The Application of Crowdsourcing and Games to Information Retrieval

Christopher G. Harris

Informatics Program
The University of Iowa
Iowa City, Iowa 52242

christopher-harris@uiowa.edu

ABSTRACT

Crowdsourcing and games with a purpose (GWAP) have each received considerable attention in recent years. These two human computation mechanisms aid humans in solving tasks that either cannot be solved or are difficult to solve using machines. Despite this increased attention, much of this transformation has been limited to a few areas of information retrieval (IR). In this paper, we examine these two mechanisms' applicability to IR. Using a traditional, or "core" IR model, we break the model into distinct steps, evaluate the literature, and define several essential criteria for the suitability of these crowdsourcing and games to each step. After finding the most suitable steps using these criteria, we choose one step and design an experiment to empirically evaluate how these mechanisms can benefit IR tasks.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Management, Measurement, Design, Experimentation, Human Factors, Verification.

1. INTRODUCTION

1.1 Crowdsourcing

Crowdsourcing has been defined as the act of taking a job traditionally performed by a designated agent (usually an employee) and making it available to an undefined, generally large group of people in the form of an open call [11]. Although crowdsourcing is not a new concept, the recent rise in attention placed on crowdsourcing is due to several factors, including the ubiquity of the Internet, an increased need to perform tasks that computers cannot do well (such as relevance judgments, geo-tagging, and image annotations), the improved worldwide reach of micropayment methods, as well as the disparity of global economic labor demand and tight local labor restrictions. In fact, the large worker supply, little regulation, and low labor costs provide crowdsourcing's strongest advantages [4].

As with many new technologies, the early days of crowdsourcing have primarily focused on the areas with the greatest need: repetitive, single-purpose tasks designed around a simple objective, such as image classification, video annotation, form-based data entry, optical character recognition, translation, and document proofreading. However, as crowdsourcing begins to mature, it has begun a transformation, creating fascinating new opportunities for leveraging real-time human computation for a range of diverse tasks.

Since Amazon.com introduced Mechanical Turk[1] in 2005, this mechanism has become a phenomenon in academic research in many disciplines as well. Consider language translation: an examination of the number of articles returned by a Google Scholar search for "language translation" shows slow but steady growth over the period 2006-2011. A similar search combining crowdsourcing-related terms combined with the search phrase "language translation" shows exponential growth over the same period (see Figure 1). This may indicate that from 2006-2011 research in language translation expanded slowly, but an increasing share of this research involved crowdsourcing.

In effect crowdsourcing has not only facilitated other research, but has become a research area of its own. This presents both new challenges and new opportunities in this intersection of people and technology. Design components such as human factors and human-computer interaction are essential to successful crowdsourcing tasks, as are consideration of psychology, law, ethics, systems testing, etc. – all of which need to be integrated with existing principles of computer architecture and application design. This may in turn provide functionality or accuracy not previously thought achievable.

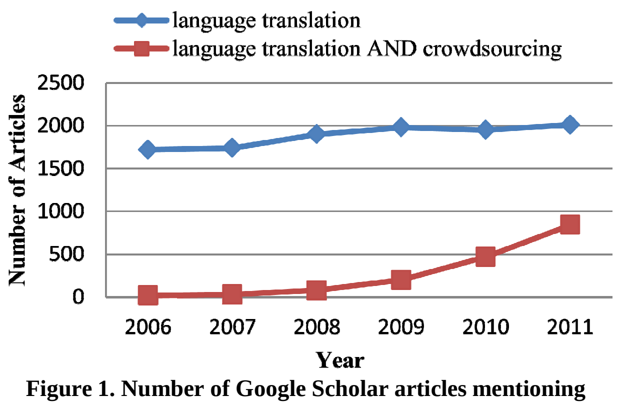

Figure 1. Number of Google Scholar articles mentioning "language translation" and "language translation" AND "crowdsourcing" by year, 2006-2011

