Discovering User Perceptions of Semantic Similarity in Near-duplicate Multimedia Files

Raynor Vliegendhart
R.Vliegendhart@tudelft.nl

Martha Larson
M.A.Larson@tudelft.nl

Johan Pouwelse
J.A.Pouwelse@tudelft.nl

Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands

ABSTRACT

We address the problem of discovering new notions of user-perceived similarity between near-duplicate multimedia files. We focus on file-sharing, since in this setting, users have a well-developed understanding of the available content, but what constitutes a near-duplicate is nonetheless nontrivial. We elicited judgments of semantic similarity by implementing triadic elicitation as a crowdsourcing task and ran it on Amazon Mechanical Turk. We categorized the judgments and arrived at 44 different dimensions of semantic similarity perceived by users. These discovered dimensions can be used for clustering items in search result lists. The challenge in performing elicitations in this way is to ensure that workers are encouraged to answer seriously and remain engaged.

Categories and Subject Descriptors:
H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval—*Search process*

General Terms: Human Factors, Experimentation

Keywords: Near-duplicates, perceived similarity, triadic elicitation, Mechanical Turk

1. INTRODUCTION

Crowdsourcing platforms make it possible to elicit semantic judgments from users. Crowdsourcing can be particularly helpful in cases in which human interpretations are not immediately self evident. In this paper, we report on a crowdsourcing experiment designed to elicit human judgments on semantic similarity between near duplicate multimedia files. We use crowdsourcing for this application because it allows us to easily collect a large number of human similarity judgments. The major challenge we address is designing the crowdsourcing task, which we ran on Amazon Mechanical Turk, to ensure that the workers from whom we elicit judgments are both serious and engaged.

Multimedia content is semantically complex. This complexity means that it is difficult to make reliable assumptions about the dimensions of semantic similarity along which multimedia items can resemble each other, i.e., be considered near duplicates. Knowledge of such dimensions is important for designing retrieval systems. We plan ultimately to use this knowledge to inform the development of algorithms that organize search result lists. In order to simplify the problem of semantic similarity, we focus on a particular area of search, namely, search within file-sharing systems. We choose file-sharing, because it is a rich, real-world use scenario in which user information needs are relatively well constrained and users have a widely-shared and well-developed understanding of the characteristics of the items that they are looking for.

Our investigation is focused on dimensions of semantic similarity that go beyond what is depicted in the visual channel of the video. In this way, our work differs from other work on multimedia near duplicates that puts its main emphasis on visual content [1]. Specifically, we define a notion of near duplicate multimedia items that is related to the reasons for which users are searching for them. By using a definition of near duplicates that is related to the function or purpose that multimedia items fulfill for users, we conjecture that we will be able arrive at a set of semantic similarities that will reflect user search goals and in this way be highly suited for use in multimedia retrieval results lists.

The paper is organized as follows. After presenting background and related work in Section 2, we describe the crowdsourcing experiment by which we elicit human judgments in Section 3. The construction of the dataset used in the experiment is given in Section 4. Direct results of the experiment and the derived similarity dimensions are discussed in Section 5. We finish with conclusions in Section 6.

2. BACKGROUND AND RELATED WORK

2.1 Near-duplicates in search results

Well-organized search results provide an easy means for users to overview search results lists. A simple, straightforward method of organization groups together similar results and represents each group with a concise surrogate, e.g., a single representative item. Users can then scan a shorter list of groups, rather than a longer list of individual result items. Hiding near duplicate items in the interface is a specific realization of near-duplicate elimination, which has been suggested in order to make video retrieval more efficient for users [11]. Algorithms that can identify near duplicates can be used to group items in the interface. One of the challenges in designing such algorithms is being able to base them on similarity between items as it is perceived by users. Clustering items with regard to general overall similarity is a possibility. However, this approach is problematic since items are similar in many different ways at the same time [7]. Instead, our approach, and the ultimate aim of our

CrowdSearch 2012 workshop at WWW 2012, Lyon, France.

