Human Aided Computer Assessment for Exhaustive Search

Christopher Hogan
H5
San Francisco, USA
chogan@h5.com

Dan Brassil
H5
San Francisco, USA
dbrassil@h5.com

Mitch Marcus
University of Pennsylvania
Philadelphia, USA
mitch@cis.upenn.edu

Abstract—Human-computer interaction has been proposed as a means for more effectively dealing with the challenges posed by information retrieval. However, within the hierarchy of information needs, certain kinds of needs are only poorly handled by existing techniques. Information needs requiring thoroughness (exhaustive research, information discovery) are not well served by existing models of human-computer information retrieval (HCIR). Information needs such as these are commonplace in legal, patent, medical and intelligence searches. In such applications, high recall with high precision is the primary consideration, and the standard model, which treats human interaction as a kind of post-processing filter, does not yield a system with desired characteristics. In this paper, we propose an alternative model of HCIR which yields systems whose properties are more closely aligned with the needs of the exhaustive research task, and describe an implementation of such a system, demonstrating its effectiveness over the standard model in a task that models real-world conditions.

Index Terms—Human-Machine Cooperation and Systems; Human-Computer Interaction; Computational Intelligence; Information Retrieval

I. INTRODUCTION

Information Retrieval (IR) tasks may be classified according to the user's information need, the reason that the user is requesting the information. One scheme for information need [1] defines the follow categories:

- **Known Item**: The user is searching for an information object already known to exist.
- **Exploratory**: The user is seeking to learn something about a topic but does not know in advance what may be important.
- **Exhaustive Search**: The user is trying to learn *everything* about a particular topic.

Much IR research has been focused on the Known Item task, and powerful methods have been developed to ensure that important, known-items are ranked near the top of retrieval lists [2]. For known-item searches, result sets are small (usually one item), queries are short (2.3 words on average [3]), and there is strong emphasis on precision, as opposed to recall.[1]

Exploratory information needs, while less well researched than known-item, are amenable to a variety of techniques that build on basic search, including query expansion, and aggregate processing of result sets [4]. Because the user is seeking to explore the information space, result sets must be small enought to be easily comprehended and precision continues to be relatively important.

Exhaustive Search information needs are characterized by a requirement for high recall and high precision. Information needs such as this are often encountered in the fields of medicine, law and intelligence [4], [5]. In these fields, there is a requirement that all of the information related to a given topic be identified. The consequences of missing information in these fields can be enormous: a missed diagnosis, legal sanction or geopolitical disaster. The consequences of retrieving too much information can be similar, if less injurious: time wasted reviewing results or added review costs. Result sets for exhaustive search, comprising all documents relative to a particular topic, can be huge. This fact, together with other factors, including subtle topic definition and substantial linguistic variability, combine to make exhaustive search substantially more challenging than other information needs.

One approach that has been explored to address the challenges of IR is to incorporate greater human interaction into the retrieval process. Human-Computer Information Retrieval (HCIR) [6] examines approaches that combine insights from information retrieval and human computer interaction. HCIR combines human cognitive capabilities and machine capabilities to produce systems that offer increased human control over the retrieval with concommittant human expenditure of effort.

Although HCIR has been applied to known-item retrieval, in the form of authority models [7], it is clear that the exploratory and exhaustive search needs are most likely to benefit from increased user interaction to better define the information request and refine the resulting retrieval.

In particular, note that there exists a relationship between the value the user assigns to a particular retrieval result and the amount of effort he or she is willing to invest in guaranteeing a high quality result. In the lower limit, a user may not assign very much value to a web search, and is therefore unwilling to invest much effort in specifying or refining their request. This translates into short queries and an unwillingness to use iterative approaches such as relevance feedback for web queries. In the upper limit, a user whose search has serious consequences, as described above, will be willing to invest a substantial amount of effort into ensuring that the correct

---

[1]Precision is the proportion of documents returned by a system that are correct, while Recall is the proportion of correct documents that are identified by a system.
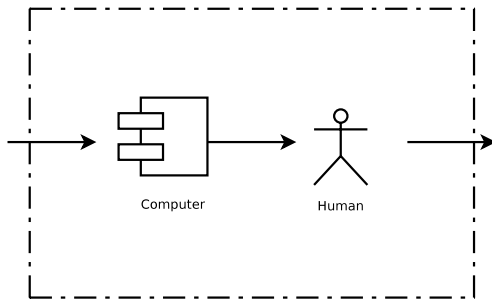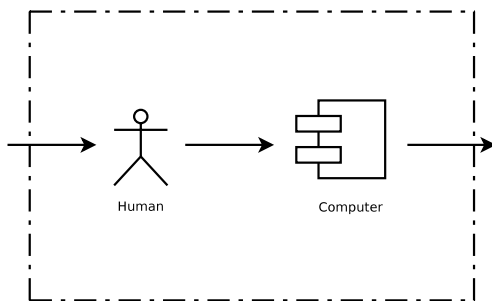
Fig. 1.   Computer Aided Human Assessment.

