Towards Large-Scale Collaborative Planning: Answering High-Level Search Queries Using Human Computation

Edith Law
Machine Learning Department
Carnegie Mellon University
edith@cmu.edu

Haoqi Zhang*
School of Engineering and Applied Sciences
Harvard University
hq@eecs.harvard.edu

Abstract

Behind every search query is a high-level mission that the user wants to accomplish. While current search engines can often provide relevant information in response to well-specified queries, they place the heavy burden of making a plan for achieving a mission on the user. We take the alternative approach of tackling users' high-level missions directly by introducing a human computation system that generates *simple plans*, by decomposing a mission into goals and retrieving search results tailored to each goal. Results show that our system is able to provide users with diverse, actionable search results and useful roadmaps for accomplishing their missions.

Introduction

Human computation is the study of systems where humans play an integral part in the computational process. For example, players of the ESP Game are essentially computing a function that maps images to labels, as a by-product of playing an enjoyable game. Effective human computation systems recognize humans' ability to easily solve some problems that are still difficult for existing AI algorithms (e.g., recognizing objects in images). By involving humans in the loop to solve these problems, human computation systems seek to accelerate the development of useful technologies (e.g., image search), and to provide new insights that help to advance AI research.

Web search is a difficult AI problem. To date, research on Web search has focused primarily on improving the relevance of search results to a query. However, people use the Web not only to retrieve information, but to solve short-term or long-term problems that arise in everyday lives. While current search engines are able to provide relevant information in response to well-specified queries, it places the heavy burden of actually solving a problem (e.g., figuring out what steps to take, how to accomplish these steps, and what queries to enter to find helpful resources) entirely on the user. For a user with a *mission* in mind, e.g., "I want to get out more," or "I need to manage my inbox better," a typical search scenario today would involve the user digging through a set of blogs, opinion or "how-to" articles on the Web in order to identify important sub-problems, and then submitting *multiple* search queries to find resources for addressing each sub-problem.

We envision the next generation of search engines to more closely resemble interactive planning systems, that are able to take in high-level mission statements (e.g., "I want to . . .," "I need to . . .") as input, and directly generate plans to achieve these missions. For example, a *simple plan* may detail specific steps to take, provide explanations for why these steps are important, and return relevant resources for accomplishing each step. A more complex plan may even include conditional branches and recourse decisions, e.g., to handle situations when a step does not work as intended. Unfortunately, the gap between the capabilities of current search engines and the envisioned next-generation search engines is huge – a system as described has to not only understand natural language mission statements, but also be equipped with large amounts of common-sense and real-world knowledge on how to solve specific problems of interest.

To fill this gap, we introduce *CrowdPlan*, a human computation algorithm that takes as input a high-level mission and returns as output a simple plan that captures the important aspects of the user's problem. CrowdPlan leverages human intelligence to decompose a *mission* into low-level *goals*,[1] which are then mapped into queries and passed onto existing search engines. The output is a simple plan consisting of a set of goals for tackling different aspects of the mission, along with search results tailored to each goal. For example, the high-level mission "I want to live a more healthy life" can be decomposed into a variety of goals, including "stop smoking," "eat healthier food," "learn to cook at home," "exercise," "follow a diet plan," "drink less alcohol," "spend time with friends," "sleep well" etc. Each of these goals, in turn, can be supported by one or more search queries. For example, "exercise" can be supported by queries such as "running shoes," "best bike routes," "personal trainer" etc. In this paper, we describe in detail how we obtain goals and queries and generate simple plans using CrowdPlan, and discuss the strength and pitfall of our approach using results from three experiments. Results show

[*] Both authors have equal contribution in this work.

[1] We adapt the definitions in (Jones and Klinkner 2008), and define a *goal* as "an atomic information need, resulting in one or more queries" and a *mission* as "a set of related information needs, resulting in one or more goals."

that CrowdPlan is a promising human computation algorithm for generating simple plans that help users accomplish their missions. We found that the subjects were split in their preference for CrowdPlan versus standard search, and that their preference can be attributed to different reasons. In particular, users who prefer standard search find the retrieved results to be more relevant, authoritative and informational; users who prefer CrowdPlan find the retrieved results to be more diverse and actionable, and that the simple plans provide useful roadmaps for accomplishing their missions and highlight aspects of the problem that might have been missed otherwise.