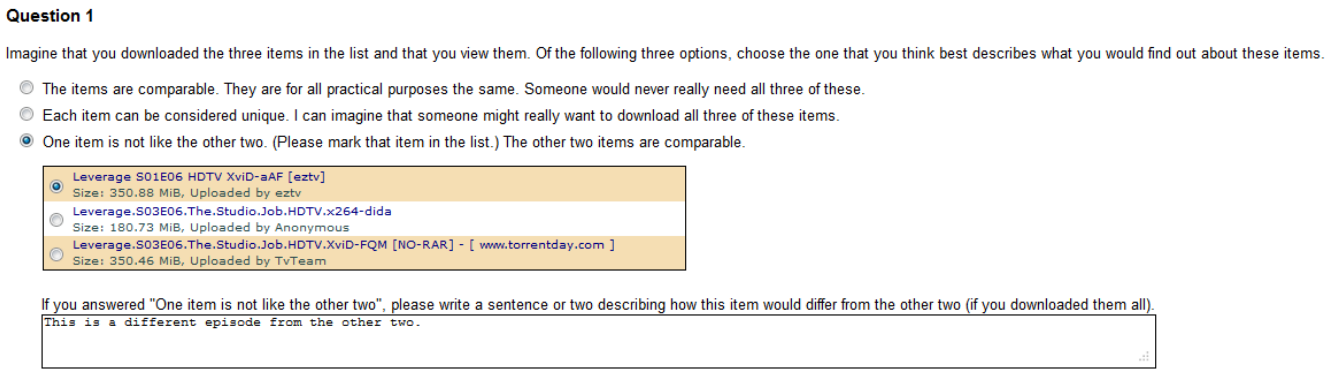

Figure 1: One of the triads of files and the corresponding question as presented to the workers.

work, is to develop near-duplicate clustering algorithms that are informed by user-perceptions of dimensions of semantic similarity between items. We assume that these algorithms stand to benefit if they draw on a set of possible dimensions of semantic similarity that is as large as possible.

Our work uses a definition of near duplicates based on the function they fulfill for the user:

> *Functional near-duplicate multimedia items are items that fulfill the same purpose for the user. Once the user has one of these items, there is no additional need for another.*

In [11], one video is deemed to be a near duplicate of another if a user would clearly identify them as essentially the same. However, this definition is not as broad as ours, since only the visual channel is considered.

Our work is related to [3], which consults users to find whether particular semantic differences make important contributions to their perceptions of near duplicates. Our work differs because we are interested in discovering new dimensions of semantic similarity rather than testing an assumed list of similarity dimensions.

2.2 Eliciting judgments of semantic similarity

We are interested in gathering human judgments on semantic interpretation, which involves the acquisition of new knowledge on human perception of similarity. Any thoughtful answer given by a human is of potential interest to us. No serious answer can be considered wrong.

The technique we use, triadic elicitation, is adopted from psychology [6], where it is used for knowledge acquisition. Given three elements, a subject is asked to specify *in what important way two of them are alike but different from the third* [8]. Two reasons make triadic eliciation well suited for our purposes. First, being presented with three elements, workers have to abstract away from small differences between any two specific items, which encourages them to identify those similarities that are essential. Second, the triadic method is found to be cognitively more complex than the dyadic method [2], supporting our goal of creating an engaging crowdsourcing task by adding a cognitive challenge.

A crowdsourcing task that involves the elicitation of semantic judgments differs from other tasks in which the correctness of answers can be verified. In this way, our task resembles the one designed in [10], which collects viewer-reported judgments. Instead of verifying answers directly, we design our task to control quality by encouraging workers to be serious and engaged. We adopt the approach of [10] of using a pilot HIT to recruit serious workers. In order to increase worker engagement, we also adopt the approach of [5], which observes that open-ended questions are more enjoyable and challenging.

3. CROWDSOURCING TASK

The goal of our crowdsourcing task is to elicit the various notions of similarity perceived by users of a file-sharing system. This task provides input for a card sort, which we carry out as a next step (Section 5.2) in order to derive a small set of semantic similarity dimensions from the large set of user-perceived similarities we collect via crowdsourcing.

