VisualizIR – A Game for Identifying and Categorizing Relevant Text in Documents

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

ABSTRACT
In this paper, we introduce VisualizIR, a game where players identify relevant document terms that match predefined categories. VisualizIR evaluates players on accuracy, recall, and precision against an established gold standard, a pooled consensus of judgments made by other players, or a weighted combination of the two. The annotated document can then viewed by any XML-compatible browser, allowing for quick identification of terms in the document related to each category. Here we describe some of the playability design tradeoffs made during the game's development, as well as our findings from two experiments conducted using VisualizIR output.

Author Keywords
VisualizIR, games with a purpose, GWAP, text categorization, user interface design, information filtering

ACM Classification Keywords
H.3.3 [Information Storage and Retrieval]: Information Filtering; H.5.m [Information Interfaces and Presentation (e.g., HCI)]: Miscellaneous; K.8.0 [Personal Computing]: Games

General Terms
Human Factors; Design; Experimentation

INTRODUCTION
Text categorization (also known as *text classification* or *topic spotting*) is the task of automatically sorting documents into a predefined set of categories. There are numerous applications of this categorization task, including automated indexing of scientific articles, identifying and classifying patents, sentiment analysis, spam filtering, identification of document genre, and authorship attribution, to name a few [3]. This type of categorization typically evaluates each document in its entirety. However, there are many practical uses for categorization of text *within* a document, such as at the term (word or phrase) level. With a term-level approach, the reader can quickly focus on a document's most relevant portions, enhancing both readability and navigation. This is especially true in domains such as biomedicine, legal e-discovery, and patent prior art evaluation. A document may be categorized by its constituent terms, but the converse is not true. Term-level categorization has all the benefits of document-level categorization, plus the ability to identify the "hot spots" in a document. For these reasons, categorization at the term-level is valuable in information retrieval.

Over the past decade there has been a rapid advancement in document-level automatic text categorization techniques; however, far less has been done in term-level categorization. Significant challenges to the term-level approach include resolution of jargon, non-standard abbreviations, acronyms, and ambiguous terms. Document-level approaches can examine the document at a high level and overlook the most challenging terms whereas term-level approaches cannot. In addition, creating detailed metadata describing the terms in each document is not only tedious but also prone to error.

Games with a Purpose
Games with a purpose (GWAP) are human computation mechanisms designed to solve tasks (including the identification of relevant text in documents). These tasks are easy for humans to accomplish, but difficult for computers. These games represent human players as nodes in a large computation, permitting tasks to be quickly accomplished in parallel. However, unlike computer processors, humans need to be provided with incentives to join a collective computation [6]. Some of the benefits of games are making mundane tasks more engaging, a lower cost per task than crowdsourcing, and less spam and noisy data as compared to crowdsourcing inputs [2]. Games have emerged as a powerfully seductive tool for encouraging people to participate in this human computation process.

VisualizIR
VisualizIR is a game designed to identify relevant terms matching predetermined categories in a document collection. Using an intuitive graphical interface, the game allows players to categorize documents, which can be then used to improve searches. Although initially designed to assist with patent searches, the game's flexibility allows the evaluation of other document features, such as *sentiment*

NordiCHI '12, October 14-17, 2012 Copenhagen, Denmark

analysis (also known as *opinion mining*, which focuses on the evaluation of opinion, sentiment, and subjectivity in text) and *anaphora resolution* (the linguistic challenge of resolving what a pronoun or a noun phrase refers to).

VisualizIR provides metrics to evaluate the ability to identify relevant text based on a *gold standard*, a *consensus vote* (a pooled consensus of judgments made by other players) or a weighted combination of the two. Using consensus vote data allows us to overcome the challenges associated with obtaining gold standard data, such as its high cost and a reliance on subject matter experts. With a sufficient number of player contributions, a consensus vote can provide results that approach gold standard data [5]. If the two differ significantly, it invites further investigation and may indicate the need for further resolution of ambiguous terms, abbreviations, or jargon, etc.

To illustrate, patent prior art searches often require years of experience to perform efficiently. Providing a game that asks trainees to categorize terms could not only provide immediate feedback and search suggestions but also indicate areas of a patent document associated with a given category of interest. Based on the particular need, scoring weights can be calibrated to focus on the performance measures used in evaluating machine learning techniques: accuracy, precision, recall, or a combination of the three.