[1] http://www.mturk.com

Frequent uses for crowdsourcing in IR tasks include relevance judgments, classification tasks, labeling and annotating text and images. Since document relevance is a highly-subjective task, redundant judgments by different assessors are often designed into the task to reduce subjectivity and bias. Crowdsourcing addresses this need as judgments can be performed inexpensively, quickly and – when the relevance task is designed well – with high quality. Likewise, classification tasks also rely on the abilities of the worker to correctly group items, so the diversity of crowdsourcing is an advantage. Finally, labeling and annotating text is often a tedious exercise, making it prone to errors. Crowdsourcing can address these potential errors by permitting additional labels to be generated through several members of the crowd, and spreading out the workload to more people, reducing the tedium. An example of this is LabelMe [28], a tool that provides for the labeling of images; several workers annotate a single image, and these annotations are then merged to create a single metadata file.

Overall, we observe that crowdsourcing is an exciting area that involves a mix of different disciplines. Although the discussion in this paper primarily focuses on how crowdsourcing and games can improve IR-related tasks, there are many other disciplines where potential benefits are possible, such as was attained in collaborative filtering algorithms with the 2009 Netflix prize.

This paper is organized as follows. In the next section, we cover related work in this area. In Section 3, we introduce a core IR model and describe each of the model's steps. In Section 4, we identify the research questions we will examine in our study and ascertain the significant problems that exist in using crowdsourcing and games for IR tasks. In Section 5, we outline our planned research, including a discussion of our research methodology, anticipated challenges, and the expected contributions of our work. We conclude our work in Section 6.

2. RELATED WORK

Although the quantity of research conducted on the utility of crowdsourcing and game mechanisms has grown rapidly in recent years, few attempts have been made to classify the types of tasks these mechanisms address. A 2011 study by Quinn and Bederson classified human computation systems, including crowdsourcing [24] but did not describe the tasks these systems address in detail. Yuen et. al. perform a more comprehensive survey on tasks in crowdsourcing [37] and a similar survey of serious games in [36]. In each survey, the authors provide a good taxonomical classification, but their taxonomy does not examine their utility with different systems, specifically IR systems.

IR models have been discussed extensively in the literature (e.g., [3, 7, 22]); however, none of these discussions address how crowd and game-based mechanisms can add value to areas of an IR model.

2.1 Crowdsourcing

There have been several recent tutorials on social computing that examine the application of crowdsourcing to areas of IR. Alonso has presented some guidelines for conducting studies using crowdsourcing platforms [1, 2]. Likewise, Ipeirotis [14] has provided some useful insight into study design. However, these discussions primarily focus on crowdsourcing studies in the most high-demand tasks: labeling, translations and transcriptions, relevance judgments, and classification. The last two are elements of our model and their work is incorporated into our research design.

Lease and Alonso also conducted several workshops that address specific issues with crowdsourcing, serious games, and IR, such as addressing spam, incentive models, and quality control, among others [1, 5, 18]. The Human Computation workshops examine issues in human computation, a broader classification of human-machine collaboration that includes both games and crowdsourcing [12, 13]. The topics covered in these workshops, although broad, provide useful insight that we employ in our research methodology.

2.2 Serious Games

A recent research study indicated there are more than half a billion people worldwide playing online games at least an hour a day, 183 million in the US alone [23]. Another study has estimated that the average American has played 10,000 hours of video games by the age of 21, equivalent to five years of working a full-time job 40 hours per week [25]. What if some of this time and energy could somehow be channeled into productive work? And better yet, what if people playing computer games could, without consciously doing so, simultaneously solve large-scale computation problems?

Serious games are human-based computation mechanisms designed to solve a specific problem outside of pure entertainment [26, 29]. Although designed to be entertaining, their main purpose usually falls within one of three categories: to train, to investigate, or to advertise [9]. They are denoted "serious games" to differentiate them from games designed strictly for leisure. A subset of serious games, designed to be played by the crowd through an open call, is denoted games with a purpose (GWAP). In this abstract, we refer to "GWAP" as "serious games" or simply, "games".

