# Applying Human Computation Mechanisms to Information Retrieval

**Christopher G. Harris**
Informatics Program
The University of Iowa
Iowa City, Iowa 52242
christopher-harris@uiowa.edu

**Padmini Srinivasan**
Computer Science Department and Informatics Program
The University of Iowa
Iowa City, Iowa 52242
padmini-srinivasan@uiowa.edu

## ABSTRACT

Crowdsourcing and Games with a Purpose (GWAP) have each received considerable attention in recent years. These two human computation mechanisms assist with tasks that cannot be solved by computers alone. Despite this increased attention, much of this transformation has been limited to a few aspects of Information Retrieval (IR). In this paper, we examine these two mechanisms' applicability to IR. Using an IR model, we apply criteria to determine the suitability of these crowdsourcing and GWAP mechanisms to each step of the model. Our analysis illustrates that these mechanisms can apply to several of these steps with good returns.

### Keywords

Crowdsourcing, information retrieval, GWAP, games with a purpose, human computation

## INTRODUCTION

Crowdsourcing, 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, is not a new concept; however, it has received considerable attention recently. Although considered inexpensive when contrasted with traditional workers, crowdworkers are still expensive relative to machine-based computation methods. The appeal of crowdworkers has largely been fostered by the increased need to perform tasks that computers cannot do at any price (such as relevance judgments, geo-tagging, and image annotations) or can only do imperfectly. Other factors influencing the growth of crowdsourcing include the ubiquity of the internet, the improved worldwide reach of micropayment methods, as well as the disparity of global economic labor demand and tight local labor restrictions.

In addition to the crowdworkers, there are more than half a billion people worldwide playing online games at least an hour a day, 183 million in the US alone. It is estimated that the average American has played 10,000 hours of video games by the age of 21 (McGonigal 2011) - 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? This redirection of a user's time and energy from pure entertainment to accomplishing a task while being entertained is a principle factor behind the GWAP development and use. GWAP is defined as a game (or set of games) played on a computer that serve an intrinsic or extrinsic purpose for the game's provider by harnessing human computation abilities in an entertaining setting. Unlike computer processors, humans need to be given an inducement or incentive to become part of a collective computation. Thus, GWAP are a powerfully seductive method for encouraging people to participate in this human computation process. These GWAP, considered a subset of the genre of games called serious games, are not mutually exclusive from crowdsourcing tasks - particularly if financial compensation is involved; however, a vast number of GWAP participants are compensated solely through entertainment in lieu of a financial payment.

The early days of crowdsourcing and GWAP have primarily focused on the areas with the greatest need: tasks which make use of knowledge or skills most humans have, but that computers are unable to duplicate or imitate. In this paper, we examine the application of the crowd and GWAP to each step of an IR model. In the next section, we build on the "core" definition of IR by introducing an IR model. Next, we describe each of the model's stages, and assess the applicability of crowdsourcing and GWAP to each step through the use of criteria.

## CORE IR MODEL

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, image search, and audio search, among others. The collections normally comprise minimally-structured or semi-structured data, which contrasts with structured data searches typically found in relational databases.

*ASIST 2012,* October 28-31, 2012, Baltimore, MD, USA.

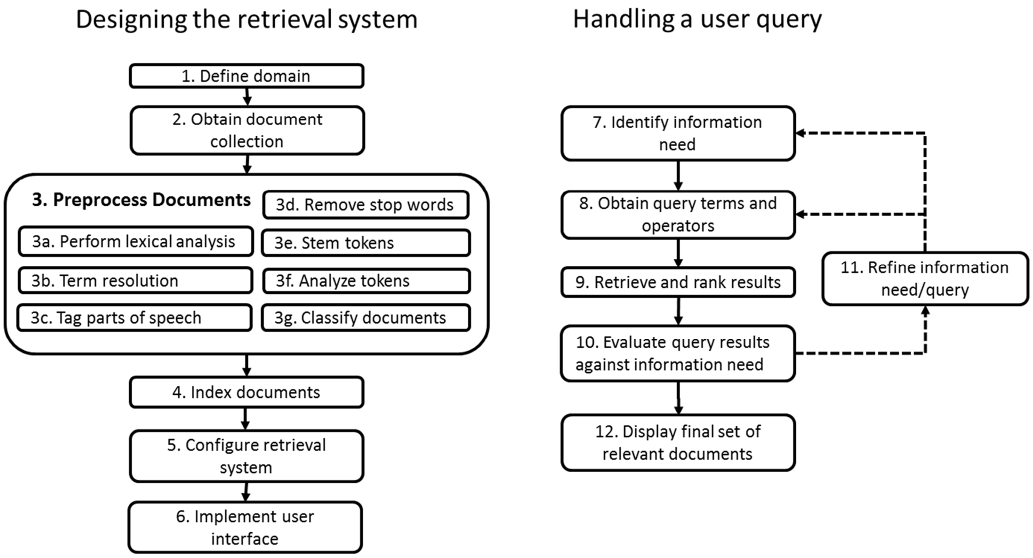
Designing the retrieval system
1. Define domain
2. Obtain document collection
3. Preprocess Documents
3a. Perform lexical analysis
3b. Term resolution
3c. Tag parts of speech
3d. Remove stop words
3e. Stem tokens
3f. Analyze tokens
3g. Classify documents
4. Index documents
5. Configure retrieval system
6. Implement user interface
Handling a user query
7. Identify information need
8. Obtain query terms and operators
9. Retrieve and rank results
10. Evaluate query results against information need
11. Refine information need/query
12. Display final set of relevant documents

**Figure 1. Steps of the Core IR Model**

Figure 1 illustrates the steps of our core IR model. Similar models have been developed as in (Lancaster and Warner, 1993). This model illustrates a typical process for establishing and 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 of the model designate the IR system design and implementation (preparatory stage) and Steps 7-12 designate the user query processing (interactive stage). In this paper, we will examine each step's objective and the expected result or output in detail. This model is then used to examine the applicability of crowdsourcing and GWAP to each step.