Related Work

Human computation systems have typically been used to distribute *independent* tasks, where the output of one worker does not affect another worker. With the development of TurKit (Little et al. 2010) and other toolkits that provide programmatic access to Amazon's Mechanical Turk (`www.mturk.com`), researchers are now equipped with the ability to write algorithms (Parameswaran et al. 2011) that involve humans in the loop – to perform basic operations, check conditions, decompose problems, compose results, generate useful heuristics for optimization, etc.

Recently, a few collaborative planning websites such as `ifwerantheworld.com` and `43things.com` have emerged that allow users to share their goals and come up with solutions together. At ifwerantheworld.com, users can enter mission statements (e.g., "If I ran the world, I would place more recycling boxes on campus"), and generate *action platforms* consisting of micro-actions that other users can perform. In contrast, at 43things.com, users can enter their resolutions or goals, and support each other (e.g., by posting suggestions or progress reports in a forum) in their attempts to accomplish their goals. Different from these approaches, we aim to build collaborative planning systems with *explicit algorithms* that exert finer control over the process – deciding what to compute, how and by whom – instead of leaving the planning process entirely up to the dynamics of the crowd.

One of the biggest challenges search engines face is understanding user intent. Search queries can be ambiguous (e.g., "jaguar," "windows," "java") and under-specified; currently, the burden is on search engines to figure out what users really mean. There has been a substantial amount of research on predicting user intent from query logs (Baeza-Yates, Calderon-Benavides, and Gonzalez-Caro. 2006; Lee, Liu, and Cho 2005; Rose and Levinson 2004). This is a difficult task for two reasons. First, intent prediction requires training data in the form of intent-to-queries mappings, which is difficult, if not impossible, to obtain. Typically, researchers resort to manually labeling hundreds to thousands of search queries (Baeza-Yates, Calderon-Benavides, and Gonzalez-Caro. 2006) in order to build such training sets. Second, search queries belonging to the same intent are often issued over a span of days or weeks, and interleaved with queries belonging to other intents, making the prediction task even more difficult (Jones and Klinkner 2008). In contrast, our system, which takes user intents (or missions) as inputs and generates search queries as outputs, can produce these intent-to-queries mappings as a by-product.

Finally, our work is related to research on diversifying search results (Santos, Macdonald, and Ounis 2010; Clarke et al. 2008; Agrawal et al. 2009; Gollapudi and Sharma 2009; Chen and Karger 2006) as a way to better answer *faceted* queries (e.g., "wedding planning"), which are queries that encompass a diverse set of information needs. Diversification of search results can be seen as a bi-criterion optimization problem (Gollapudi and Sharma 2009) that aims to maximize relevance and diversity in the search results, while minimizing redundancy. This is also sometimes called aspect or subtopic (Zhai, Cohen, and Lafferty 2003) retrieval, where each query is associated with a topic, and the goal is to return a ranked list of search results that provide good coverage of the subtopics of the query topic. There are two approaches to diversification (Santos, Macdonald, and Ounis 2010) – the implicit approach assumes that similar documents cover similar aspects and attempts to demote the ranking of similar documents already covered. In contrast, the explicit approach aims to discover the aspects of a query explicitly, then retrieve relevant documents belonging to each of those aspects (Santos, Macdonald, and Ounis 2010). Our work also takes the explicit approach; however, we elicit the help of human workers to decompose high-level queries into aspects that are not merely *related* (Santos, Macdonald, and Ounis 2010) to the mission, but that are *actionable*, i.e., help users take concrete steps towards achieving their missions.

CrowdPlan

The CrowdPlan algorithm takes a high-level user mission m and generates a simple plan $\mathcal{P}_m$ for accomplishing the mission. A simple plan (e.g., Figure 3(b)) consists of a set of tuples $(g_i, \mathcal{R}_i)$, where g_i is a goal relevant to the mission and $\mathcal{R}_i$ is a set of resources, e.g., search results, associated with the goal g_i. Figure 1 depicts the CrowdPlan algorithm, showing the human-driven and machine-driven operations in grey and white boxes respectively, including:

- **decompose**: given a high-level mission m and a set of previous goals $\{g_1, \ldots g_n\}$, this operation generates an additional goal g_{n+1} that is relevant for the mission, but different from already stated goals.
- **rewrite**: given a high-level mission m and a goal g_n, this operation generates a search query q_n for finding web resources that help to achieve the goal g_n.
- **assess**: given a high-level mission m and a set of tuples (g_n, q_n), $n = 1 \ldots N$, this operation returns an assessment vector $\vec{a} = \{0, 1\}^N$ where bit i indicates whether the search query q_n is likely to return good search results towards accomplishing goal g_n.
- **filter**: given a set of assessment vectors $\vec{a}_1, \ldots, \vec{a}_L$ provided by L workers, this operation aggregates the votes and returns a set of the highest quality search queries $\{q_{i_1}, \ldots, q_{i_K}\}$ to retain.
- **search**: given a search query q_{i_j}, this operation retrieves a set of search results $\mathcal{R}_{i_j}$ associated with the query.

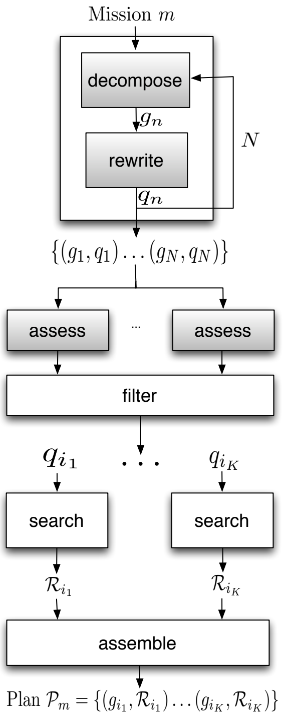


Figure 1: CrowdPlan

- **assemble**: this operation returns a simple plan that consists of a set of tuples $\{(g_{i_1}, \mathcal{R}_{i_1}) \dots (g_{i_K}, \mathcal{R}_{i_K})\}$ to present to the user. Note that this plan can be presented to the user using different forms of visualization.

Each of the human-driven operations (shown in grey in Figure 1) – *decompose*, *rewrite*, and *assess* – is associated with a small task that is distributed to workers on Mechanical Turk (Turkers).[2] The *decompose* and *rewrite* operations are combined into a single HIT (human intelligence task), wherein a worker is given a high-level mission and a set of existing goals, and paid 10¢ to first generate an additional goal relevant to the mission, and then rewrite the goal as a search query. Combining these two consecutive operations into the same HIT allows a Turker to work off his or her own goal when formulating a query (instead of having to interpret and rewrite someone else's), thus simplifying the computation. For each mission, we obtain up to 10 goal-query pairs. The *assess* operation is associated with a HIT wherein a worker is paid 10¢ to cross out any search queries that are unlikely to take a step towards accomplishing the mission, and to discuss how useful the remaining queries are. Each search query is clickable and links directly to a webpage containing the search results returned by Google for that query.

The machine-driven operations include *filter*, *search* and *assemble*. The *filter* operation eliminates potentially problematic search queries as follows. Each query is assigned a removal score $s_q = n_q + vn_q - vp_q$, where n_q is the number of people who gave a negative assessment for that query, vn_q is the number of people who reviewed the search query (by clicking on the link to bring up the search results) before giving a negative assessment for that query, and vp_q is the number of people who have reviewed the search query before giving a positive assessment for the query. This scoring scheme essentially gives a query an additional vote (in favor of it being filtered) for each judge who has actually reviewed the search query carefully before giving a negative assessment. We request five *assess* HITs per mission, and filter out a query if its score is ≥ 3. The remaining queries are ranked by their scores in ascending order.

The *search* operation uses the Google Search API to retrieve eight search results for each query. The *assemble* operation then puts together a simple plan, consisting of goals and search results, to display to the user. Starting with the highest quality queries (i.e., those with the lowest removal score) and continuing in a round-robin fashion, we collect one search result per query until we have a set of 10 *unique* search results. If we encounter a search result that has been returned for more than one query, we associate the result with the lower ranked query (since lower ranked queries are unlikely to have as many good results) and use the second best result for the higher ranked queries. Note that the round-robin selection process ensures that the search results for the same query are well separated, allowing the diversity of the results to show through first. The sort order also puts the better search queries higher in the ranking.