Some examples of serious games include the ESP Game [31], considered the first human computation system. It has been subsequently adopted as the Google Image Labeler. Its objective is to collect image labels for Google Images. In addition to basic image annotation, Peekaboom [35] is designed to have players determine object locations within images. Phetch [32, 33] is designed to enhance image descriptions and therefore improve web accessibility, particularly with image searches. In a similar way, Matchin [10] helps search engines rank images that fit a given set of criteria. These games improve the inputs to IR, and could be akin to the preprocessing step for non-text retrieval.

The underlying concept of the ESP Game has been applied to other aspects of information retrieval. For instance, TagATune [17] provides annotation for sounds and music, which in turn can improve searches for audio clips. Verbosity [34] and Common Consensus [20] collect "common-sense" knowledge. This common-sense information can be provided as input to situations involving reasoning and is a step towards enhancing interactive user interface design.

More directly relevant to IR, Page Hunt [21] is a single-player game that displays a random web page to a player; the player's task is to come up with query terms and operators that would bring up the web page in the top few results (e.g., the top 5) of a search engine. Page Race and Page Match are two variations of the Page Hunt game that are two-player competitive and collaborative games, respectively. These games provide valuable feedback on query design.

3. INFORMATION RETRIEVAL MODEL

In this section, we introduce an IR model and describe each of the model's stages. We then examine the applicability of crowdsourcing and games to each stage.

A "core" definition of information retrieval (IR) is the science of finding relevant material that satisfies an information need from within a large collection. This searchable material may be generated either by users or through computer applications. Although IR has its origins in finding relevant text, in recent years the scope has expanded to multimedia search (e.g., [6, 19], image search (e.g., [15, 27]), and audio search (e.g., [8, 30]), among others. The collections normally comprise minimally-structured or semi-structured data, which contrasts with structured data searches typically performed in relational databases.

IR also covers other kinds of data and functions beyond what is specified in the core definition provided above. For example, it covers browsing or filtering document collections and annotation tasks (such as the addition of metadata).

Figures 2 and 3 illustrate the steps of our core IR model. Similar IR models have been developed, such as that by Lancaster and Warner [16]. This model illustrates a typical process for establishing an IR application for searching a document collection. Although there are several other aspects to IR not shown in this core IR model, (e.g., document translation); the model forms a reasonable starting point for examining the fit of crowdsourcing and GWAP. Steps 1-6 (Figure 2) designate the IR system design and implementation (preparatory stage) and steps 7-12 (Figure 3) designate the user query processing (interactive stage). Later, we will examine each of these steps in more detail, including each step's objective and the expected result or output.

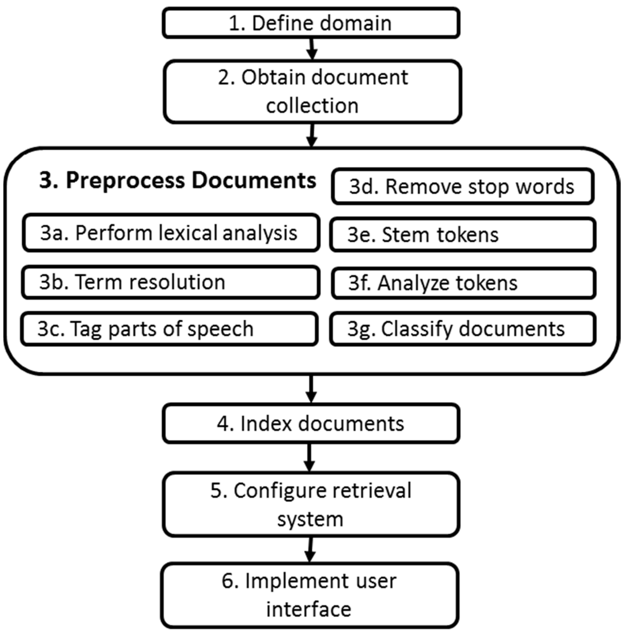

Figure 2. Designing the retrieval system application.

