consideration that were improperly classified by the SVM.

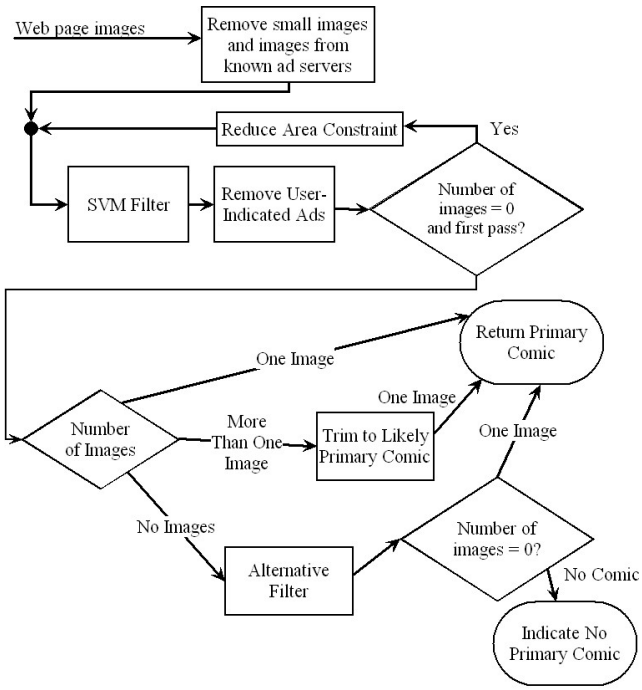

Figure 2 - ComicsMiner Architecture

The final stages of the filter behave differently based on the number of images selected by the SVM. If the SVM identifies a single image, this image is the one presented to the user as the primary comic from the given page. If no primary comic is found, an alternative method of primary comic selection is applied to the images from the page. If the SVM selects more than one image, a second filter method is applied to select the image that is most likely to be the primary comic.

The alternative selection method uses simple rules to select the primary comic. These rules work well on traditional comic strips. The rules first discard all images that are at page position (0, 0). These images are typically generated in JavaScript and are not visible or are decorative or spacing images. The rules then check for images with an area and aspect ratio that fall within the bounds of typical web comic strips. Finally, any image that is unlikely to be a line drawing based on the avg-delta value from Table 1 is also discarded. The first image passing all the rules is selected as the primary comic.

If the SVM selects more than one image on a page, a second filter is applied to select the image that is most likely to be the primary comic. As in the alternative selection method, the second filter discards images at page position (0, 0). If more than one image remains, the filter selects the image closest to the top left part of the page.

Test Set	Number of Comics	Number of Hosts	Comic Types
Test Set 1	87	49	50 Traditional, 20 Web-Only, 9 Editorial, 5 News, 3 Commerce
Test Set 2	199	188	20 Traditional, 180 Web-Only, 2 Editorial, 2 News, 2 Commerce

Table 2 - Description of Test Sets

Evaluation

Two tests were preformed using the primary comic selection method. Table 2 describes the two test sets. Each test consists of a set of comic web sites from which the primary comic is to be extracted daily over a two month period to evaluate how well the ComicsMiner software works in the dynamic web comic environment. The first test consisted of 87 manually selected sites containing comic sites and a small number of news and commerce sites. (News and commerce sites were included in the training set as samples of sites with no comics.) The

second test consisted of 199 sites randomly selected from a master list containing the sites in the first list and a list of over 400 web-only comic sites derived from the Webcomic List web site (Young 2011), a site that contains a list of recently updated web-only comics.

Because many traditional comics are hosted at conglomerate comic sites, pages on the same host typically share a standard format, which is likely to make primary comic recognition easier in Test Set 1. Test Set 2, which has fewer sites on common hosts, provides a more difficult test for the system.

Both test sets include traditional comics, web-only comics, commerce sites, editorial cartoon sites, and news sites. Traditional comics typically have one of three formats: a horizontal format with several panels such as the daily Pearls Before Swine strip, a single panel strip like the daily Family Circus strip and some editorial cartoons, and a larger Sunday format. Web comics come in a variety of sizes and formats, and some are very large relative to traditional comic strips.

A separate SVM was constructed for each test set using images from the sites to be mined. The training set for the first test set was collected over a six month period and contains 905 images, while the test set for the second set was collected over two weeks and contains 1340 images.

During testing, the system presents a single image to the user for each comic strip on a daily basis. The image is either an image selected from the site or a "no comic" image that indicates no comic is available. Each image is manually rated as correct, a false negative, or a false positive. Advertisement images selected from a site with no comic are scored as a false positive. If a site contains a primary comic, but either another image from the site or the "no comic" indication image is presented, the primary comic image is scored as a false negative.