The design choices we made in creating this particular algorithm, or *workflow*, was influenced heavily by our observations of how workers responded to the task. For example, the *decompose* operation could have followed a top-down approach, in which workers first provide a coarse representation of the mission (e.g., "I want to throw a Thanksgiving party") by naming a few goals that encompass the entire solution (e.g., "cook dinner," "invite people," "plan activities"), then provide successively finer-grained subgoals to accomplish each of the goals. However, in our initial experiments we found that Turkers did not operate at that level of abstraction, i.e., they often provided goals that did not require further decomposition. Therefore, we made the *decompose* operation to be more akin to an iterative, brainstorming task in which workers are asked to come up with concrete goals towards accomplishing the mission.

Our algorithm is implemented in Javascript, and uses TurKit (Little et al. 2010) to interface with Mechanical Turk.

Experiments

In order to evaluate how well our system can answer high-level queries, we asked 14 subjects to each give us two mission statements in the form of (i) a new year resolution or life goal, and (ii) a concrete task that they want to accomplish. Subjects are mostly recent college graduates who did not major in computer science, and were told that we are working on an information retrieval system that can help answer high-level search queries. Subjects were told that their

[2] We envision that the CrowdPlan algorithm can eventually be embedded as part of collaborative planning websites that have access to tens of thousands of human *volunteers*; but in this paper, we use Amazon Mechanical Turk (AMT) as a platform to recruit human subjects for our experiments.

	New Year Resolutions / Life Goals
1	cook at home more often
2	manage my inbox better
3	become healthier by working out more
4	run a marathon
5	find an academic job in a good research university in the US
6	become a competitive amateur triathlete.
7	be a good (new) mother
8	start song writing
9	get into petroleum engineering/natural gas field
10	get outside more
11	take a trip to the space
12	be happier
13	lose 80 pounds
14	keep in better touch with high school friends
	Concrete Tasks
1	choose a wedding DJ
2	book a great honeymoon for August 7-14
3	figure out where to go on a week-long sailing vacation with nine friends
4	buy a new pair of dress pants
5	survive the Jan-Feb crazy conference deadlines
6	start exercising and follow an appropriate training program (to become a competitive amateur triathlete)
7	finish the bathroom and laundry room in our basement
8	see if Honda will fix my seatbelt for free
9	kick my friend in the arse
10	find a place to live in Toronto
11	finish Need For Speed Hot Pursuit game
12	shower daily
13	go to the market and buy groceries
14	change address on my car insurance policy

Figure 2: mission statements submitted by subjects

missions may be shown publicly, but did not know that human computation is involved. Figure 2 shows the high-level missions we received, which range from very concrete, actionable tasks (e.g., "change address on my car insurance policy") to less specific, long-term aspirations (e.g., "be happier," "be a good (new) mother").

Through a series of three experiments, we evaluate the effectiveness of the CrowdPlan algorithm for generating simple plans that help users accomplish their missions. In particular, the experiments are aimed at addressing the following questions: (1) Can existing search engines (e.g., Google) handle high-level mission statements as search queries? (2) How did the users perceive search results returned by CrowdPlan compared to those obtained by rewriting mission statements as search queries? (3) In what ways are simple plans better or worse than using a standard search interface at helping users accomplish their mission?

Study I: Mission Statement versus Query Rewrites

In the first experiment, our objective is to gauge the capabilities of existing search engines (specifically, Google) at handling high-level search queries, by comparing the relevance of the search results when the mission statements are used directly as search queries (referred to as MQ for "mission queries"), versus when they are rephrased by humans into a set of search queries (referred to as RQ for "rewrite queries"), which are then used to generate the search results. In addition to verifying that current search engines lack the ability to handle natural language mission queries, this study also helps to establish RQ as a strong baseline to compare CrowdPlan against in the subsequent studies.

To rephrase each mission into search queries, we paid five Mechanical Turk workers 5¢ to independently rephrase the mission as a search query and explain why their proposed query is likely to generate good search results for the mission. Rewrite queries are subjected to a similar *assess-filter-search-assemble* process as depicted in Figure 1. Table 1 shows some examples of high-level missions and their associated rewrite queries. We see here and in our results that search queries generated from rewriting often contain words (e.g., "shopping," "store," "review," "apparels," "exercise," "fitness") that were not originally in the mission statement.