From Figure 2, we can make several observations. First, the domain, which contains the boundaries and nature of the task, is defined. Next, we identify and obtain a document collection. Then in Step 3, we preprocess the collection's documents prior to indexing. This preprocessing consists of several steps, including lexical analysis, term resolution (resolving abbreviations and acronyms), part-of-speech tagging, stop word removal, stemming and analysis of tokens, and the classification of documents. Once preprocessing is completed, we determine the appropriate indexing strategy, essential configuration parameters and index the preprocessed document collection, and configure the retrieval system application. The last aspect of the system design is the creation of the search interface.

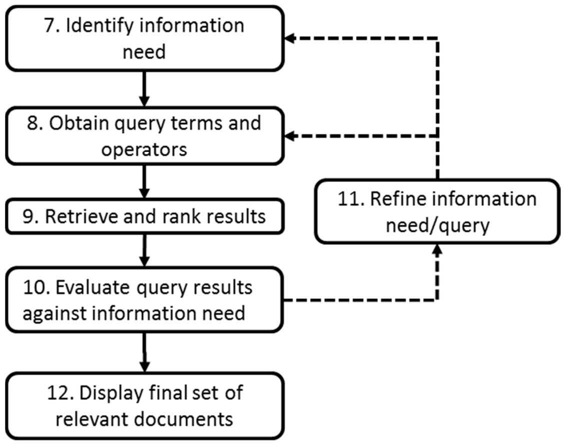

Figure 3. Handling a user query

Once the system has been implemented, it is ready for searches. Figure 3 illustrates the steps involved to handle a query. The user defines her or his information need and provides search terms and operators through the search interface. The interface passes this information to the search engine, which retrieves and ranks the results. The user would then evaluate the query results against her or his information need; if it was not what the user had anticipated, she or he would refine the query terms and re-issue the query until satisfied with the results. Our model then enters its final stage, and the final ranked list is displayed. These steps illustrated are similar to a traditional relevance feedback model as described in the literature, (e.g., [3]).

4. RESEARCH QUESTIONS

Given the model described in Section 3, we wish to examine how applicable crowdsourcing and game mechanisms are to each step. Therefore, the objective of our study is to examine the following research questions:

1. Which steps of this core IR model are suitable for crowdsourcing? For game-based formats?
2. What criteria are important to determine this suitability? Are all criteria equally important?
3. For the steps we identify as highly suitable for either crowdsourcing or game-based formats, why haven't they yet been exploited?
4. How can we identify and incorporate game design techniques into our tasks to improve quality?

These four research questions address some of the issues facing this research area – a definition of the types of tasks to which crowdsourcing and/or games can apply. The rapid growth of crowdsourcing and game-based approaches to solving human computation issues has been impressive, but this growth has not been evenly applied to IR. We infer from this uneven growth and from our own knowledge of IR that not all aspects of our model are equally suitable for crowdsourcing and game-based formats.

Thus, our first question examines why the application of these mechanisms is not more uniformly distributed among the steps. The second research question addresses the need for a systematic way of determining fit for each mechanism. We believe the best way to arrive at this determination of fit is through an objective set of criteria and weighting on this criteria. Our third question studies this fit and empirically examines each step of our IR model through previous research and our own experiments. Our fourth question examines the ability to design a GWAP that aligns game objectives with task objectives. These four questions, taken together, can guide researchers to determine the applicability of crowdsourcing and games to a specific task.

5. PROPOSED METHODOLOGY

In this section, we discuss our research methodology, anticipated challenges, and the expected contributions of our work.

5.1 Step Suitability Determination

To provide some insight into our first research question, we initially review the literature for studies involving crowdsourcing and games as applied to each of our identified steps. Finding a large amount of related research that relates to a step may likely provide insight into the criteria that may be used to determine suitability.