Test 1 Results		
	SVM Stage	Full Filter
Accuracy	99.45%	99.80%
Recall	98.56%	99.80%
Precision	99.50%	99.44%
Test 2 Results		
Accuracy	99.02%	99.69%
Recall	97.30%	99.36%
Precision	99.70%	99.35%

Table 3 – Two-Month Average Test Results

Table 3 shows the results of the two-month tests for both test sets. Results for the full filter show that the additional extraction rules improve the SVM accuracy and recall while reducing the precision only slightly. This result is by design: it is assumed that a user would like to see all the comics requested even if that means a few advertisements are also included.

Test Set 1 Using Test Set 2 SVM		
	SVM Stage	Full Filter
Accuracy	96.82%	99.51%
Recall	90.31%	99.45%
Precision	98.80%	98.65%
Test Set 2 Using Test Set 1 SVM		
Accuracy	94.55%	98.63%
Recall	83.59%	96.26%
Precision	99.38%	99.11%

Table 4 - Tests Switching SVM Models

The accuracy of a daily run is based on the number of correct selections relative to the total number of images considered by the SVM. Precision is $tp/(tp+fp)$, and recall is $tp/(tp+fn)$, where tp is the number of true positives, fp is the number of false positives, and fn is the number of false negatives. The number of true positives is the number of primary comic images correctly scored by the SVM and heuristic filters. Precision and recall are reported as percentages.

Rather than training an SVM just for the set of comics of interest for a particular user, a simpler solution would be to train and use a single SVM. A short nine day test was performed using the SVMs trained for Test Sets 1 and 2 on the opposite test set sites. Table 4 gives the results. The Test Set 2 SVM performed better on the more traditional comic Test Set 1 than the Test Set 1 SVM did on Test Set 2 which contained a greater variety of comics. The second SVM performed better because it was trained on a more diverse set of comics. Although it performed well on a wide range of comics, greater accuracy is still obtained by training the SVM on the sites to be mined.

New Sites Using the SVM from Test Set 2		
	SVM Stage	Full Filter
Accuracy	95.44%	98.24%
Recall	86.69%	94.93%
Precision	99.85%	99.85%

Table 5 - Filter Performance for New Sites

Another short test was performed using the SVM from the second test set. All the active comic sites in the master list of web comics that were not used for training either SVM, a set of 265 sites, were crawled for three days with

the results given in Table 5. The SVM stage of the filter has diminished accuracy as compared with using an SVM on the same sites from which training information was obtained. The full filter still performs fairly well; however, taking SVM training data from the sites to be crawled improves filter performance. It should be noted that the training set for Test Set 2 was collected over a short period. Collecting training data over a longer period would likely improve the SVM performance.

Related Work

Existing applications for this task are somewhat brittle, requiring human modification of extraction parameters as comic sites change (Tanvekski 1996, Medico 2001, Dunham and Bowditch 2003, Parker 2007). Some of these applications use hand-crafted regular expressions to identify which HTML tag contains the source address of a comic image while others reference the n^{th} image of a particular size on a page. These extractors are 100% accurate until a site layout is changed, and then no comics are extracted from the site until the user reconfigures the software. Another approach is to use the size and aspect ratio of comics to decide which images are comics (Cline 2008); however, this approach tends to allow a number of false positives, i.e., advertising images that are marked as being a comic, and does not work well with some of the larger format web-only comics. This software is about 15% less accurate at identifying comics than the ComicsMiner software.

The NPIC system of Wang and Kan (2006) classifies non-photographic web images into a number of categories including cartoons. Their system uses both low-level image features and the textual context of web images in their classifier system. Cartoon images in this work appear to refer to traditional comics and the system does not differentiate between the primary comic on a page and other images. It would tend to classify some web-only comics incorrectly, e.g., those created with 3D rendering software or those based on photographic images. However, the use of low-level image features in the NPIC system is broader than in the ComicsMiner. On tests categorizing images obtained using Google image searches, NPIC was 97.6% accurate identifying the typical line drawing type of cartoon and 81% accurate on cartoons from Wikipedia.

Lienhart and Hartman (2002) developed a technique to differentiate between comics/cartoons and presentation slides/scientific posters. This technique used information about text lines in the images, information about borders, and image aspect ratio to perform the categorization.

Future Work

The methodology described in this paper demonstrates that an accurate web comics extractor can be constructed to create daily personalized comics pages by mining the web for the primary comics from the specified sites.

