IST FP7 231507

D5.1 Multimedia test collection

File name	PuppyIR-D5.1-MultimediaTestCollection-v1.0.pdf
Author(s)	Karl Gyllstrom (KUL) and Carsten Eickhoff (TUD)
Work package/task	D5.1
Document status	final
Version	1.0
Contractual delivery date	30 September 2010
Confidentiality	Public
Keywords	dataset, Mechanical Turk, child-appropriateness
Abstract	Description of data sets created and applied to evaluation by WP3.

Table of Contents

Executive Summary 3

1 Introduction..... 4

2 TRUK 5

2.1 Collection process 5

2.2 Format and composition..... 5

2.3 Caveats 5

3 Granular ratings 7

3.1 Collection process 7

3.2 Format and composition..... 7

3.3 Corpus statistics..... 7

4 Images depicting interactive/engaging media types 8

4.1 Collection process 8

4.2 Format and composition..... 9

5 YouTube videos..... 10

5.1 Collection process 10

5.2 Format and composition..... 10

5.2.1 Video..... 10

5.2.2 User..... 11

5.2.3 Comment 11

5.3 Corpus statistics..... 11

References..... 12

Executive Summary

We describe a group of large datasets collected for research into various aspects of our work in the WP3.

- A collection of human annotations of the child-appropriateness of web pages, which we used to evaluate automated methods. This is directly related to work done by WP3 toward the automated detection of child-appropriateness of web pages.
- A collection of human annotations similar to the web page ratings above. These are more granular their assessments.
- A collection of annotated images. These images are categorized according to children's topic (e.g., dinosaurs) and media type (e.g., coloring pages, mazes).

This dataset is related to the following deliverables:

- **D3.1** This deliverable outlines a set of open datasets. We have made extensive use of some of these datasets, including the Simple Wikipedia (D3.1, 3.1.1), DMOZ (D3.1, 3.3.1), and ClueWeb09 (D3.1, 3.3.1).
- **D1.4** This deliverable specifies hosting of datasets to be used internally to PuppyIR to assist research. We have adhered to these guidelines in disseminating the material internally.

1 Introduction

Test collections are essential to information retrieval (IR) research. They provide standards by which IR-systems are evaluated and compared among one another. Unlike general IR, which has a substantial volume of publicly available test collections, research in children's information interaction is devoid of open datasets.

In this report, we document some datasets that have been created within PuppyIR. We list some advantages of this work. First, the creation of usable datasets has been essential to evaluation of the research we have already executed within PuppyIR. Next, we plan to release some of these datasets to the community, which itself offers two advantages: it provides positive exposure to the PuppyIR project, and it helps further others research in this essential space.

We document the following datasets:

1. A collection of child-appropriateness ratings of web pages by human users (Section 2).
2. A similar collection of granular ratings including specific age-level and topical assessments (Section 3).
3. A collection of images depicting children's topics and child-friendly media types such as mazes (Section 4).
4. A collection of ratings for existing Youtube videos indicating the child-appropriateness of the content depicted by the videos (Section 5).

The datasets outlined in this report can be used by researchers in two main ways. First, and most simply, they can serve as corpora of existing web objects (e.g., pages, images, videos) that are determined by at least one human rater to be appropriate for children. For example, these could be immediately used by a practitioner who implements or maintains a children's web search engine that constrains search results to web pages that are known or presumed to be for children. Second, it can be used to bootstrap or validate new research, including new methods to identify children's web pages, or language-based methods for which an existing data set is necessary. The latter use constitutes the use of these datasets as a test collection.

2 TRUK

The TRUK (Testing Retrieval Using Kids) data is a collection of human evaluations on the child-appropriateness of web pages; specifically, a collection of web page URLs alongside a single human assessment of the child-appropriateness of those URLs on a Likert scale. The data was originally collected for the AgeRank evaluation [8] for the purpose of verifying the accuracy of the child-appropriateness labels assigned by the algorithm. AgeRank is an algorithm that automatically assigns child-appropriateness ratings to web pages by following HTML anchor links.

2.1 Collection process

Human ratings were collected via Amazon's Mechanical Turk [2]. Mechanical Turk is an online, international marketplace spanning *Requesters* and *Workers*. Requesters submit typically small tasks, for which they make an offer of payment for completion by a Worker. An example of such a task might be asking the Worker to verify that the tag for an image is accurate. The marketplace enables such tasks to be completed quickly and inexpensively.

The collection work took place in April-May, 2010. A large number of pages within ClueWeb09 [1] were assigned child-appropriateness labels by AgeRank [8][1]. From this pool we randomly selected a large number of URLs. These URLs were sent to Mechanical Turk alongside a request for a human assessment of the URL in terms of child-appropriateness.

This assessment is of the form -3 to 3, where -3 is very inappropriate for children, 3 is very appropriate, and 0 is neutral. Users could input -4 if the URL was not available.

2.2 Format and composition

Data file is in text format, where each line contains a single page assessment. This line is of the form:

> ClueWeb node ID(tab)ClueWeb doc ID(tab)URL(tab)score

The first two columns pertain to ClueWeb notation and their documentation should be consulted if these qualities are needed. Otherwise, the URL and score are sufficient for those who are not constrained to ClueWeb09.