One notable observation is much of the current research in crowdsourcing has been applied to areas where computers perform poorly, such as obtaining document collections (Step 2) and evaluating query results against an information need (Step 10). There are also a few research studies that use crowdsourcing to identify an information need (Step 7) and to obtain query terms and operators (Step 8).

With games, the scope of existing research is even narrower, with fewer examples to borrow from. As with crowdsourcing, evaluating query results against an information need (Step 10) is relatively well-represented in the literature, but there is little previous work in the use of games to obtain a document collection (Step 2). There is limited work in the classification preprocessing step (Step 3g) and identify information need (Step 7) and obtaining query terms and operators (Step 8).

Examining this further, we see that there are some criteria that are essential to crowdsourcing and game design, some that are beneficial, and some additional criteria that only apply to game-based formats. We discuss these criteria in the next section.

5.2 Evaluation Criteria

Our approach to determining crowdsourcing and game applicability to each step is to initially apply mandatory criteria for each step. These criteria are determined by making a common-sense evaluation of the process involved in each step as well as a review of relevant literature. We regard Step 12 as a terminal step and do not evaluate it further. Our two mandatory criteria are:

<u>Criterion 1:</u> Can the mechanism (either crowdsourcing or games) handle the scale of the task?

To illustrate our scalability requirement, consider one of the preprocessing tasks – stemming tokens (Step 3e). In this step, we "stem" or reduce inflected or derived terms to a stem, base or root form, usually through the use of a stemming algorithm. To accomplish this step using the crowd or a game, millions of tokens would need to be evaluated and stemmed in a similar way. Clearly, machines can perform this step far more efficiently and so the stemming step would fail our scalability test. Thus, all seven of the preprocessing steps (Steps 3a - 3g) do not scale for crowdsourcing or game design, given the large number of items to be evaluated. Indexing (Step 4) and the ranking and retrieval (Step 9) also fail our scalability test.

<u>Criterion 2:</u> Does the mechanism require specialized or local knowledge to complete?

A step may require extensive local knowledge, such as an understanding of user expectations of the IR system, the existing and expected system capabilities, or other local constraints that neither the crowd nor game players could be made aware of in a reasonable amount of time. The domain definition (Step 1) usually requires system designers with considerable knowledge of the local system. Likewise, the configuration of the retrieval system application (Step 5) requires specialized knowledge of IR systems, such as the ability to tune and configure parameters and develop a search strategy. Thus, these are unlikely to benefit from an "open call" and are subsequently eliminated from further consideration.

For the steps not eliminated by the scalability and localized or specialized knowledge tests, we consider the following criteria:

<u>Criterion 3</u>: Can the outputs provided by the mechanism be effectively and efficiently tested and integrated into the IR model?

The following two additional criteria apply only to games:

<u>Criterion 4</u>: Can the mechanism be designed to be entertaining yet accomplish the objectives of the task?

<u>Criterion 5</u>: Can the mechanism be designed to provide an evaluation of performance and a score aligned with the task's objective?

Each of these questions is designed to be answered in a "yes"/"no" format, although Criteria 4 and 5 are not applicable for crowdsourcing and therefore we mark it "N/A".

With Criterion 3, we wish to determine if the mechanism outputs can effectively be tested and integrated into our model without affecting the inputs of the steps occurring later in our model. Criterion 4 determines the potential of making the task engaging or entertaining. Criterion 5 evaluates if it is possible to score and reward performance in "real time" for proper execution in a task.

5.3 Human Value-Added Computation Model

Our initial test, which examined if each step in our IR model could scale using the crowd and games, was evaluated based on having either the crowd or the game complete all work associated with the task. Perhaps the humans could provide a supplemental role, like handling the most difficult components, or provide quality control? In other words, what if we took a human computation approach to those tasks, where computers and humans each apply their strengths in a hybrid-like fashion?