The first step of our model is to define the domain, containing the boundaries and nature of the retrieval task. Next, we identify and obtain a suitable document collection. Then we preprocess the collection's documents prior to indexing. This preprocessing typically consists of several steps, including lexical analysis, term resolution, 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. The last aspect of the system design is the creation of the retrieval system search interface.

Once this IR system has been implemented, it is ready for searches. The user defines her or his information need and provides search terms and operators through the search interface. The search 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 (or perhaps even modify the information need) until satisfied with the results. Our model then enters its final stage, displaying the final ranked list.

Tables 1 and 2 contain the task objective and expected output for each step, along with pointers to recent work in each area. In the "Step No." column of these two tables, we also define who is typically responsible for each step according to the following: (SD) – System Designer, (SYS) –System, (U) – User. Next, we assess the applicability of our mechanisms to each step.

## ASSESSING CROWD AND GWAP APPLICABILITY

Our approach to determining crowdsourcing and GWAP applicability to each step is to initially examine two criteria. We do not evaluate Step 12 of our model, since this represents a terminal step.

### Initial Assessment

Our assessment begins by examining two criteria for each step.

**Criterion 1:** *Can the mechanism (either crowdsourcing or GWAP) handle the scale of the task?*

To illustrate this scalability requirement, consider one of the preprocessing tasks – stemming tokens (Step 3e). To accomplish this step using the crowd or a GWAP, millions of tokens would need to be evaluated and stemmed in a standardized manner. Clearly, machines perform this step far more efficiently and so the stemming step would not meet our scalability criterion. Indexing (Step 4) and the ranking and retrieval (Step 9) also do not meet our scalability criterion.

| Step No | Step Description | Task/Objective | Expected Output |
|---|---|---|---|
| **1 (SD)** | Define domain | Define the scope of information for retrieval tasks | A definition of the domain (boundaries and nature of the contents) applicable to task |
| **2 (SD)** | Obtain document collection | Determine the documents to be indexed and available for retrieval. Web spidering, web link analysis, etc., could be done at this stage. | List of documents or document sources, and, if applicable, a set of inter-document links. |
| **3 (SD)** | Preprocess documents | Provide data enrichment | Contains multiple objectives and expected outputs. See **Table 2** for additional details. |
| **4 (SYS)** | Index documents | Create an index for all documents | A set of indexed documents |
| **5 (SD)** | Configure retrieval system | Determine the best search strategies and parameters for an anticipated set of tasks | Best search strategy and parameters within a chosen retrieval system for the collection |
| **6 (SD)** | Implement user interface | Have an interface that users can use to meet their needs | A usable interface to all anticipated user group interaction with retrieval documents |
| **7 (U)** | Identify information need | Identify the user's information need | Information need |
| **8 (U)** | Obtain query terms and operators for the user's search | Obtain the initial query terms and any operators (e.g., Boolean) and apply them. | Initial query terms and operator weights |
| **9 (SYS)** | Retrieve and rank user query results | Provide a ranked set of relevant documents based on the submitted query | Ranked set of retrieval documents |
| **10 (U)** | Evaluate query results against information need | Assess the ranked results (Step 9) against the information need (Step 7) | Relevance judgments |
| **11 (U)** | Refine information need or query | Refine information need (Step 7) or query (Step 8) based on the evaluation findings in Step 10 | Refinement of information need or query (such as the introduction of new terms to reduce ambiguity, removal of terms to increase diversity) |
| ***Steps 7 to 11 above are repeated until no further refinements are requested*** | | | |
| **12 (SYS)** | Display final ranked set of relevant documents | Obtain and display the most correct set of relevant documents | Final ranked set of relevant retrieval documents |

**Table 1. Core IR model step description, including recent work in crowdsourcing and games.**

| Step No | Step Description | Task/Objective | Expected Output |
|---|---|---|---|
| **3a (SYS)** | Perform lexical analysis | Break each document into tokens for analysis | A set of tokens from the documents in the collection |
| **3b (SYS)** | Term resolution | Resolve term acronyms and abbreviations | A set of tokens with abbreviations and acronyms resolved |
| **3c (SYS)** | Tag term parts of speech | Determine and tag the part of speech for each token | Part of speech (POS) tags for each token in the collection |
| **3d (SYS)** | Remove stop words | Remove specific tokens from the collection | The set of tokens determined in step 3a, less those tokens in our stop list |
| **3e (SYS)** | Stem tokens | Reduce each token to a stemmed form | A stemmed set of tokens from the collection |
| **3f (SYS)** | Analyze tokens | Perform analysis on document tokens as an input into Step 5 (configure system) | An analysis of the collection, such as document statistics, diversity of tokens, etc. to determine an appropriate indexing strategy |
| **3g (SYS)** | Classify documents | Assign documents to one or more classes | A set of documents contained in each class |

**Table 2. Preprocessing steps of the core IR model, including recent work in crowdsourcing and games.**

Examining Tables 1 and 2 further, we see several steps of our IR model fail this important scalability test. All seven of the preprocessing steps (Steps 3a - 3g) do not scale for either crowdsourcing or GWAP design, given the large number of items to be evaluated. Preprocessing several million documents does not scale sufficiently for using crowdsourcing or GWAP approaches. Some crowd-based mechanisms, such as Soylent (Bernstein et. al, 2010) and Turkit (Little et, al, 2010) are designed for preprocessing tasks such as these, but even these platforms would likely face significant scaling issues even on a moderately-sized IR system.