The crowdsourcing task aims to achieve workers' seriousness and engagement with judicious design decisions. Our task design places particular focus on ensuring task credibility. For example, the title and description of the pilot makes clear the purpose of the task, i.e., research, and that the workers should not expect a high volume of work offered. Further, we strive to ensure that workers are confident that they understand what is required of them. We explain functional similarity in practical terms, using easy-to-understand phrases such as "comparable", "like", and "for all practical purposes the same". We also give consideration to task awareness by including questions in the recruitment task designed to determine basic familiarity with file-sharing and interest level in the task.

3.1 Task description

The task consists of a question, illustrated by Figure 1, that is repeated three times, once for three different triads of files. For each presented triad, we ask the workers to imagine that they have downloaded all three files and to compare the files to each other on a functional level. The file information shown to the workers is taken from a real file-sharing system (see the description of the dataset in Section 4) and are displayed as in a real-world system, with filename, file size and uploader. The worker is not given the option to view the actual files, reflecting the common real file-sharing scenario in which the user does not have the resources (e.g., the time) to download and compare all items when scrolling through the search results.

The first section of the question is used to determine whether it is possible to define a two-way contrast between

the three files. We use this section to eliminate cases in which files are perceived to be all the same or all different. This is following the advice on when not to use triadic elicitation that is given in [9]. Specifically, we avoid forcing a contrast in cases where it does not make sense.

The following triad is an example of a case in which a two-way contrast should not be forced:

Despicable Me The Game
VA-Despicable Me (Music From The Motion Picture)
Despicable Me 2010 1080p

These files all bear the same title. If workers were forced to identify a two-way contrast, we would risk eliciting differences that are not on the functional level, e.g., "the second filename starts with a V while the other two start with a D". Avoiding nonsense questions also enhances the credibility of our task.

In order to ensure that the workers follow our definition of functional similarity in their judgment, we elaborately define the use-case of the three files in the all-same and all-different options. We specify that the three files are the same when someone would never need all of them. Similarly, the three files can be considered to be all different from each other if the worker can think of an opposite situation where someone would want to download all three files. Note that emphasizing the functional perspective of similarity guides workers away from only matching strings and towards considering the similarity of the underlying multimedia items. Also, we intend the elaborate description to discourage workers to take the easy way out, i.e., selecting one of the first two options and thereby not having to contrast files.

Workers move on to the second section only if they report it is possible to make a two-way contrast. Here they are asked to indicate which element of the triad differs from the remaining two and to specify the difference by answering a free-text question.

3.2 Task setup

We ran two batches of Human Intelligence Tasks (HITs) on Amazon Mechanical Turk on January 5th, 2011: a recruitment HIT and the main HIT. The recruitment HIT consisted of the same questions as the regular main HIT (Section 3.1) using three triads and included an additional survey. In the survey, workers had to tell whether they liked the HIT and if they wanted to do more HITs. If the latter was the case, they had to supply general demographic information and report their affinity with file-sharing and online media consumption.

The three triads, listed below, were selected from the portion of the dataset (Section 4) reserved for validation. We selected examples for which at least one answer was deemed uncontroversially wrong and the others acceptable.

- Acceptable to consider all different or to consider two the same and one different:

  Desperate Housewives s03e17 [nosubs]
  Desperate Housewives s03e18 [portugese subs]
  Desperate Housewives s03e17 [portugese subs]

  Here, we disallowed the option of considering all files to be comparable. For instance, someone downloading the third file would also want to have the second file as these represent two consecutive episodes from a television series.

- Acceptable to consider all different:

  Black Eyed Peas - Rock that body
  Black Eyed Peas - Time of my life
  Black Eyed Peas - Alive

  Here, we disallowed the option of considering all files to be comparable as one might actually want to download all three files. For the same reason, we also disallowed the option of considering two the same and one different.

- Acceptable to consider all same or to consider two the same and one different:

  The Sorcerers Apprentice 2010 BluRay MKV x264 (8 GB)
  The Sorcerers Apprentice CAM XVID-NDN (700 MB)
  The Sorcerers Apprentice CAM XVID-NDN (717 MB)

  Here, we disallowed the option of considering all files different. For instance, someone downloading the second file would not also download the third file as these represent the same movie of comparable quality.

The key idea here is to check whether the workers understood the task and are taking it seriously, while at the same time not to exclude people who do not share a a similar view onto the world as us. To this end, we aim to choose the least controversial cases and also admit more than one acceptable answers.