To evaluate these questions, we consider a "human value-added" criterion on those steps we eliminated for their inability to scale (Criterion 1). This new criterion evaluates whether the crowd or game mechanism can be designed to have humans add value to the task. We believe several preprocessing tasks could benefit from having humans take on those parts of the step that machines cannot complete, such as term resolution (Step 3b) or those steps where a human quality control effort could provide value, such as the classification step (Step 3g). The discussion of how each criterion applies to each step is not included here due to space considerations.

In our research going forward, we plan to examine the applicability of crowdsourcing and games to these areas and see (1) if it is possible to expand crowdsourcing and games to the IR steps in our model, (2) how easy or difficult this can be for each approach, and (3) which of the two approaches (crowdsourcing or games) is the more suitable for a particular step.

The task to refine the information need or query (Step 11) meets our criteria nicely for both crowd and game mechanisms. There are only a few instances where these mechanisms have been used to satisfy this step. We wish to compare how different agents perform the query reformulation task and if crowdsourcing or games might lead to better retrieval performance.

We plan to evaluate several hypotheses that involve the four research questions discussed in Section 4. We plan to evaluate and compare the following agents in our studies:

- layperson or "casual user"
- librarian or "expert user"
- crowdsourcing worker or "crowdworker"
- algorithms

We will examine the effects of initial queries against reformulated queries for these four types of agents, and examine how game-based mechanisms can improve the performance of meeting information needs for three of them: casual users, expert users and crowdworkers. We plan to use topics (modified to represent information needs) and data collections provided in the TREC 2004 Robust task. Our intent is to measure retrieval performance using the following metrics: recall, precision, and F-score.

5.4 Expected Contributions

The anticipated contributions of our research are as follows. First, using a traditional IR model, we identify the objectives and outputs of each step. Second we evaluate the crowdsourcing and game-based fit for tasks at each step of our model, using criteria we develop from empirically evaluating other work, as well as observing the characteristics that may impede these mechanisms. Third, we introduce several research questions that will guide our evaluation of the steps in our model that appear suitable for crowd or game involvement yet remain underrepresented in research..

5.5 Anticipated Challenges

The following are challenges we anticipate with the study mentioned in Section 5.3. First, we need to ensure the criteria are comprehensive and meaningful. Second, we need to ensure the samples chosen for each agent group are representative of the population. We also need to interleave the experiments to remove testing and selection bias.

6. CONCLUSION

We have introduced crowdsourcing and serious games as mechanisms to accomplish IR-related tasks. We have defined and described a core IR model, to which we plan to examine the suitability of applying crowdsourcing and games criteria. We also plan to further investigate additional crowdsourcing and game-based research that has been applied to each step of this core IR model. We then discussed some future research designs we anticipate applying to the relationship between our IR model and both crowdsourcing and game-based mechanisms.

Although every step of our IR model may not suitable for game-based or crowdsourcing mechanisms, we do believe there are significant gains that can be obtained. If we can demonstrate significant gains using these mechanisms, we believe this contribution will provide a lasting benefit to Information Science.

7. ACKNOWLEDGMENTS

I appreciate my dissertation advisor, Padmini Srinivasan, as well as my committee members for their contributions and suggestions for this work. I would also like to thank the reviewers and organizers of JCDL 2012 Doctoral Consortium for their valuable feedback and support.

8. REFERENCES

[1] Alonso, O. and Lease, M. Crowdsourcing 101: putting the WSDM of crowds to work for you. *Proceedings of the fourth ACM international conference on Web search and data mining* (WSDM '11). ACM, New York, USA, pp. 1-2. DOI=10.1145/1935826.1935831

[2] Alonso, O. and Lease, M. Crowdsourcing for information retrieval: principles, methods, and applications. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (SIGIR '11). ACM, New York, USA, pp. 1299-1300. DOI=10.1145/2009916.2010170

[3] Baeza-Yates, R. and Ribeiro-Neto, B. *Modern information retrieval*. Addison-Wesley New York, 1999.