Fig. 2.   Human Aided Computer Assessment.

result is achieved. Thus, not only is exhaustive search most likely to benefit from human-computer interaction, but it and its applications are most likely to sustain the substantial human effort needed to make effective use of interaction.

In this paper we explore HCIR models as applied to the problems posed by exhaustive search. We propose a classification scheme for HCIR systems, and motivate the reasons that lead us to choose one model over the other. We then describe our implementation of the preferred model, and provide experimental results that compare the two classes of models.

II. HCIR Paradigms

There are two distinct models for incorporating human interaction into information retrieval that can be differentiated by the nature of the relationship between the human and the computer. The key difference in relationship is one of dominance: which of the computer or the human is dominant in the particular model. The resulting system takes on the characteristics of the dominant entity, with consequences for the application of such a system to particular information needs.

The first model is one where the human is dominant to the computer. This model arises in a straightforward manner where the human processes are viewed as being distinct from those of the computer: the computer produces artifacts which are then taken up by the human. A common example is a search engine plus user, considered as a system: the combination of the search engine together with the human review of the results extends the capabilities beyond those of the search engine alone. This model gives rise to a form of Computer Aided Human Assessment (CAHA), where the capabilities and limitations of the human are the primary driver of system results. A CAHA system is depicted in Figure 1.

In the second model, the computer is dominant to the human. In this model, human response is solicited by, and incorporated directly into the computer system, but the computer's reponse is determinative of the output. While typical of certain kinds of artificial intelligence systems (*e.g.* [8]), this approach has not been extensively used in information retrieval. This model represents a form of Human Aided Computer Assessment (HACA). An HACA system is depicted in Figure 2.

The key difference between these models is whether it is the computer or the human whose capabilities are primary. Systems corresponding to these models will share the characteristics of the dominant entity. Some important characteristics are:

- **Scalability:** How well does the resulting system deal with large amounts of data, particularly large numbers of relevant documents?
- **Consistency:** How consistent is the system in producing the same output given a particular input?
- **Flexibility:** How does the system react to changes in specification?

Clearly, it is desirable that any system exhibit these characteristics, and indeed, both the CAHA model and the HACA model incorporate elements with these characteristics. However, due to assymetries in the way that scalability, consistency and flexibility are generated and transmitted by computers and humans, the architectures of the two models give rise to distinctly different systems.

Scalability is a property at which computers excel and humans fail. Beyond this, however, scalability that arises by use of a computer is not maintained when a human is required to analyze the results. Thus, a scalable computation followed by a non-scalable human interaction will not be scalable, while a non-scalable human interaction followed by a scalable computation will be scalable. Thus, the CAHA system will not, in general, be scalable, while the HACA system will be.

Consistency is another property at which computers outperform humans but which is destroyed by human intervention. A purely algorithmic approach exhibits perfect consistency given a particular input. Humans exibit no such consistency. In general this means that no combination of humans and computers, neither $Human \rightarrow Computer$ nor $Computer \rightarrow Human$ will be necessarily consistent. The benefit of the HACA system, then, is not that it guarantees consistency, but rather that it provides the opportunity, through careful design, to introduce consistency. CAHA, on the other hand, provides no such opportunity, making consistency an unlikely property of such a system.

Finally, flexibility is a property which is fundamentally characteristic of humans, and for which computers have limited capacity. However, due to their enhanced consistency, computers are remarkably good transmitters of flexibility, while at

the same time poor generators of the same. Thus, both HACA and CAHA systems will exhibit flexibility, the CAHA system because the human is primary, while the HACA because the computer reliably transmits the human capacity for flexibility.

CAHA and HACA represent two approaches to the challenge of incorporating human interaction into information retrieval systems. With respect to the desirable properties of an information retrieval system, HACA can achieve the same degree of flexibility as CAHA, and offers substantially improved scalability and consistency.

III. AN HACA SYSTEM

In this section, we describe an HCIR system that instantiates the HACA properties described earlier. One of the key ways to ensure that a system will exemplify HACA is that human interaction with the system is front-loaded, *i.e.*, precedes the computation, or at least precedes the vast majority of the computation. For an IR system, this effectively means that the computer must be independently capable of providing the desired level of accuracy *per se*, without additional human post-processing. For problems that require high recall, such as the exhaustive research problem, one-shot, ranked retrieval approaches, such as OKAPI have not historically exhibited sufficiently high precision and recall to allow them to fulfill this role unaided.