In addition to improving document search for humans, VisualizIR can also be used to train automatic classifiers. In some applications, such as sentiment analysis, automatic classifiers can be difficult to train; for example, if a product review contains an multi-faceted comparison between two very similar products, sentiment classifiers are likely to misinterpret the negative terms used in the comparison [4]. VisualizIR encourages players to provide context for the sentiment for a single entity; for example, if a good actor is cast in a bad movie, reviews of other movie features (the script, the director, or other actors) might confuse the classifier. By evaluating the review from the context of the actor, we can extract the relevant terms.

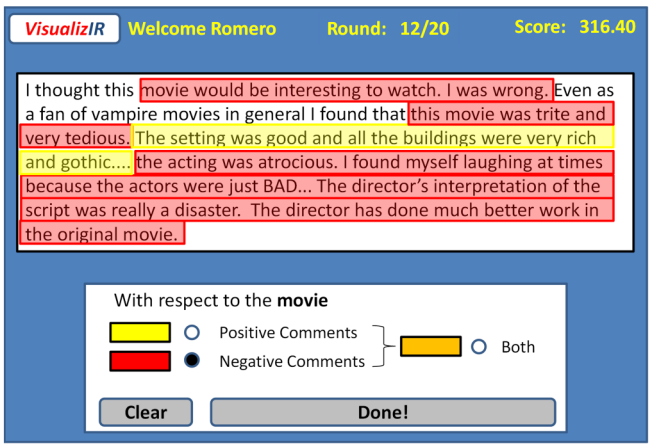

Figure 1. The VisualizIR interface. A player is tasked with highlighting terms relevant to a single entity, *movie*.

THE VISUALIZIR INTERFACE

In the design of any interface there are tradeoffs. In this section we discuss some of the design considerations we encountered during the creation of VisualizIR.

Game Structure

Each game is divided in to multiple rounds. In an earlier empirical evaluation, we created a game with 5 rounds and increased the number of rounds by five each time. Using player feedback, we determined that 20 rounds was optimal for maintaining *flow* – a concept describing the game's balance between boredom (i.e., too slow or not challenging enough) and frustrating (i.e., too fast, confusing or challenging) [1]. Throughout the game's 20 rounds, the same two categories are evaluated, but each round presents a different document from our collection. These documents are shown in random order; therefore, prior knowledge about the information contained within each document is not expected and player bias is minimized.

At the end of each round, the player's score is calculated as shown in Figure 2. At the end of the game, the scores for each round are tallied and a final score and overall player rank is then displayed. If the player's score is among the top 10, they are invited to add their name to the leaderboard.

Evaluation Metrics

As an information retrieval (IR) tool, we evaluate using three standard IR metrics: precision (exactness), recall (completeness), and accuracy, defined as follows:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where TP = number of true positive terms identified, TN = true negative terms identified, FP = false positive terms identified, and FN = false negative terms identified. Our

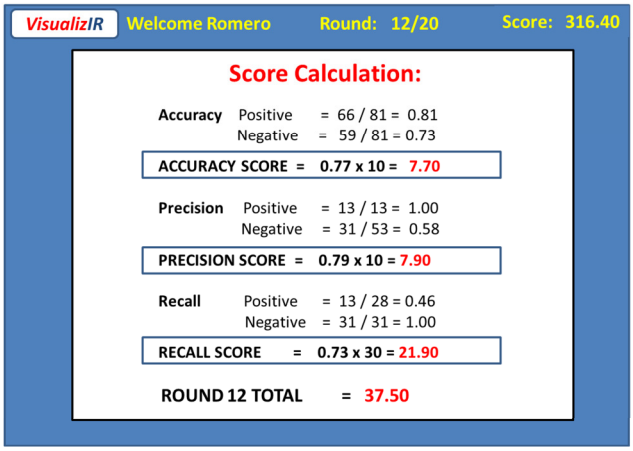

Figure 2. Score calculation for each metric after a round. The weights of 10 for accuracy, 10 for precision, and 30 for recall are established in the game settings.

metrics evaluate all terms equally without regard to their *informativeness* (typically infrequently-occurring terms contain more information than frequently-occurring terms).

DESIGN CONSIDERATIONS

Game Interface

VisualizIR was created in Adobe Flash allowing us to develop and refine the game quickly. To enhance player competitiveness, we included a leaderboard that displays the names of the highest scorers for "bragging rights'.

Figure 1 illustrates the user interface. In this example, players select one of the radio buttons at the bottom of the screen and then highlight text in a document using a mouse. The color of the highlighting indicates if a player believes a term is relevant to a first category (yellow for 'positive comments'), a second category (red for 'negative comments'), or orange for terms relevant to both categories. Non-highlighted terms are those the player indicates are not relevant to either category.