**Criterion 2:** *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 GWAP players could be made aware of in a reasonable amount of time. The domain definition (Step 1) usually requires system designers with considerable local experience. Thus it does not meet this criterion and is not suitable to complete using the crowd or GWAP. Likewise, the configuration of the retrieval system (Step 5) requires specialized knowledge of IR systems, such as the ability to tune and configure parameters and develop a search strategy. Thus, it is unlikely to benefit from an "open call" and is eliminated as well.

### Secondary Assessment

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

**Criterion 3:** *Can the process performed by the mechanism be integrated in a timely and cost-effective manner? (can be integrated)*

The following two additional criteria apply only to GWAP:

**Criterion 4:** *Can the mechanism be designed to be entertaining yet accomplish the objectives of the task? (fun yet meets objective)*

**Criterion 5:** *Can the mechanism be designed to provide an evaluation of performance and a score aligned with the task's objective? (scoring aligns with objective)*

Each of these criteria is designed to be answered in a "yes"/"no" format. This format allows a measurement of each mechanism's suitability for each step of our IR model. With Criterion 3, we wish to determine if the mechanism can integrate with the process associated with a particular step, particularly with respect to temporal demands.

For GWAP, we evaluate two additional criteria. Criterion 4 determines the potential of making the task engaging or entertaining. This criterion evaluates whether the concept of *flow* (Csikszentmihalyi 1991) can be maintained while still attaining the step's primary objective. It would be challenging to implement a user interface (Step 6), for example, as an engaging GWAP. Criterion 5 evaluates if it is possible to provide a numerical representation of player achievement and evaluate a player's performance in "real time" for proper execution in a task. Scoring is only meaningful for GWAP and not applicable to crowdsourcing. Some tasks, such as the one where the GWAP format obtains relevance feedback on queries (Step 10), can be scored in real time; other tasks, such as obtaining a document collection (Step 2), require a longer period of time before their outputs can be evaluated and scored, making that step less suitable for a GWAP format with this criteria as well.

We examine the IR tasks with these criteria. We score the mechanisms for suitability based on the percentage of "yes" answers using the following equation.

$$score = \frac{\text{criteria answered as "yes"}}{\text{criteria answered as either "yes" or "no"}}$$

We give low, medium and high suitability ratings based on the scores achieved.

Examining the six remaining steps using Criteria 3-5, we observe that for crowdsourcing, all six are rated high. For GWAP, one is rated high, three are rated medium, and two are rated low. Below we discuss each of the six steps covered in Table 3.

### Step 2 – Obtain Document Collection

**Crowdsourcing:** High; **GWAP:** Low

This step involves producing a collection of documents for an identified domain and the output from this step is a collection of documents or document sources. The crowd can assist with this step, provided that it can be done within a reasonable amount of time. The amount of time considered "reasonable" is dependent on the system requirements, mandated deadlines, and other criteria determined at the local level. If the task has a tight time deadline, a "divide and conquer" approach can be applied to quickly locate documents for the collection.

Designing this step into a GWAP-based format would

| Step | Criterion | 3 | 4 | 5 | Suitability Rating |
|---|---|---|---|---|---|
| | Mechanism | Can Be Integrated | Fun Yet Meets Objective | Scoring Aligns w/ Objective | |
| 2 – Obtain Document Collection | Crowdsourcing | Yes | N/A | N/A | High |
| | GWAP | No | No | No | Low |
| 6 – Implement User Interface | Crowdsourcing | Yes | N/A | N/A | High |
| | GWAP | Yes | No | No | Low |
| 7 – Identify information need | Crowdsourcing | Yes | N/A | N/A | High |
| | GWAP | Yes | Yes | No | Medium |
| 8 – Obtain query terms and operators | Crowdsourcing | Yes | N/A | N/A | High |
| | GWAP | Yes | Yes | No | Medium |
| 10 – Evaluate query results against an information need | Crowdsourcing | Yes | N/A | N/A | High |
| | GWAP | Yes | Yes | Yes | High |
| 11 – Refine information need or query | Crowdsourcing | Yes | N/A | N/A | High |
| | GWAP | Yes | Yes | No | Medium |

**Table 3. Assessment of the suitability of applying crowdsourcing and games to the core IR model**

likely be more problematic. First, it would be difficult to make this step engaging; second, to apply real-time scoring of document collection suitability would also be challenging. Although Criterion 3 (can be integrated) can be met by the crowd-based approach, we do not consider the GWAP-based approach a good fit for this step.

No examples using the crowd to determine sources for a document collection have been found in the literature. However, we note this task plays to one of crowdsourcing's advantages: obtaining a potentially diverse set of responses to a question.

There is a distinction between having the crowd locate items for a collection and having the crowd create new data. A number of crowdsourcing studies are involved with the latter, particularly in activities involving labeling and geo-tagging (the OpenStreetMap project (McCann, Doan et al. 2003) is one such example), but there are relatively few involved with the former. These data creation tasks are not within the scope of our core IR model.

The literature does not provide any known studies describing GWAP formats where players identify documents for a collection. There have been some GWAP created to accumulate common-sense knowledge such as Verbosity (Von Ahn, Kedia et al. 2006) and Common Consensus (Lieberman, Smith et al. 2007) games. However, these GWAP-based mechanisms are primarily used for the clarification of facts (e.g., "how tall was Abraham Lincoln?"), instead of locating documents for a collection (e.g., "which resource(s) did you use to determine Abraham Lincoln's height?").