The evaluation tests showed that training the SVM with data from the sites to be mined provides an accurate system. An improvement on the system would be to create an SVM that was accurate when presented with sites not in the training set. A more extensive library of images will be constructed and a new SVM trained to test if more extensive training data can improve the performance of the SVM over a number of sites not in the training set.

Another area of investigation involves using text close to images to help determine the primary comic. Some sites label comics with the date and name in proximity to the primary comic. Buttons or links indicating the previous or next comic in proximity to a larger image would also be useful in locating the primary comic.

Bigham, Kaminsky, and Ladner (2006) and Guo, Kato, Sato, and Hoshino (2006) have used Optical Character Recognition (OCR) techniques to attempt to read text that is part of web images, extracting information for image categorization or making the text available via text-to-speech processing, respectively. Although OCR processing on typical web images is not perfect, some text strings extracted from images could be used to help determine the image function on the page.

Arai and Herman (2010) have worked on extracting individual panels from comics. Identifying images with frames would be a helpful feature in selecting which images are comics and which are not. Not all comics are subdivided into frames, and some comics have different formats relative to the number of frames from day to day. However, an image with frames is more likely to be a comic than one without frames.

References

Arai, K. and Herman, T. 2010. Method for Automatic E-Comic Scene Frame Extraction for Reading Comic on Mobile Devices. *Seventh International Conference on Information Technology.* 370-375.

Bigham, J., Kaminsky, and R., Ladner, R. 2006. WebInSight: Making Web Images Accessible. *ASSETS 2006.* 181-188.

Cai, D., Yu, S., Wen, J.-R., and Ma, W.-Y. 2003. VIPS: A Vision-based Page Segmentation Algorithm. Microsoft Technical Report MSR-TR-2003-79.

Chang, C-C. and Lin, C-J. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology,* 2(3). 27:1-27:27.

Cline, B. E. 2008. Automatic Extraction of Comics from Web Pages. [Online] http://www.benjysbrain.com/misc/comics/.

Dunham, D. and Bowditch, E. 2003. ComicEZ. [Online] http://a-sharp.com/comicez/.

Feng, H., Shi, R., and Chua, T.-S. 2004. A Bootstrapping Framework for Annotating and Retrieving WWW Images. In *Proccedings of the ACM International Conference on Multimedia.* 960-967.

Frankel, C., Swain, M. J., and Athitsos, V. 1996. WebSeer: An Image Search Engine for the World Wide Web, Technical Report 96-14. Computer Science Department, The University of Chicago, Chicago, IL.

Guo, Q., Kato, K., Sato, N., and Hoshino, Y. 2006. An Algorithm for Extracting Text Strings from Comic Strips. *International Conference on Computer Graphics and Interactive Techniques: ACM SIGGRAPH 2006.* Research poster.

Hattori, G., Hoashi, K., Matsumoto, K., and Sugaya, F. 2007. Robust Web page Segmentation for Mobile Terminal Using Content-Distances and Page Layout Information. *WWW 2007.* 361-370.

Hu, J. and Bagga, A. 2003. Identifying Story and Preview Images in News Web Pages. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition,* 640-644. Edinburgh, IEEE Computer Society Press.

La Cascia, M., Sethi, S., and Sclaroff, S. 1998. Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web. In *IEEE Workshop on Content-Based Access of Image and Video Libraries.* IEEE Computer Society, Los Alamitos, CA. 24-28.

Medico, A. 2001. Dailystrips. [Online] http://dailystrips.sourceforge.net/1.0.27/readme.html.

Nakapan, W., Halin, G., Bigon, J.-C., and Wagner, M. 2004. Extraction of Building Product Image from the Web. *International Journal of Intelligent Systems,* 19:65-78.

Parker, T. 2007. Comics Grabber. [Online] http://tevp.net/projects/comicsgrab/.

Tanvekski, T. 1996. Web Comic Strip Reader. [Online] http://sourceforge.net/projects/webcomicreader/.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory.* New York: Springer Verlag.

Yi, L., Liu, B., and Li, X. 2003. Eliminating Noisy Information in Web Pages for Data Mining. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 296-305. Washington, DC.

Young, A. 2011. The webcomic list [Online]. http://www.thewebcomiclist.com.

Wang, F. and Kan, M.-Y. 2006. NPIC: Hierarchical Synthetic Image Classification Using Image Search and Generic Features. *Image and Video Retrieval, Lecture Notes in Computer Science.* 4071/2006, 473-482.