Missions	Rewrites
buy a new pair of dress pants	branded trousers and apparels dress pants new dress pants shopping dress pants store dress pants review
become healthier by working out more	fitness strategies exercise techniques tips instruction on daily fitness workout daily workout tips increase exercise frequency

Table 1: Missions rewritten as search queries

For each mission, we asked a set of workers (~7-10) to compare 10 search results returned by MQ versus 10 search results returned by RQ. In the evaluation task, workers are given a high-level mission, and shown two sets of search results in randomized order. They are asked to evaluate the search results and decide which set of search results is (i) more relevant to the mission, and (ii) will actually help someone accomplish this mission. Workers are also asked to explain their votes.

Results show that RQ improves both the relevance and usefulness of the search results. For 17 out of 28 missions, the RQ search results are judged to be strictly more relevant, with an average overall lead of 5.17 votes, although the results are not statistically significant ($p < 0.07$) using the Friedman test. For 20 out of 28 missions, the RQ search results are also judged to be strictly more *useful* towards helping the user accomplish the mission, with an average overall lead of 4.35 votes. Here, the difference is statistically significant according to the Friedman test, with $p < 0.001$.

Study II: CrowdPlan versus Query Rewrites

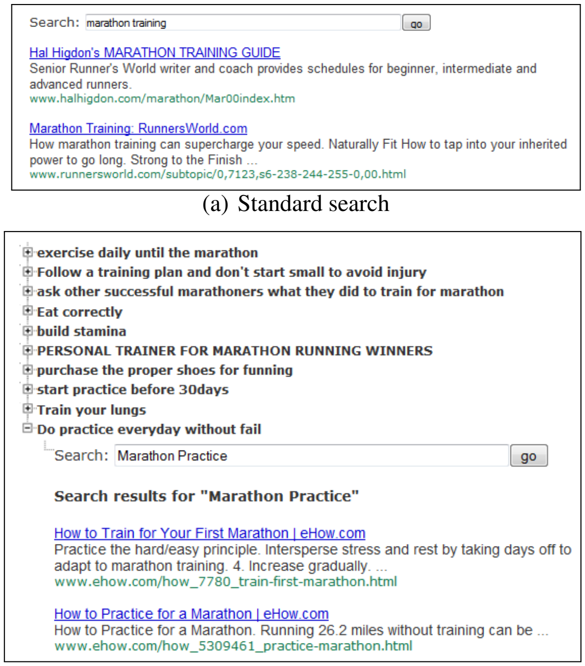
In the second study, our goal is to understand how users perceive the search results returned by (i) our proposed CrowdPlan algorithm, which decomposes high-level mission into goals and rewrites goals into search queries, versus (ii) rewriting the mission statement into search queries without decomposition (RQ).

We asked the 14 subjects to evaluate the CrowdPlan and RQ search results for one of their missions. First, subjects are shown each set of search results and asked to rate (on a 4-point scale) how useful each result is for helping them accomplish their mission. Half of the subjects are shown the RQ search results first, while the other half are shown the CrowdPlan results first. After reviewing the search results individually, subjects are then asked to compare the two sets of search results side by side, and asked (i) which they prefer, (ii) which will help them accomplish their tasks, and (iii) which is more diverse.

Using the paired t-test, we found no statistical difference between the relevance score of the CrowdPlan (mean=1.05 ± 0.66) and RQ (mean=1.27 ± 0.66) search results. In terms of preference for the search results, 7 out of 14 subjects preferred CrowdPlan over RQ. In terms of helpfulness towards accomplishing the mission, 6 out of 14 subjects preferred CrowdPlan over RQ. More importantly, 12 out of 14 subjects thought that CrowdPlan results are more diverse than RQ results.

The most important finding in this study are the *reasons* behind why users are split in their preference for CrowdPlan versus RQ search results. Many subjects who preferred the RQ results commented that the search results are more relevant and authoritative. In contrast, those who preferred the CrowdPlan results pointed out that the search results are more diverse and contain actionable items (e.g., "The results seemed to have more pages that involved actually doing things, whereas [the RQ] results were merely informational," "It had two absolute gems that were exactly what I was looking for. RQ just had general information that was less helpful to me."), or pinpointed certain aspect of the problem that RQ results did not capture (e.g., "Some of them are a better fit for what I'm looking for, [such as] ideas for hiking, inexpensive workout equipment options, how to become motivated to work out, etc").