### Step 6 – Implement the User Interface

**Crowdsourcing:** High; **GWAP:** Low

This step includes three aspects of user interface implementation: the design (the look and feel of the interface), the functional integration (the ability to allow a user to specify conditions or weights to be used in the retrieval process) and testing (the ability for the interface to correctly make use of all aspects of the retrieval system). Although it is possible to crowdsource the functional integration, this is normally not done due to trust issues – few system designers would allow anonymous crowd participants to have access to the crucial parts of a retrieval system; thus is best left to local experts who are held accountable and are familiar with the nuances of the specific retrieval system. However, the interface design and testing are both suitable for crowdsourcing.

The user interface design could be handled in several ways. First, the design could be crowdsourced if clear guidelines can be provided to several designers, who work on it in parallel, and the most suitable design would be accepted for use. Having several crowd participants working on designs at once, a design is able to be created more quickly. Likewise, there would be less risk of delays in delivery; if one of the designers does not produce a suitable interface prior to an established deadline, it is likely there will be still be several other suitable designs for consideration.

Testing the interface would also allow for the "divide and conquer" approach – many software beta tests use a form of crowdsourcing to report any issues with a software product quickly and effectively. Likewise, a user interface designer could build many interfaces and have the crowd test them systematically, which is a method used by search engine interface designers.

It would be more difficult to create a GWAP to design an interface and still make it fun; likewise, the ability to score user interface design in real time would be challenging. The functional aspects of the interface also are not suitable because of the difficulty of making it fun, as well as the same privacy issues that apply to the crowd. Testing would be easier to turn into a GWAP – the user could be given a task using the interface and the quality and time taken to obtain the results could be measured and scored.

With the exception of open-source software initiatives, such as the Mozilla software project[1], no examples have been found in the literature involving all three aspects of implementing the interface (design, functionality, and testing); however, many crowdsourcing platforms, such as elance.com, guru.com and odesk.com allow for this type of crowd-based development model and have been applied successfully to user interface design and testing. In contrast, with GWAP based mechanisms, there have not been any to date that address user interface implementation or testing.

### Step 7 – Identify the User's Information Need

**Crowdsourcing:** High; **GWAP:** Medium

Step 7 is the first IR step that typically involves the user. In this step, the goal is to identify the information need of the user as an expressed (usually written) statement. A user's information need might be to "find the highest-rated sources of pizza in Baltimore" and this step involves how to correctly communicate this need to our document retrieval system (i.e., Rated by whom? Do sources mean restaurants or supermarkets and the cooking of friends as well? Does "Baltimore" also imply neighboring areas such as Catonsville?).

Both crowdsourcing and GWAP are able to perform this task, and are given scores of "high" and "medium" respectively. Identifying the information need would integrate well into our IR system, provided that the integration could be done in a timely manner. For example, if a user required assistance with identifying a need for finding the best pizza in Baltimore, the threshold of time might be a few minutes at most; if they needed assistance with something more abstract and less time-sensitive, they might be willing to wait a bit longer for a good solution from the crowd.

[1] http://www.mozilla.org

Finding participants in the crowd at the time of need may not always be feasible. One possible solution is to have a portion of the crowd "on-call" at a set minimum fee – Yan *et. al.*'s CrowdSearch is a crowd-based image search system that discusses such an on-call system to reduce the latency time (Yan, Kumar et al. 2010). Given the appropriate temporal conditions, we believe that it sufficiently meets the integration criterion.

It would be possible to design a GWAP that is both fun but still meets the task's objective, as the PageHunt game has demonstrated. One challenge for GWAP is to score this in real-time; to provide a score would require us to somehow anticipate all player inputs and score them in advance, which is not trivial. PageHunt, for example, only scores on the time taken to supply the words, a very restricted approach to scoring that may limit the GWAP's appeal to potential players.

In the literature, several collaborative approaches have been examined in crowdsourcing, including a study on collecting user interaction information with search interfaces (Zuccon, Leelanupab et al. 2011) by defining user information tasks, capturing the interaction, and applying post-search analysis to evaluate if the original information need could be determined from the analysis. Another study found that crowdsourcing can be effectively used in localized (geographically-constrained) searches (Paiement, Shanahan et al. 2010). If a user is searching for "pizza", a determination of their true information need may be "pizza available now and within 5 miles from my current location." This evaluation of what the user enters and their true intent was examined to evaluate any disparity.

There are only a few GWAP that have been created to evaluate the information needs of users and associate them with a query or set of queries; the Page Hunt game, (Ma, Chandrasekar et al. 2009), provides a set of query results and asks the player to determine the search terms (and thus an understanding of the user's information need). Page Match and Page Race are multi-player variations of Page Hunt, and associate an information need with search results. The information obtained from these GWAP can be used as input into understanding information needs of users.

**Step 8 – Obtain query terms and operators**

**Crowdsourcing:** High; **GWAP:** Medium

This step involves mapping the user's information need into a format the search engine can correctly understand. It involves the optimal use of query terms along with associated operators (e.g., Boolean operators).

Crowdsourcing and GWAP mechanisms are given high and medium scores, respectively, based on our criteria. The condition of timeliness of the crowd with the integration criterion discussed in Step 7 applies in this step as well. It is suitable to use the crowd for this task as long as the amount of time to get a response is appropriately considered. In most cases, few users will wait more than a few minutes for query terms to be produced; if the crowd can assist with this task quickly, it can be considered feasible. In other situations that use batched data, such as "provide a query that summarizes the news on the international currency market for the previous 24 hours and provide it by the following morning", there is a relatively larger time window in which to define and provide the query terms and operators. Given the temporal conditions, we believe that it can sufficiently meet the integration criteria for both mechanisms.