A. Design Choices

Because the traditional approach yields limited gains, we have made two fundamental design choices that enable our IR system to function as HACA:

- Document Classification, rather than Document Retrieval
- Iteration

Our proposed system treats the problem of identifying documents as one of document classification, rather than document retrieval. In a document retrieval system, a user provides a query which specifies the set of documents to be retrieved. The task of the system is to retrieval documents that match the query. Such retrieval systems typically offer rapid response for relatively short queries, and exhibit a standard trade-off between precision and recall. Unlike document retrieval, a document classification system attempts to generalize from a set of pre-categorized training documents to assign categories to unseen documents. Such systems can exhibit high precision and recall, but at the cost of significant input data requirements: typically hundreds to thousands or more training examples, as opposed to a paragraph-length query.

By itself, however, the choice of a classification, rather than retrieval model is insufficient to guarantee the level of performance necessary to ensure a functioning HACA system. We therefore introduce the design element of iteration. In a non-iterative IR system, the user has only one opportunity to formulate his or her input (a query in the case of retrieval or training documents in the case of classification). In such a system, the user cannot benefit from information about the output the system to improve results. An iterative design ameliorates this shortcoming by enabling multiple passes of
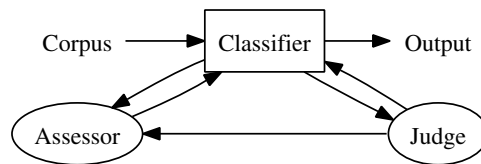

Fig. 3. Iterative, Classification-based HCIR Architecture

user input. In this way, a user can derive information from the output of the system in order to better represent his or her information need to the system. Our system is able to make good use of many (≈ 10) iterations of user input to drive performance to desired levels.

Design choices such as these are necessarily dictated by the expectations of the user *vis à vis* his or her information need. For a known-item or exploratory search, users' expectations are such that rapid system response is relatively more important than accuracy. In such cases, the overhead incurred by requiring training documents and iteratively refining them is unacceptable. For exhaustive research, on the other hand, and especially within certain domains, users are willing to trade off response time in exchange for substantially improved performance. In such cases, the greater effort required to craft training exemplars and to refine the results they generate will be more than made up for by the improved performance of the resulting system.

B. Architecture

The architecture of our iterative, classification-based HCIR system is shown in Figure 3. In addition to the classification component, also pictured are the interations and processes that drive classification. Note that computational components are shown with rectangles, while human interactions are depicted using ovals.

One important characteristic of this architecture is that classification directly determines final output. That is, after any number of iterations of human interaction have been completed during development, it is the output of classification, a computation, that constitutes the system output. Thus, the fundamental requirement of HACA, that the computation is not further elaborated by the human, is fulfilled.

In order to generate training documents so as to effectively generalize the user's notion of categorization, the classification component requests human interaction in order to assess documents as relevant or not relevant for the category under consideration. The human performing this assessment represents, or at least has access to, the ultimate user of the system in order to determine what relevance means within the context of user's information need.

Iteration is implemented by feeding output from the classification component back into the assessment iteration. More explicitly, iteration is effected by requesting human interaction for a process of measurement whereby a portion of documents in any given iteration are reviewed for accuracy by human judges. Discrepancies between the outputs of the classification

Algorithm:
loop
    Classifier requests that Assessor categorize documents
    Assessor responds with categorized documents
    Classifier learns from categorized documents
    Classifier categorizes entire Corpus
    Classifier requests that Judge evaluate documents
    Judge evaluates documents
    if performance target has been reached **then**
        stop, return Output from Classifier
    end if
end loop

Fig. 4. Operation of the HACA System

and the true relevance, as determined by the judges are scrutinized to determine how assessments should be modified in order to better align the outputs of classification with those of the judges.

The operation of the system is show in Figure 4.

IV. EVALUATION

We evaluated the above system in the context of the TREC Legal Track Interactive Task. The Legal Track was established by TREC to evaluate approaches to information retrieval with application to the problems encountered in the legal world. The Interactive Task was formulated in order to provide a more realistic setting for the information retrieval task, and includes the following features: [9]

- A Topic Authority is the sole determiner of relevance
- Teams have 10 hours to interact with the Topic Authority
- Teams must submit a binary classification for every document in the population
- Teams can appeal assessment decisions with Topic Authority making the final decision

The Interactive Task was evaluated on the IIT Complex Document Information Processing (CDIP) Test Collection, comprising 6,910,192 documents released under the tobacco "Master Settlement Agreement".