We deemed the recruitment HIT to be completed successfully if the following conditions were met:

- No unacceptable answers (listed above) were given in comparing files in each triad.
- The answer to the free-text question provided evidence that the worker generalized beyond the filename, i.e., they compared the files on a functional level.
- All questions regarding demographic background were answered.

Workers who completed the recruitment HIT, who expressed interest in our HIT, and who also gave answers that demonstrated affinity with file sharing, were admitted to the main HIT.

The recruitment HIT and the main HIT ran concurrently. This allowed workers who received a qualification to continue without delay. The reward for both HITs was $0.10. The recruitment HIT was open to 200 workers and the main HIT allowed for 3 workers per task and consisted of 500 tasks in total. Each task contained 2 triads from the test set and 1 triad from the validation set. Since our validation set (Section 4) is smaller than our test set, the validation triads were recycled and used multiple times. The order of the questions was randomized to ensure the position of the validation question was not fixed.

4. DATASET

We created a test dataset of a 1000 triads based on popular content on The Pirate Bay (TPB),[1] a site that indexes content that can be downloaded using the BitTorrent [4] file-sharing system. We fetched the top 100 popular content page on December 14, 2010. From this page and further

[1] `http://thepiratebay.com`

Table 1: Dimensions of semantic similarity discovered by categorizing crowdsourced judgments

Different movie vs. TV show	Different movie
Normal cut vs. extended cut	Movie vs. trailer
Cartoon vs. movie	Comic vs. movie
Movie vs. book	Audiobook vs. movie
Game vs. corresponding movie	Sequels (movies)
Commentary document vs. movie	Soundtrack vs. corresponding movie
Movie/TV show vs. unrelated audio album	Movie vs. wallpaper
Different episode	Complete season vs. individual episodes
Episodes from different season	Graphic novel vs. TV episode
Multiple episodes vs. full season	Different realization of same legend/story
Different songs	Different albums
Song vs. album	Collection vs. album
Album vs. remix	Event capture vs. song
Explicit version	Bonus track included
Song vs. collection of songs+videos	Event capture vs. unrelated movie
Language of subtitles	Different language
Mobile vs. normal version	Quality and/or source
Different codec/container (MP4 audio vs. MP3)	Different game
Crack vs. game	Software versions
Different game, same series	Different application
Addon vs. main application	Documentation (pdf) vs. software
List (text document) vs. unrelated item	Safe vs. X-Rated

queried pages, we only scraped content metadata, e.g., filename, file size and uploader. We did not download any actual content for the creation of our dataset.

Users looking for a particular file normally formulate a query based on their idea of the file they want to download. Borrowing this approach, we constructed a query for each of the items from the retrieved top 100 list. The queries were constructed automatically by taking the first two terms of a filename, ignoring stop words and terms containing digits. This resulted in 75 unique derived queries.

The 75 queries were issued to TPB on January 3, 2011. Each query resulted in between 4 and 1000 hits (median 335) and in total 32,773 filenames were obtained. We randomly selected 1000 triads for our test dataset. All files in a triad correspond to a single query. Using the same set of queries and retrieved filenames, we manually crafted a set of 28 triads for our validation set. For each of the triads in the validation set, we determined the acceptable answers.

5. RESULTS

5.1 Crowdsourcing task

Our crowdsourcing task appeared to be attractive and finished quickly. The main HIT was completed within 36 hours. During the run of the recruitment HIT, we handed out qualifications to 14 workers. This number proved to be more than sufficient and caused us to decide to stop the recruitment HIT prematurely. The total work offered by the main HIT was completed by eight of these qualified workers. Half of the workers were eager and worked on a large volume of assignments (between 224 and 489 each). A quick look at the results did not raise any suspicions that the workers were under-performing compared to their work on the recruitment HIT. We therefore decided not to use the validation questions to reject work. However, we were still curious as to whether the eager workers were answering the repeating validation questions consistently. The repeated answers allowed us to confirm that the large volume workers were serious and not sloppy. In fact, the highest volume worker had perfect consistency.