Likewise, it would be straightforward to design a GWAP that is both fun and still meets this step's objective. However, to score these terms and operators in "real time" for a GWAP would likely be a challenge. One way real-time scoring could be accomplished is by running these queries against the database and scoring their performance using some approximation methods (overlap, diversity, etc.) and making these scores available for use with the GWAP format. For those queries entered by the player but which have not been scored in advance, a randomized score within a preset range of values would be provided in "real time". This randomized score would be approximately equivalent to the user's mean per-query score. Thus, if a score had not been pre-calculated for a user query submitted during the GWAP, we still maintain engagement, or flow, without providing a significant scoring advantage or disadvantage to the user.

There has not been any relevant research that directly addresses this step using crowdsourcing. One study used the crowd to enhance database queries (Franklin, Kossmann et al. 2011). The crowd is used to modify unsuccessful database queries with additional query terms and operators. In another study using GWAP formats, players were asked to provide search inputs and refinement on search terms for an information need using a GWAP, Koru, (Milne, Nichols et al. 2008). Humans could play a larger role through crowdsourcing or GWAP to enhance queries in this step.

**Step 10 – Evaluate query results against the user's information need**

**Crowdsourcing:** High; **GWAP:** High

In this step, we examine the results provided in Step 9 against our information need identified in Step 7. This step is the one where our human-based mechanisms can provide considerable value. Relevance judgments comparing query results to the user's information need score highly for both mechanisms. Again, the time to obtain the relevance assessment needs to be considered based on the user's expectations; for a shorter time window, such as if the user is urgently waiting for a response, both mechanisms would be a poor match. For other needs, such as information filtering, using the example given in Step 8 (obtaining news on international currency markets), crowd-based mechanisms would likely perform well. As with Steps 7 and 8, given the temporal conditions, we believe that the mechanisms can sufficiently meet the integration criterion.

With GWAP, the task can be made enjoyable, as observed with the results obtained with the GeAnn (Games for Engaging Annotations) game (Eickhoff, Harris et al, 2012) Likewise, if pre-established gold standard judgments are used, scoring the judgments can be made in real time. With GeAnn, the consensus of user judgments determines the gold standard.

There are a number of papers that discuss the effectiveness of using the crowd to evaluate query results against the information need. Again, in the batch experiment setting, one study examined whether the crowd can assess as well as TREC evaluators (Alonso and Mizzaro 2009): their conclusion using TREC data was that the quality of crowdsourcing raters was as good as the experts (TREC assessors). They note that it is extremely important to carefully design the experiment and collect feedback from the crowd in the relevance assessment process. Other papers examine relevance assessment as well, an examination of inter-annotator agreement in crowdsourcing relevance assessments in (Nowak and Rüger 2010) also conclude that using the crowdsourcing mechanism for assessments is not statistically different from expert assessors. A similar conclusion was made by Grady and Lease (Grady and Lease 2010). Last, Whitehill *et. al.* compared expert annotators with the crowd and it was determined that quality differences were not statistically significant, but the differences in time required and cost were indeed significant between the two groups (Whitehill, Ruvolo et al. 2009).

Additionally, numerous crowdsourcing studies examine the relevance of other types of documents to an information need, such as ranking Twitter feeds (Naveed, Gottron et al. 2011), ranking music results (Urbano, Morato et al. 2010), Rankr, a generalized crowdsourcing ranker (Luon, Aperjis et al.), the use of Learning to Rank methods (Kumar and Lease 2011) to name a few. Therefore, a number of studies have examined the ranking of the document collection based on these crowdsourced judgments. Likewise, with games, there are several GWAP-based tools that allow a user to Matchin (Hacker and Von Ahn 2009), Picture This (Bennett, Chickering et al. 2009) work with images to help associate and rank each collection based on a set of tags.

With GeAnn a set of categories is provided to describe the association of a given keyword with a text snippet (Eickhoff, Harris, et.al, 2012). The categories can be used to assess the relationship between the text snippet and the keyword. GeAnn, accumulates these assessments for a far lower cost-per-assessment than would be possible using crowdsourcing; for the TREC Crowdsourcing track, 10,350 labels were assessed at a quality level comparable to other crowdsourcing tasks for a total cost of $3.74. We believe there is significant room for new GWAP to be introduced.

**Step 11 – Refine information need or query**

**Crowdsourcing:** High; **GWAP:** Medium

In this step, we refine either the information need that was produced in Step 7 or the query provided in Step 8, based on the evaluation of the findings, i.e., retrieved/ranked items, from Step 10. This involves the refinement of information that is based on feedback from the query results returned (i.e., relevance judgments on a few documents); if the user's information need is satisfied, we move to Step 12, if not, the user refines their information need and repeat Steps 8-10. As with Steps 7 through 10, timeliness is an essential consideration for integration; given the temporal conditions, we believe that both of these mechanisms can sufficiently meet the integration criterion.

Consider the information need example mentioned in Step 7. If, after the query was issued, the user was not able to obtain important information from the results (e.g., due to allergy concerns, they need to find a gluten-free pizza in Iowa City and this information was not available in any of the search results), the nature of the initial query has changed to now include a conjunction of search terms related to "gluten-free". Depending on the user's expectation, the time needed to locate this pizza, as well as the time to find crowd participants who are able to help, the crowd may or may not be easy to integrate into our model. Likewise, with our refinement on international currency news summaries mentioned in the discussion in Step 8; perhaps we notice some important correlation of tomorrow's currency markets with today's oil prices. We may ask for a query refinement to be done by the end of the current day, providing a window of several hours. Thus, in this second example time is less of an issue for our refinement and would be a far easier condition to meet.