Number of Categories

We initially considered having players evaluate three separate categories in each round. Categories are not mutually exclusive; therefore 2^N-1 category decisions need to be made by each player in each round. We experimented with a number of interfaces to best represent the intersection between the three categories, including a Venn diagram (Figure 3). From a playability perspective, players indicated the interfaces was confusing and distracted them from the task. In response, we restricted each round to the evaluation of two categories. Players found the examination of two categories far more engaging and scores across all three metrics increased in response to this change.

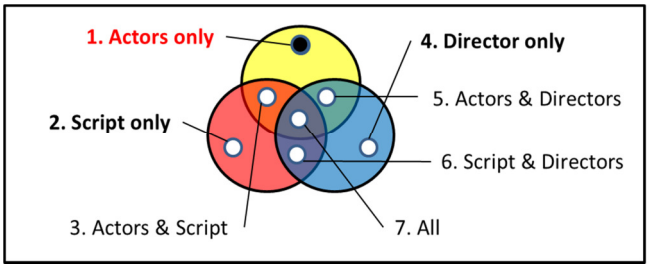

Figure 3. An early VisualizIR interface with three categories. In initial testing, players found this interface confusing.

Use of Stopwords

In our initial design, we used a stopword list to remove unimportant terms (e.g., 'a', 'and', and 'the') from the text prior to calculating metrics. Evaluating the player-highlighted text after stopword removal is arguably more meaningful, since only the most informative terms are used in the calculation. In some early experiments on playability, players indicated scoring after stopword removal was confusing and unintuitive. For this reason, the current configuration of VisualizIR does not remove stopwords.

Settings

As mentioned previously, we designed the game to be flexible and allow for a variety of configuration settings, including a weighted balance between (1) accuracy, precision, and recall and (2) gold standard or consensus vote judgments. The consensus vote determination allows a ground truth to be established by consensus opinion of other players, which has been shown to be accurate even when done by non-expert raters [5]. To avoid cold start issues (i.e., using player consensus as our ground truth without a sufficient number of judgments from which to draw inferences), VisualizIR requires a minimum of 5 votes before a document can be scored using consensus voting.

Timed Rounds

Initially each round of VisualizIR was timed. This provided an additional scoring incentive (given in the form of a bonus) for highlighting categories as quickly as possible. We found that timed players completed rounds more quickly; however, the resulting metrics – particularly scores for accuracy and precision – were significantly lower than untimed versions of the game.

Storage

All player markup information is stored in a database allowing for greater portability. As a result of how information is stored, VisualizIR is able to display a document with highlighted markups in any XML-compatible browser for quick and easy navigation. This is accomplished by embedding XML tags to identify the beginning and end of the markup for each category. In order to quickly calculate metrics at the completion of each round, we also include the start and end position of each player's annotation in a separate table. A batch script is run periodically in the background to aggregate and update the consensus votes for each document.

EVALUATION

In this section, we discuss two experiments using VisualizIR, one to examine its ability to train a sentiment classifier, another to evaluate the utility of VisualizIR XML output in assessing information in documents.

Sentiment Analysis Study

We conducted an experiment to examine sentiment in 640 Blu-ray player reviews, each containing at least 100 words, obtained in April 2012 from Amazon.com. A subset of 320 reviews was evaluated using the VisualizIR game a minimum of 5 times each[1]. We used a binary SVM^{light} classifier trained using 10 manually-created seeds on the same 320 Blu-ray reviews as our baseline. This classifier was then run on a test set of 320 unseen Blu-ray player reviews. A human volunteer evaluated the polarity of each

[1] A nice side effect of using the game-based mechanism is that 22% of players completed additional games with no expectation of compensation. This increased the number of annotations per review from 5.0 to 6.1.

of the 320 reviews in the test set to establish a ground truth. We then retrained our sentiment classifier using the 320 reviews annotated with VisualizIR and re-ran the SVM^{light} classifier on the test set. Using this approach, we are able to evaluate sentiment analysis on the document's constituent terms as a determination of the overall sentiment of the review. We then compared the results obtained with our VisualizIR-retrained classifier with our baseline. Our results showed the VisualizIR-trained classifier was able to correctly determine polarity significantly better than the baseline (two-tailed t-test, $p < 0.001$). For the 320 reviews evaluated, the baseline classifier correctly identified sentiment polarity in 251 reviews (78.4%); in contrast, the VisualizIR-trained documents correctly identified sentiment polarity in 303 reviews (94.7%).