Ratings are limited to web pages in HTML format. We do not include ratings for multimedia objects such as images and video. This is due to the fact that such objects are not contained within the ClueWeb09 corpus.

The size of the file is approximately 156KB.

2.3 Caveats

We address some potential limitations with these data that users should consider.

The assessments in this work are time-sensitive. The web is, of course, dynamic in nature and the contents of pages to which these URLs refer may change, possibly rendering some assessments obsolete.

[1] ClueWeb09 is a massive corpora of web pages crawled during 2009. This dataset has become popular in the WWW/Information retrieval communities due to its vast size, realism, and recency.

Further, they were based on the ClueWeb crawl from 2009, so pages may have changed between the time of the crawl and the time of the assessment.

The use of Mechanical Turk is controversial due to concerns about quality of the workers, who are paid quite little for assessments. From our hand-inspection of some answers, it is clear that the workers were not always right (or at least did not share our opinion of pages). We did not collect multiple ratings per page to assess inter-rater agreement because it made more sense to use the funds for more unique ratings (with the idea that more data reduces statistical variation). We believe that collectively these ratings are good, but that individual assessments should not be considered authoritative.

3 Granular ratings

This dataset is similar in nature to TRUK, but the assessments are more granular. This was collected for evaluation in [6] (see deliverable D3.3 for a more thorough description of this work).

3.1 Collection process

A set of 1565 web pages was sampled randomly from the DMOZ web taxonomy [3][2]. and collected suitability judgements for this set. (At least 5 independent judges rated each web page). Each judge answered the following survey questions per page:

1. Is the web site suitable for children (8-12 years)?
2. Is the web site suitable for very young children (3-7 years)?
3. Is the web site's general topic interesting for children?
4. Was the web site specifically designed for children?
5. Rate the web site's quality on a scale from 1 (very bad) to 4 (very good).

3.2 Format and composition

The corpus contains the annotators' judgements in the following format: Each line represents one distinct web page consisting of:

- Column 1: The web site's suitability for children (3-12 years). 0:1
- Column 2: The web site's suitability for young children (3-6 years). 0:1
- Column 3: The web site's topic's relevance for children. 0:1
- Column 4: Was the web site designed especially for children? 0:1
- Column 5: The web site's quality. 1 (very low quality):4 (very high quality)
- Column 6: Standard deviation of page quality.
- Column 7: Inter-annotator agreement for the page's child suitability. 1:0
- Column 8: The web site's document id (identifies the relevant html file).
- Column 9: The web site's URL.

3.3 Corpus statistics

We collected a number of key statistical figures to give a deeper understanding of the corpus:

[2] DMOZ is a large collection of topics for which web searchers are be interested (e.g., *sports*). Within each category are collections of links to web pages (not from DMOZ but from the web at large) pertaining to the topic. This collection is authored and edited by a set of vetted volunteers. For research purposes, it serves as a useful organization of web pages by human-assigned categories.

Kid web page share according to judges:	0.65
Inter-annotator agreement on suitability:	0.68
Inter-annotator agreement on suitability for young children:	0.67
Inter-annotator agreement on page interestingness:	0.67
Inter-annotator agreement on page focus towards children:	0.67
Annotator agreement with DMOZ label:	0.52
Annotator agreement with DMOZ label adult:	0.45
Annotator agreement with DMOZ label kid:	0.64
Standard deviation of quality judgements per page:	0.81
Kid class precision of judgements (against DMOZ):	0.39
Kid class recall of judgements (against DMOZ):	0.64
Kid class precision of judgements (against majority vote):	0.77
Kid class recall of judgements (against majority vote):	0.7

Table 1: Corpus Statistics

4 Images depicting interactive/engaging media types

This dataset is a collection of images, with each image depicting a *topic* and *media type*. Here, a topic is a concept in which a child might be interested (e.g., dinosaurs), while media type is the nature in which the topic is depicted, spanning various interactive formats such as mazes, maps, flags, crossword puzzles, coloring pages, music sheets, etc. An example image in this collection depicts a maze in the shape of a dinosaur.

This data was collected to evaluate *collAge* [7], a system designed to complement children's search results for a query with different media types that might be appealing to children, and induce their interaction with the topic. For example, if a child issues the query "dinosaurs" and gets results such as dinosaur mazes, dinosaur treasure maps, or dinosaur anatomy depictions, their interest may be piqued to interact with those media types and engage with the topic.

4.1 Collection process

We generated a set of media-queries for which to find images representing both the query topic and one of the various media types. First, we created a set of topical queries by collecting the titles of leaf subdirectories under the top-level topic "Kids and Teens" from the *Open Directory Project* [3], which included titles such as "dinosaurs" and "Egypt". Though not necessarily reflective of queries that children would naturally generate, these queries suffice for the creation of a test set.