Integration of these refinements using GWAP would be more challenging than using the crowd, but it could be done; to accomplish this integration task, we can limit these refinements to several suggestions and have the user choose one or more. With regard to a fun/engaging format, we believe query refinement can meet this criterion. However, we find that scoring in "real time" for this step might be a challenge, so we mark the scoring criterion as a "no" for GWAP. Indeed, this area is of particular interest to us, and we plan to conduct a series of experiments in this area.

Prior research discussing the use of human computation mechanisms with query reformulation is very limited. Since evaluation of query terms and operators is implied in this step, many of the studies described in Step 8 would also apply to this step, such as the aforementioned Koru game.

### The Value-Added Effects of Human Mechanisms

Our initial criterion (Criterion 1), which examined if each step in our IR model could scale using the crowd and GWAP, was evaluated based on turning the entire task over to these mechanisms. Steps 3a – 3g, 4, and 9 were eliminated due to scaling issues. What if we took a hybrid approach to those tasks, where computers and humans together apply their strengths? We note that even the best tokenization, stemming, and part of speech tagging tasks generally have accuracy rates near 90 percent, e.g., (Ravi and Knight 2009). For other machine-based pre-processing tasks such as term resolution, this accuracy dips down to approximately fifty percent, e.g., (Ng 2007). We see value in having the machine perform a significant share of the work in some steps, then have the crowd or GWAP-based mechanisms assist with a subset of the task – the portions with which the machine has the most difficulty.

With this new criterion, we consider a human value-added criterion on the document preprocessing steps (Steps 3a – 3g), the indexing step (Step 4) and the retrieval and ranking step (Step 9), which we eliminated earlier for their inability to scale for human mechanisms. This new criterion

| Step No | Step Description | Assessment of Applying Crowdsourcing | Assessment of Applying Games with a Purpose |
|---|---|---|---|
| **1** (SD) | Define domain | ○○○ Requires local/special knowledge to perform. | ○○○ Requires local/special knowledge to perform. |
| **2** (SD) | Obtain document collection | ●●● High – finding new data sources to apply to existing domains is something the crowd could readily assist with. | ●○○ Low – Making this a fun GWAP, is a challenge; making it possible to score in real time would be difficult |
| 3 (SD) | Preprocess documents | See **<u>Table 5</u>** for specific details on each of the preprocessing steps | |
| 4 (SYS) | Index documents | ○○○ Does not scale, little ability to add human value | ○○○ Does not scale, little ability to add human value |
| 5 (SD) | Configure retrieval system | ○○○ Requires local/special knowledge to perform. | ○○○ Requires local/special knowledge to perform. |
| 6 (SD) | Implement user interface | ●●● High – having the crowd help design user interfaces is promising for design and testing. | ●○○ Low – Although a GWAP could be integrated for testing, the challenge of creating a user interface might lack the excitement needed and scoring would be difficult. |
| 7 (U) | Identify information need | ●●● High – crowdsourcing is particularly useful to help define complex information needs or to assist novice users | ●●○ Medium – the challenge to the GWAP format is to allow the GWAP to be scored in real time. |
| 8 (U) | Obtain query terms and operators for the user's search | ●●● High crowdsourcing is particularly useful to obtain query terms and operators | ●●○ Medium – the challenge to the GWAP format is to allow the GWAP to be scored in real time. |
| 9 (SYS) | Retrieve and rank user query results | ○○○ Does not scale, little ability to add human value | ○○○ Does not scale, little ability to add human value |
| 10 (U) | Evaluate query results against an information need | ●●● High – getting assistance from the crowd to compare results and information need is a crowdsourcable task | ●●● High – performing relevance assessments can be done in real time and made interesting. |
| 11 (U) | Refine information need / query | ●●● High – having the crowd evaluate or aid in refinement of the information need (step 8) based on a set of retrieval results (step 11) is appropriate as a crowd task | ●●○ Medium – identifying the information need can be made into an interesting GWAP. The challenge is to make it possible to score it in real time. |
| *Steps 7 to 11 above are repeated until no further refinements are requested* | | | |
| 12 (SYS) | Display final ranked set of relevant documents | This is a termination step and just displays the best result; therefore no assessment done | This is a termination step and just displays the best result; therefore no assessment done |

**Table 4. Assessment of Applying of Crowdsourcing and GWAP to Accomplish Steps in our Core IR Model**

| Step No | Step Description | Assessment of Crowdsourcing | Assessment of Applying Games with a Purpose |
|---|---|---|---|
| 3a (SYS) | Perform lexical analysis | ●○○ Low – The error rate on this step is low, except in specialized domains (such as chemical terms). There is limited human value added. | ●○○ Low – The error rate is generally low, though it would be easy to make this into a GWAP. There is limited human value added. |
| 3b (SYS) | Term resolution | ●●● High – The human value-added component is high, given a substantial machine error rate. | ●●● High – This could be made into a GWAP that could be fun and evaluated against a lexicon in real time. |
| 3c (SYS) | Tag term parts of speech | ●●○ Medium – This could be evaluated by the crowd. The low machine error rate keeps this from being rated high. | ●●○ Medium – This is also a task that could be evaluated as a GWAP. |
| 3d (SYS) | Remove stop words | ●○○ Low – The low error rate limits the human value added. Creating a stop list may be slightly more valuable | ●○○ Low – It would be a challenge to turn this into a GWAP, since it involves examining suitable terms in a very large document collection and thus is not practical |
| 3e (SYS) | Stem tokens | ●○○ Low – Stemming follows rigorous rules, and the human value added is low | ●○○ Low – It would be difficult to turn a stemming task into a fun GWAP |
| 3f (SYS) | Analyze tokens | ●○○ Low – This token analysis usually involves examining aggregate information, which humans provide limited value, particularly since it requires specialized knowledge | ●○○ Low – Since it requires specialized knowledge, it would be difficult to make into a GWAP, and difficult to score in "real time" |
| 3g (SYS) | Classify documents | ●●● High – Humans can provide a training set, or perform quality assurance on machine-classified documents. | ●●● High – Classification is a task that is easy to turn into a GWAP, and thus the human value added is high |