The workers produced free-text judgments for 308 of the 1000 test triads. The other 692 triads consisted of files that were considered either all different or all similar. Workers fully agreed on which file differed from the other two for 68 of the 308 triads. Only two judgments out of the three given judgments agreed which file was different for 93 triads. For the remaining 147 triads no agreement was reached. Note that whether an agreement was reached is not of direct importance to us since we are mainly interested in just the justifications for the workers' answers, which we use to discover the new dimensions of semantic similarity.

5.2 Card sorting the human judgments

We applied a standard card sorting technique [9] to categorize the explanations for the semantic similarity judgments that the workers provided in the free-text question. Each judgment was printed on a small piece of paper and similar judgments were grouped together into piles. Piles were iteratively merged until all piles were distinct and further merging was no longer possible. Each pile was given a category name reflecting the basic distinction described by the explanations. To list a few examples: the pile containing explanations "The third item is a Hindi language version of the movie." and "This is a Spanish version of the movie represented by the other two" was labeled as *different language*; the pile containing "This is the complete season. The other 2 are the same single episode in the season." and "This is the full season 5 while the other two are episode 12 of season 5" was labeled *complete season vs. individual episodes*; the pile containing "This is a discography while the two are movies" and "This is the soundtrack of the movie while the other two are the movie." was labeled *soundtrack vs. corresponding movie*.

The list of categories resulting from the card sort is listed in Table 1. We found 44 similarity dimensions, many more than we had anticipated prior to the crowdsourcing experiment. The large number of unexpected dimensions we discovered support the conclusion that the user perception of semantic similarity among near duplicates is not trivial. For example, the "commentary document versus movie" dimension, which arose from a triad consisting of two versions of a motion picture and a text document that explained the movie, was particularly surprising, but nonetheless important for the file-sharing setting.

Generalizing our findings in Table 1, we can see that most dimensions are based on different instantiations of particular content (e.g., quality and extended cuts), on the serial nature of content (e.g., episodic), or on the notion of collections (e.g., seasons and albums). These findings and generalizations will serve to inform the design of algorithms for the detection of near duplicates in results lists in future work.

6. CONCLUSION

In this work, we have described a crowdsourcing experiment that discovers user-perceived dimensions of semantic similarity among near duplicates. Launching an interesting task with the focus on engagement and encouraging serious workers, we have been able to quickly acquire a wealth of different dimensions of semantic similarity, which we otherwise could not have thought of. Our future work will involve expanding this experiment to encompass a larger number of workers and other multimedia search settings. Our experiment opens up the perspective that crowdsourcing can be used to gain a more sophisticated understanding of user perceptions of semantic similarity among multimedia near-duplicate items.

7. REFERENCES

[1] A. Basharat, Y. Zhai, and M. Shah. Content based video matching using spatiotemporal volumes. *Computer Vision and Image Understanding*, 110(3):360–377, June 2008.

[2] P. Caputi and P. Reddy. A comparison of triadic and dyadic methods of personal construct elicitation. *Journal of Constructivist Psychology*, 12(3):253–264, 1999.

[3] M. Cherubini, R. de Oliveira, and N. Oliver. Understanding near-duplicate videos: a user-centric approach. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 35–44, New York, 2009. ACM.

[4] B. Cohen. Incentives build robustness in BitTorrent. In *Workshop on Economics of Peer-to-Peer systems*, 2003.

[5] C. Eickhoff and A. P. de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*, 15, 2012. To appear.

[6] F. Fransella and R. Bannister. *A Manual for Repertory Grid Technique*. Wiley, 2nd edition, 2003.

[7] M. A. Hearst. *Search User Interfaces*. Cambridge University Press, 1st edition, Sept. 2009.

[8] G. A. Kelly. *The Psychology of Personal Constructs, volume one: Theory and personality*. Norton, New York, 1955.

[9] G. Rugg and P. McGeorge. The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems*, 14(2):80–93, 1997.

[10] M. Soleymani and M. Larson. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. In *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, pages 4–8, 2010.

[11] X. Wu, A. G. Hauptmann, and C.-W. Ngo. Practical elimination of near-duplicates from web video search. In *Proceedings of the 15th international conference on Multimedia*, MM '07, pages 218–227, New York, 2007. ACM.