Ease of Navigation Study

We also conducted an ease-of-navigation study (N=32, 7 female, average age = 29, min = 20, max = 47, mean computer experience = 7 years). Each subject was provided with ten randomly-chosen Blu-ray product reviews from the 320 annotated for the sentiment analysis study. Subjects were also provided with two adjacent interfaces displaying identical reviews: One interface that highlights terms matching those entered into a search tool, and another interface that displays the highlighted sentiment results from VisualizIR. This second interface used consensus-vote determined polarity from our sentiment analysis study displayed in a format similar to the one shown in Figure 1.

Subjects were asked to identify polarity of specific features in each of the ten reviews. We logged each subject's activity time with each of the two interfaces in order to evaluate which one was utilized more during the task. After each review, we also asked subjects to indicate a simple binary preference between the two interfaces.

In the 320 documents evaluated, subjects indicated a preference using VisualizIR in 246 (76.9%), which is significantly different from random $\chi^2(1, N=320) = 92.45$, $p < 0.001$. An evaluation of the logged usage for each tool showed no statistical difference (average 12.8 seconds per search using the baseline search tool, vs. 11.5 seconds per search using the VisualizIR output). However, this doesn't provide a complete picture – as subjects progress through the ten tasks, the logs show they abandon the search tool and rely more on the VisualizIR-created output and accomplish this task more quickly as a result. Figure 4 illustrates this switch occurs for most subjects prior to the fourth task. This observation likely indicates a preference for the increased efficiency of VisualizIR to navigate a document and determine sentiment.

CONCLUSION

We have described VisualizIR, a term-based assessment game, which evaluates a player's ability to categorize text in documents. The player's assessment of terms related to a category can be scored using accuracy, precision and/or recall using a gold standard, a pooled consensus of judgments made by other players, or a weighted combination of the two. We also discussed some playability and design considerations encountered during game development. We conducted two experiments using VisualizIR: first, we evaluated its ability to aid a sentiment analysis classifier to determine product reviews polarity; second, we compared the navigational ease of VisualizIR output with a standard document search tool. These experiments show VisualizIR is a versatile game that not only improves document readability but also gathers useful metadata, which is often tedious to obtain using human annotators and often unobtainable using automatic methods.

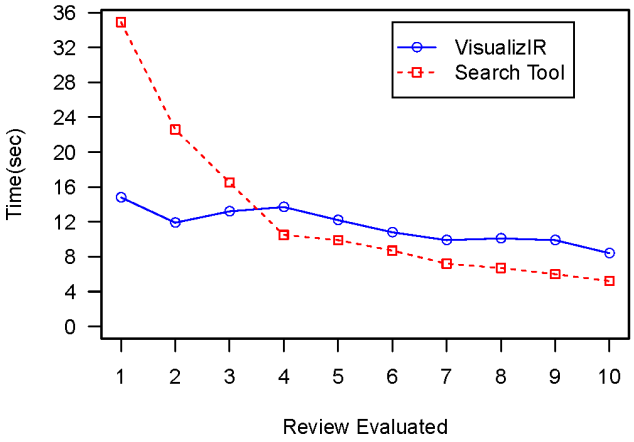

Figure 4. Time spent using each interface for a series of tasks

We understand there are numerous aspects to VisualizIR that remain untapped. In future work, we plan to further enhance the playability aspects of the game, incorporate multimedia, and explore the use of VisualizIR as an active learning tool for search and indexing strategies.

REFERENCES

1. Csikszentmihalyi, M. Flow: The psychology of optimal experience. Harper Perennial, New York, 1991.
2. Eickhoff, C., Harris, C. G., de Vries, A.P., and Srinivasan, P. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In Proc of the 35th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR'12). ACM, New York. 2012.
3. Medlock, B. Investigating classification for natural language processing tasks. VDM Verlag, 2008.
4. Prabowo, R. and Thelwall, M. Sentiment analysis: A combined approach. Journal of Informetrics, 3:2, 2009. pp 143-157
5. Snow, R., O'Connor, B., Jurafsky, D. and Ng, A. Y. Cheap and fast-but is it good?: evaluating non-expert annotations for natural language tasks. In Proc of the Conference on Empirical Methods in Natural Language Processing. ACL, Stroudsburg, PA, USA, 2008.
6. von Ahn, L. and Dabbish, L. Designing games with a purpose. Comm of the ACM, 51:8, 2008. pp 58-67.