**Table 5. Assessment of Applying of Crowdsourcing and GWAP to Accomplish Preprocessing Steps**

evaluates whether the crowd or GWAP mechanism can be designed to have humans add value to the task.

**Steps 3a to 3g – Document Preprocessing**

**Human value-added:** Varies (see Table 5)

The preprocessing of documents involves the seven steps illustrated in Table 2. If done appropriately, there is considerable value that can be added. To use the term resolution step (Step 3b) as an example, we could allow the machine to do the bulk of the work, matching each token against a lexicon. Provided that our lexicon contained a sufficiently large number of terms and were matched appropriately to the terms used in our document collection, we would have a small set of terms that could not be matched to lexicon entries. This set of terms, or a portion thereof, could have crowd or GWAP-based mechanisms applied if the number of terms to be evaluated was reasonable. The goal in each mechanism would be to provide assistance with those the machine could not resolve, or to provide quality control (including error estimates) on the steps performed by the machine. Likewise, our mechanisms could handle the other preprocessing steps too by serving as a quality assessment tool to evaluate the outputs from this automatic classifier. This will provide additional value to our IR system in terms of better processing output accuracy.

As in Step 3b, several studies have used the crowd for entity recognition and resolution tasks, such as the aforementioned CrowdDB (Franklin, Kossmann et al. 2011), as well as work by Su *et. al.* (Su, Pavlov et al. 2007) and Robson *et. al.* (Robson, Kandel et al. 2011). Several tasks have used the crowd for abbreviation resolution, such as Finin *et. al.* do in (Finin, Murnane et al. 2010).

Document classification is the task of coming up with a grouping of the documents based on their content. Once documents are assigned to groups, users can include or exclude documents belonging to one or more of these groups. Given a set of topics, user information needs, or other user context specifications, classification can be applied to decide to which particular class (or classes) a document (or set of documents) might belong. A standard approach is to manually classify a subset of documents and then automatically classify new documents based on "learning" from the manual classification. Using this approach, we could utilize the crowd for classifying the subset of documents for which the machine has difficulty to determine a class. These crowd-classified outputs could be used to refine machine-based classifiers.

With the exception of classification of documents and term resolution, very little previous research has been applied to either crowdsourcing or GWAP in the preprocessing steps. There are numerous labeling tasks that involve applying labels to objects in both crowdsourcing and in GWAP-based formats – when a specific set of labels (classes) are given to the user, it is *classification* (as opposed to free-form labeling which is considered *annotation*). A few GWAP, such as ESP game (Von Ahn and Dabbish 2004) come close to classification; however most GWAP involve the free-form labeling approach (users can tag an item with whatever label seems most suitable, and are not required to use a pre-selected set of labels).

With the addition of human processing mechanisms, we are able to address the most challenging instances of these preprocessing tasks without being overwhelmed by the majority of tasks that computers can address without human intervention. No research has been found that addresses these quality enhancement approaches through either crowdsourcing or GWAP-based mechanisms.

## Evaluation of Crowd and GWAP Mechanisms to our IR Model

In this section, we combine the evaluation of our criteria made in the preceding sections into a summary chart. Tables 4 and 5 summarize the assessment of applying crowdsourcing and GWAP mechanisms to each step of our core IR model. In Table 5, we apply these mechanisms to our seven preprocessing steps and demonstrate the ability to provide value. These assessments assume that crowd and GWAP mechanisms take a "reasonably-sized" subset of the output; in most steps, the tokens marked for human involvement would automatically be determined by the machine-based methods. For example, in Step 3b, the machine output would be a list of the tokens that did not match an entry in the lexicon; the list would be provided to our mechanisms to allow human value to be added.

## Future Research Directions

Crowdsourcing has been applied to some areas of IR, however this application is focused in a few areas, primarily those involving relevance judgments, image labeling, and domain and data source discovery. Much of the recent work in crowdsourcing is focused on spam and bot reduction, methods of encouraging better collaboration techniques and application to new IR domains, such as multimedia. We find that many areas of our core IR model remain untapped, including query reformulation and the preprocessing phases. The areas of greatest interest to IR are those steps where suitability of crowdsourcing is highest but where little research has been performed to date. Steps 7 and 10 have a high applicability to crowdsourcing, and have had substantial research already applied. Steps 8, 9, and 11 examine the refinements to either the information need and/or to the query, but have each had scant attention despite high crowdsourcing suitability. Step 3, contains many tasks where human value can be applied to the most difficult situations that computers cannot determine accurately, such as in term resolution, document classification and, part-of-speech tagging. Overall, we see several opportunities in these steps for crowd and GWAP-based mechanisms.