However, some subjects commented that CrowdPlan brought up aspects of the problem that is not exactly what they are looking for. For the mission "I want to cook at home more often," CrowdPlan returned search results for buying good cookware, researching what kinds of food freezes well, cooking shows to watch, etc; however, the subject commented that he or she was looking specifically for recipes. For the mission "I want to get outside more," CrowdPlan returned search results for taking up gardening, birdwatching, daily walks, geocaching, and adopting a dog, while the subject commented that he or she was looking for "websites geared toward more active outdoor activities in natural surroundings." These observations suggest two things. First, there is a need for personalization – even though the user intents are clear, the kinds of solutions that each user is seeking can vary. Additional context may allow CrowdPlan to tailor the simple plans more appropriately to each user. Second, CrowdPlan tends to return search results covering a diverse set of issues related to the mission, and a potential drawback of the increased diversity is decreased comprehensibility. The steps for solving a particular problem sometimes appear tangential, or even irrelevant, if they are not properly explained. For example, one of the suggested queries for spending time outdoors is "ALTA," which refers to a nonprofit tennis organization. The goal that is associated with this query is actually "take up tennis." Without this explanation, it is difficult for the user to know why the search result for ALTA would be relevant to his or her high-level mission.


Search: marathon training

Hal Higdon's MARATHON TRAINING GUIDE
Senior Runner's World writer and coach provides schedules for beginner, intermediate and advanced runners.
www.halhigdon.com/marathon/Mar00index.htm

Marathon Training: RunnersWorld.com
How marathon training can supercharge your speed. Naturally Fit How to tap into your inherited power to go long. Strong to the Finish ...
www.runnersworld.com/subtopic/0,7123,s6-238-244-255-0,00.html

(a) Standard search

⊞ exercise daily until the marathon
⊞ Follow a training plan and don't start small to avoid injury
⊞ ask other successful marathoners what they did to train for marathon
⊞ Eat correctly
⊞ build stamina
⊞ PERSONAL TRAINER FOR MARATHON RUNNING WINNERS
⊞ purchase the proper shoes for funning
⊞ start practice before 30days
⊞ Train your lungs
⊟ Do practice everyday without fail

Search: Marathon Practice

Search results for "Marathon Practice"

How to Train for Your First Marathon | eHow.com
Practice the hard/easy principle. Intersperse stress and rest by taking days off to adapt to marathon training. 4. Increase gradually. ...
www.ehow.com/how_7780_train-first-marathon.html

How to Practice for a Marathon | eHow.com
How to Practice for a Marathon. Running 26.2 miles without training can be ...
www.ehow.com/how_5309461_practice-marathon.html

(b) Simple plan in list view for the mission "run a marathon"

Figure 3: Standard Search versus Simple Plan

Study III: Simple Plans versus Standard Search

One of the benefits of the simple plans generated by CrowdPlan is that they provide an explanation (in the form of goals) for the search results returned to the user. To study the effect of explanations, for each mission, we asked 10 Turkers to rate the relevance of the CrowdPlan results, where half of the workers were given explanations and the other half were not. Turkers were paid 20¢ per HIT. Results show that when given explanations, workers judged the search results to be more relevant, both in terms of the average relevance score and discounted cumulative gain (Järvelin and Kekäläinen 2002) (RS = 1.92, DCG = 9.7), than when not given explanations (RS = 1.75, DCG = 8.7). For both metrics, the difference between the relevance judgments for explained and unexplained results are statistically significant using the paired t-test test ($p < 0.006$ and $p < 0.005$).

In light of this observation, we created a simple list-view visualization of the simple plan (see Figure 3(b)), which displays the decomposed goals for the mission, the search query associated with each goal, and a short list of five search results. Users can expand on each goal to view the current search results, and modify the search query to generate more search results associated with a particular goal.