Given this context, we evaluated on the following topic:

Topic 103. All documents which describe, refer to, report on, or mention any "in-store," "on-counter," "point of sale," or other retail marketing campaigns for cigarettes.

It should be noted that this task and topic represent a substantially challenging form of exhaustive search. Results are evaluated using the F1 measure, balancing the need for precision and recall. In order to perform well on this metric, it is necessary to achieve both high precision (which is easy) and high recall (which is difficult). The topic is underspecified and subtle: just what exactly constitutes "retail", "marketing" and "campaigns" is unclear: substantial human interaction will be necessary to communicate with the topic authority and communicate relevance to the system. Finally, the scale of the task (786,862 relevant documents in a population of 6,910,192) and time allocated for completion (12 weeks) is such that a

	Recall	Precision	F1
HACA	**0.786***	**0.830**	**0.808***
CCPool	0.528*	0.422	0.469*
CAHA$_1$	0.236*	0.733*	0.358*
CAHA$_2$	0.085*	0.697	0.152*
CAHA$_3$	0.042	0.823*	0.080

TABLE I
EVALUATION OF HACA VS. CAHA[2]

purely human approach would be impossible: the scalability of computational methods are absolutely necessary.

To run the experiment, human interaction was needed to make assessments on 7992 documents, providing substantial evidence for classification The resulting system determined that, of the 6,910,192 documents in the population, 608,807 (8.8%) were relevant, and the remaining 6,301,385 (91.2%) non-relevant.

These results were evaluated in comparison to other systems using independent human assessments of relevancy provided by law students who reviewed a stratified sample of 6500 documents provided by four systems [10]. Human assessments were checked for accuracy by the topic authority according to an adjudication procedure. According to the sample, an estimated 786,862 (11.4%) documents were relevant for the topic in the population.

In addition to the HACA system described in this paper, three CAHA systems were also compared (CAHA$_1$, CAHA$_2$ and CAHA$_3$), as well as a pool of non-HCIR systems (CCPool). Results are shown in Table I.

The HACA system performs best among all systems across all three metrics, although only Recall and, as a consequence, F1 are significantly higher than the next higher system, CCPool. In general, all systems perform well on Precision, with the exception of CCPool, which would not be expected to exhibit high precision because it aggregates results from a pool of systems. However, there are only three significantly distinct groups of systems based on precision: {HACA, CAHA$_3$}, {CAHA$_1$, CAHA$_2$} and {CCPool}. Finally, note that it is on Recall that HACA truly shines: not only does it exceed the CAHA systems, but also exceeds CCPool, which would be expected to perform well on the recall metric.

V. CONCLUSION

We have proposed a novel classification scheme for human-computer information retrieval systems that distinguishes Computer-Aided Human Assessment (CAHA) from Human-Aided Computer Assessment (HACA), and have shown how the HACA model better exhibits the desirable properties of scalability, consistency and flexibility than the CAHA model. We have exhibited an implementation the HACA system, explaining how design choices made for the system contribute to the desired goals. Finally, we provided results of a large-scale evaluation that demonstrate the increased effectiveness of the HACA model.

[2]Entries with * are significantly higher than the next lower value at the 0.05 level.

REFERENCES

[1] L. Rosenfeld and P. Morville, *Information Architecture for the World Wide Web*. O'Reilly, 2002.

[2] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Computer Networks and ISDN Systems*, 1998.

[3] T. Lau and E. Horvitz, "Patterns of search: Analyzing and modeling web query refinement," in *Proceedings of the seventh international conference on User modeling*, 1999.

[4] B. Shneiderman, "The future of information discovery," Keynote at Information Retrieval/Search Technologies Seminar, 2008.

[5] C. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[6] G. Marchionini, "Toward human-computer information retrieval," *Bulletin of the American Society for Information Science and Technology*, June/July 2006.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford Infolab, 1999.

[8] R. Burgener, "20q: The neural network mind reader," Presented at Goddard Space Flight Center Engineering Colloquium, 2006.

[9] J. Baron, B. Hedin, D. Oard, and S. Tomlinson, "TREC-2008 legal track interactive task — guidelines," Available online at: http://trec-legal.umiacs.umd.edu/2008InteractiveGuidelines.pdf, 2008.

[10] D. Oard, B. Hedin, S. Tomlinson, and J. Baron, "Overview of the TREC 2008 legal track," in *Proceedings of The Seventeenth Text REtrieval Conference (TREC-2008)*, 2008.