## CONCLUSION

In this paper, we have introduced crowdsourcing and GWAP as mechanisms to accomplish IR-related tasks. For each, we have defined and described a core IR model, to which we examined the suitability of applying crowdsourcing and GWAP. We then described existing crowdsourcing and GWAP-based research already applied to each step of this core IR model and found several areas where IR could be enhanced with these human computation methods. Although every step of information retrieval may not suitable for GWAP-based or crowdsourcing mechanisms, we do believe there are significant gains that can be obtained. If we can demonstrate such significant gains using these mechanisms, we believe this contribution will provide a lasting benefit to Information Science.

## REFERENCES

Alonso, O. and Mizzaro, S. (2009) *Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment*. In Proc SIGIR'09, ACM, New York, NY.

Bennett, P. N., Chickering, D. M. et al.. (2009) *Learning consensus opinion: mining data from a labeling game*. In Proc. WWW '09. ACM, New York, NY pp 121-130.

Bernstein, M. S., Little, G. et. al. (2010) *Soylent: a word processor with a crowd inside*. In Proc UIST'10. ACM, New York, NY, pp 313-322.

Csikszentmihalyi, M. (1991). *Flow: The psychology of optimal experience*, Harper Perennial Press

Eickhoff, C., Harris, C. G., de Vries, A.P., and P. Srinivasan. (2012) *Quality through flow and immersion: gamifying crowdsourced relevance assessments*. In Proc. SIGIR 2012, ACM, New York, NY.

Finin, T., W. Murnane, et al. (2010). *Annotating named entities in Twitter data with crowdsourcing.* In Proc. NAACL HLT 2010, Association for Computational Linguistics**,** Stroudsburg, PA, USA. pp 80-88.

Franklin, M., D. Kossmann, et al. (2011). *CrowdDB: Answering queries with crowdsourcing.* In Proc. SIGMOD (2011). ACM, New York, USA. pp 61-72.

Grady, C. and M. Lease (2010). *Crowdsourcing document relevance assessment with Mechanical Turk*, ACL, Stroudsburg, PA, USA

Hacker, S. and L. von Ahn (2009). *Matchin: eliciting user preferences with an online game*, In Proc CHI'09, ACM. New York, USA, pp 1207-1216.

Kumar, A. and M. Lease (2011). *Learning to rank from a noisy crowd*. In Proc SIGIR'11, ACM, New York, pp 1221-1222.

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

Lieberman, H., D. Smith, et al. (2007). *Common Consensus: a web-based game for collecting commonsense goals*. In Proc. IUI'07, ACM, New York, NY.

Little, G., Chilton, L. B., et. al. (2010). TurKit: human computation algorithms on mechanical turk. In *Proc.* UIST '10. ACM, New York, NY, pp. 57-66.

Luon, Y., C. Aperjis, et al. *Rankr: A Mobile System for Crowdsourcing Opinions.*

Ma, H., R. Chandrasekar, et al. (2009). *Improving search engines using human computation games*, In Proc. HCOMP'09, ACM, New York, NY.

McCann, R., A. Doan, et al. (2003). *Building data integration systems: A mass collaboration approach*, AAAI, New York, NY.

McGonigal, J. (2011). *Reality is broken: Why games make us better and how they can change the world*, Penguin Pr.

Milne, D., D. M. Nichols, et al. (2008). *A competitive environment for exploratory query expansion*, In JCDL'08, ACM, New York, NY.

Naveed, N., T. Gottron, et al. (2011). *Searching Microblogs: Coping with Sparsity and Document Quality*. In Proc. CIKM'11. ACM, New York, NY pp 183-188.

Ng, V. (2007). *Semantic class induction and coreference resolution*. In Proc. ACL'07, ACL, Stroudsburg, PA, USA pp. 536-543.

Nowak, S. and S. Rüger (2010). *Reliable Annotations via Crowdsourcing*. In Proc. MIR'10. ACM, New York, NY. pp. 557-566.

Paiement, J. F., J. G. Shanahan, et al. (2010). *Crowd Sourcing Local Search Relevance.* In Proc. CrowdConf 2010. ACM, New York, NY.

Ravi, S. and K. Knight (2009). *Minimized models for unsupervised part-of-speech tagging*, In Proc. ACL'09, ACL, Stroudsburg, PA, USA. pp 504-512.

Robson, C., S. Kandel, et al. (2011). *Data collection by the people, for the people*. In Proc. CHI'11. Vancouver, BC, Canada, ACM**,** New York, pp 25-28.

Su, Q., D. Pavlov, et al. (2007). *Internet-scale collection of human-reviewed data.* In Proc. WWW'07. Banff, Alberta, Canada, ACM**,** New York. NY, pp 231-240.

Urbano, J., J. Morato, et al. (2010). *Crowdsourcing preference judgments for evaluation of music similarity tasks*. In Proc. CSE'10, ACM, New York, NY. pp 9-16.

von Ahn, L. and L. Dabbish (2004). *Labeling images with a computer game*, In Proc. CHI'04, ACM, New York, NY.

von Ahn, L., M. Kedia, et al. (2006). *Verbosity: a game for collecting common-sense facts*, In Proc. CHI'06, ACM, New York, NY, pp 75-78.

Whitehill, J., P. Ruvolo, et al. (2009). "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise." *Advances in Neural Information Processing Systems* **22**: pp 2035-2043.

Yan, T., V. Kumar, et al. (2010). *CrowdSearch: exploiting crowds for accurate real-time image search on mobile phones*, In Proc. MobiSys'10, ACM, New York, NY.

Zuccon, G., T. Leelanupab, et al. (2011). Crowdsourcing Interactions. In *CSDM'11.* ACM, New York, NY p 35.