[4] Bederson, B. B. and Quinn, A. J. Web workers unite! addressing challenges of online laborers. *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems* (CHI EA '11). ACM, New York, NY, USA, pp. 97-106. DOI=10.1145/1979602.1979606

[5] Carvalho, V. R., Lease, M. and Yilmaz, E. Crowdsourcing for search evaluation. *SIGIR Forum* 44, 2 (January 2011), 17-22. DOI=10.1145/1924475.1924481

[6] Chang, S. F., Huang, Q., Huang, T., Puri, A. and Shahraray, B. Multimedia search and retrieval. *Multimedia Systems, Standards, and Networks*1999), 559-584.

[7] Croft, W. B., Metzler, D. and Strohman, T. *Search engines: Information retrieval in practice*. Addison-Wesley, 2010.

[8] Foote, J. An overview of audio information retrieval. *Multimedia Systems*, 7, 1 1999, pp. 2-10.

[9] Groh, F. Gamification: State of the Art Definition and Utilization. *Proceedings of the 4th Seminar on Research Trends in Media Informatics* (RCMI'12), pp. 39-46.

[10] Hacker, S. and Von Ahn, L. Matchin: eliciting user preferences with an online game. *Proceedings of the 27th international conference on Human factors in computing systems (CHI '09)*. ACM, New York, USA, pp. 1207-1216. DOI=10.1145/1518701.1518882.

[11] Howe, J. *The Rise of Crowdsourcing*. Wired, June, 2006.

[12] Ipeirotis, P., Chandrasekar, R. and Bennett, P. *Proceedings of the ACM SIGKDD Workshop on Human Computation*. ACM, New York, NY, USA. 2010. pp. 1-2.

[13] Ipeirotis, P. G., Chandrasekar, R. and Bennett, P. A report on the human computation workshop (HComp 2009). *SIGKDD Explor. Newsl.* 11, 2 (May 2010), pp. 80-83. DOI=10.1145/1809400.1809416

[14] Ipeirotis, P. G. and Paritosh, P. K. Managing crowdsourced human computation: a tutorial. *Proceedings of the 20th international conference companion on World wide web* (WWW '11). ACM, New York, USA, pp. 287-288. DOI=10.1145/1963192.1963314

[15] La Cascia, M., Sethi, S. and Sclaroff, S. *Combining Textual and Visual Cues for Content-Based Image Retrieval on the World Wide Web*. Technical Report. Boston University, Boston, MA, USA.

[16] Lancaster, F. W. and Warner, A. J. *Information Retrieval Today*. Information Resources Press, Arlington, VA, USA, 1993.

[17] Law, E. and Von Ahn, L. Input-agreement: a new mechanism for collecting data using human computation games. *Proceedings of the 27th international conference on Human factors in computing systems* (CHI '09). ACM, New York, USA, pp. 1197-1206. DOI=10.1145/1518701.1518881

[18] Lease, M. and Yilmaz, E. Crowdsourcing for information retrieval. *SIGIR Forum* 45, 2 (January 2012), 66-75. DOI=10.1145/2093346.2093356

[19] Lew, M. S., Sebe, N., Djeraba, C. and Jain, R. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2, 1 2006, pp. 1-19.

[20] Lieberman, H., Smith, D. and Teeters, A. Common Consensus: A Web-based Game for Collecting Commonsense Goals. *Workshop on Common Sense for Intelligent Interfaces, ACM International Conference on Intelligent User Interfaces* (IUI-07), Honolulu, January 2007.

[21] Ma, H., Chandrasekar, R., Quirk, C. and Gupta, A. *Improving search engines using human computation games*. Improving search engines using human computation games. *Proceedings of the 18th ACM conference on Information and knowledge management* (CIKM '09). ACM, New York, USA, 275-284. DOI=10.1145/1645953.1645990

[22] Manning, C. D., Raghavan, P. and Schutze, H. *Introduction to information retrieval*. Cambridge University Press Cambridge, UK, 2008.