For each of these queries, we created a set of media queries, which are a combination of the query terms with the media types. We had a preselected set of media types (see Table 4.2 for the list) for this purpose. For example, given the query "dinosaur", we would generate the queries, "dinosaur paintings", "dinosaur coloring pages", "dinosaur crossword puzzles", etc. All of these combined queries were issued to Google's Image search, which provides a set of images as results. From each set of search results, we drew the first (i.e., most highly ranked). This enabled us to construct a set of media-query/image result pairs.

For each pair in this set, we constructed an HTML page displaying the media term(s) (e.g., "maps") and image result from Google Image search, a question asking if the image depicted an instance of the term (e.g., a map), and a "Yes/No" input form for users to record answers. These pages were uploaded to

Amazon's Mechanical Turk service, which presented the pages to human users. This produced a binary validity assessment for each media-query/image result pair. We removed from the set each entry whose assessment was negative.

4.2 Format and composition

The format of the dataset is a simple listing of URLs pertaining to web images including their media type (e.g., music sheet). The composition of the dataset is depicted in Table 4.2.

Media type	Number in dataset
music sheet	47
connect-the-dots	33
painting	128
coloring page	178
map	161
flag	98
anatomy	68
interactive game	42
maze	29
tracing page	31
word puzzle	24
crossword puzzle	21
Total	860

Table 2: Composition of image dataset.

5 YouTube videos

Audio-visual content is often more appealing for children than textual resources as it does not demand fully developed literacy skills. We created a collection of meta data entries for shared video content on YouTube [4]. The data set was collected for the evaluation phase of an automatic approach towards filtering shared video content for children [5].

5.1 Collection process

Our research corpus consists of 12,673 YouTube videos which were collected in January 2010. The crawling process was started from a range of seed queries that were deemed typical specimen of either children's topics or adult topics. For each of the queries the top 3 videos were visited and crawled. The crawling step collected the relevant meta data for feature extraction. Starting from the initial set, at each point we queued the top 5 related videos for successive crawling.

While the average video in our collection had a total of 360 user comments, there are outliers which solely feature as many as 281,571 comments for famous pieces of popular culture. In order to reduce the crawling and processing time of such videos to reasonable dimensions we capped the number of comments fetched at $\delta_{threshold} =$ 4950 comments. At that point more than 96.8% of the videos do not have additional comments. Without effecting most of the videos the computational load could thus be significantly reduced. Out of the whole collection an initial sample of 1000 videos (~50% suitable for children and ~50% for adult audiences) have been rated concerning their child-friendliness by a domain expert with a background in childcare and education.

5.2 Format and composition

The corpus does not contain any audio-visual material but consists exclusively of textual resources. The data set is stored in a relational database with a size of 887 MB. The three main entity types that were extracted are videos, users and comments.

5.2.1 Video

Each video entry consists of 11 data fields:

1. YouTube video ID
2. The video uploader's user name
3. Title
4. Description
5. Publication date
6. Number of ratings
7. Average rating
8. Number of views
9. Number of users having marked the video as "favourite"

10. Uncapped number of comments
11. Suitability label {kid|adult}

5.2.2 User

Each user entry consists of 6 data fields (some of which may be empty if the user did not provide the relevant information on his profile):

1. User name
2. Age (not mandatory)
3. Location (not mandatory)
4. Gender (not mandatory)
5. Number of subscribers to video channel
6. Total number of views for this user's content

5.2.3 Comment

Each comment entry consists of 6 data fields:

1. Unique ID
2. Video ID
3. Author name
4. Publication date
5. Textual content
6. Comment rating

5.3 Corpus statistics

The following table shows an overview of key statistical figures for this corpus:

Average video rating:	4.13
Average number of ratings per video:	1214
Average number of views per video:	559000
Average number of favourite declarations per video:	1908
Average uncapped number of comments per video:	360
Average video description length in characters:	289
Average comment length in characters:	77.82
Average number of subscribers per user:	4881
Percentage of completely filled user profiles:	90.3%

Table 3: YouTube Corpus Statistics

6 References

[1] The ClueWeb09 dataset. `http://boston.lti.cs.cmu.edu/Data/clueweb09/`.

[2] Mechanical Turk. `https://www.mturk.com/mturk/welcome`.

[3] The Open Directory Project - Kids & Teens. http://www.dmoz.org/kids_and_teens/, 2010.

[4] YouTube. http://www.youtube.com, 2010.

[5] C. Eickhoff and A. P. de Vries. Identifying Suitable YouTube Videos for Children. In *3rd Summit on Networked and Electronic Media (NEM)*, Barcelona, Spain, 2010.

[6] C. Eickhoff, P. Serdyukov, and A. P. de Vries. A Combined Topical/Non-topical Approach to Identifying Web Sites for Children. In *Proceedings of the 4th International Conference on Web Search and Data Mining (WSDM), Hong Kong, China*. ACM, 2011.

[7] K. Gyllstrom and M.-F. Moens. A picture is worth a thousand search results: finding child-oriented multimedia results with collage. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 731–732, New York, NY, USA, 2010. ACM.

[8] K. Gyllstrom and M.-F. Moens. Wisdom of the ages: toward delivering the children's web with the link-based AgeRank algorithm. In *CIKM '10*, New York, NY, USA, 2010. ACM.