[23] McGonigal, J. *Reality is broken: Why games make us better and how they can change the world*. Penguin Press, 2011.

[24] Quinn, A. J. and Bederson, B. B. Human computation: a survey and taxonomy of a growing field. *Proceedings of the 2011 annual conference on Human factors in computing systems (CHI '11*). ACM, New York, USA, pp. 1403-1412. DOI=10.1145/1978942.1979148

[25] Richards, C. *Teach the world to twitch: An interview with Marc Prensky, CEO and founder Games2train. com. Futurelab (Dec. 2003)*. New York, NY, USA.

[26] Ritterfeld, U., Cody, M. J. and Vorderer, P. *Serious games: Mechanisms and effects*. Taylor & Francis, 2009.

[27] Rui, Y., Huang, T. S. and Chang, S. F. Image retrieval: Current techniques, promising directions, and open issues. *Journal of visual communication and image representation*, 10, 1 1999, pp. 39-62.

[28] Russell, B. C., Torralba, A., Murphy, K. P. and Freeman, W. T. LabelMe: a database and web-based tool for image annotation. *International journal of computer vision*, 77, 1 2008, pp. 157-173.

[29] Susi, T., Johannesson, M. and Backlund, P. Serious games– An overview. *Skövde: University of Skövde (Technical Report HS-IKI-TR-07-001*. 2007.

[30] Typke, R., Wiering, F. and Veltkamp, R. C. A survey on music information retrieval systems. *Proceedings of the International Conference on Music Information Retrieval*, 2005, pp 153–160.

[31] von Ahn, L. and Dabbish, L. Labeling images with a computer game. *Proceedings of the SIGCHI conference on Human factors in computing systems* (CHI '04). ACM, New York, USA, pp. 319-326. DOI=10.1145/985692.985733

[32] von Ahn, L., Ginosar, S., Kedia, M. and Blum, M. Improving image search with phetch. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (Honolulu, Apr. 15-20). IEEE Press, New York, 2007, pp. 1209- 1212.

[33] von Ahn, L., Ginosar, S., Kedia, M., Liu, R. and Blum, M. Improving accessibility of the web with a computer game. *Proceedings of the SIGCHI conference on Human Factors in computing systems* (CHI '06), Rebecca Grinter, Thomas Rodden, Paul Aoki, Ed Cutrell, Robin Jeffries, and Gary Olson (Eds.). ACM, New York, USA, pp. 79-82. DOI=10.1145/1124772.1124785

[34] von Ahn, L., Kedia, M. and Blum, M. Verbosity: a game for collecting common-sense facts. *Proceedings of the SIGCHI conference on Human Factors in computing systems* (CHI '06), Rebecca Grinter, Thomas Rodden, Paul Aoki, Ed Cutrell, Robin Jeffries, and Gary Olson (Eds.). ACM, New York, USA, pp. 75-78. DOI=10.1145/1124772.1124784

[35] von Ahn, L., Liu, R. and Blum, M. Peekaboom: a game for locating objects in images. *Proceedings of the SIGCHI conference on Human Factors in computing systems*(CHI '06), Rebecca Grinter, Thomas Rodden, Paul Aoki, Ed Cutrell, Robin Jeffries, and Gary Olson (Eds.). ACM, New York, USA, pp. 55-64. DOI=10.1145/1124772.1124782

[36] Yuen, M. C., Chen, L. J. and King, I. A Survey of Human Computation Systems. *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04* (CSE '09), Vol. 4. IEEE Computer Society, Washington, DC, USA, pp. 723-728. DOI=10.1109/CSE.2009.395

[37] Yuen, M. C., King, I. and Leung, K. S. A survey of crowdsourcing systems. *SocialCom '11: Proceedings of The Third IEEE International Conference on Social Computing.* IEEE Computer Society, 2011. IEEE, pp. 766 – 773.