To evaluate the effectiveness of this interface, we asked our 14 subjects to spend 3 minutes using (i) a standard search engine (Figure 3(a)) and then (ii) a simple plan in list view (Figure 3(b)) to find web resources to help them achieve their high level missions. We then asked subjects to compare the two interfaces in terms of how well they help them accomplish their goal. We again find a split in opinion – 7 out of 14 subjects prefer our search interface over the standard interface. Subjects who preferred the standard search interface commented that it was more "straightforward" to use, generated more "one-stop" search results (i.e., general purpose websites with links to resources), and that the simple plans generated some search results that are irrelevant to what they were looking for specifically. Here are some comments:

- I like the idea behind simple plans, but I find it more straightforward to use a regular search tool.
- The standard search tool was better because I knew enough about what I wanted that I could type in more specific searches.
- I think many good websites will give me a one-stop shop for marathon information. The simple plan was fairly comprehensive although perhaps too specific.
- I can see how search tool simple plans could be more useful in a multistep resolution. It wasn't quite so for something as ambiguous as songwriting. The standard search engine was much better in this case. Some of the hits in the simple plan didn't make any sense.

In contrast, subjects who preferred simple plans over standard search results had the following comments:

- The simple plan actually organized my search for me, into discrete and doable steps. The standard search tool left me to do all the creative parsing and generation of search terms. I felt that the simple plan gave me a roadmap to the entire space by my mentioning something in that space.
- The simple plan gave me some good ideas for concrete steps to take that would help me accomplish my goal. Therefore, the search queries were more focused, and the overall process more effective.
- The simple plan made me consider a variety of different aspects of achieving my goal, rather than me having to come up with search terms.
- The simple plan provided an outline of different aspects of completing my task.
- The simple plan gave me a birds-eye view of useful search queries from which to pick. the recommendations were really useful. My reaction to some of them was "oh, I didn't think of that. good point!" The simple plan solves to some degree the problem of unknown unknown, which is that in order to find something you need to know you need it. This problem makes the standard interface of limited use, because you need to know a priori what you have to do in order to find instructions on how to do it. But the simple plan, being broader in its results, suggests things you didn't think of.

These comments are revealing for several reasons. First, they show that not all missions require decomposition, and given mission statements rewritten as well-specified search queries, the standard search engines may already be quite good at retrieving relevant results. Second, they show that simple plans are effective in three ways – making users aware of aspects of the problems they had not originally thought of, providing an organized roadmap for solving a problem, and suggesting concrete, actionable steps towards accomplishing the mission.

Conclusion

We introduced CrowdPlan, a human computation algorithm for answering high-level search queries. While from a planning perspective this is a relatively simple scenario where the users are given only simple plans that show a set of goals relevant to the mission and the associated web resources for supporting each goal, CrowdPlan nevertheless serves as an interesting and useful system. Moreover, this work serves as an initial exploration into collaborative planning, wherein future works will seek to generate more elaborate plans (e.g., fully ordered plans with conditional plans to handle unforseen situations) using humans in the loop for particular applications of interest.

Our results show that the current search engines do a poor job of handling mission statements as search queries. In particular, we showed that search results generated by rewriting mission statements into search queries produce significantly more useful results. In addition, the CrowdPlan algorithm, which decomposes missions into goals and rewrites each goal into search queries, generated search results that were preferred by half of our subjects because they contain diverse, actionable steps for accomplishing the missions. Finally, we found that search results, when accompanied with an explanation helps users navigate the space of their problem. As future works, we plan to work on personalized simple plans, as well as other helpful visualizations for displaying the simple plans to users.

Acknowledgments

We thank Rob Miller for his support. Edith Law and Haoqi Zhang are generously funded by a Microsoft Graduate Research Fellowship and NDSEG Fellowship respectively.

References

Agrawal, R.; Gollapudi, S.; Halverson, A.; and Leong, S. 2009. Diversifying search results. In *WSDM*.

Baeza-Yates, R.; Calderon-Benavides, L.; and Gonzalez-Caro., C. 2006. The intention behind web queries. In *SPIRE*, 98–109.

Chen, H., and Karger, D. 2006. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR*, 429–436.

Clarke, C.; Kolla, M.; Cormack, G.; Vechtomova, O.; Ashkan, A.; Buttcher, S.; and MacKinnon, I. 2008. Novelty and diversify in information retrieval evaluation. In *SIGIR*.

Gollapudi, S., and Sharma, A. 2009. An axiomatic approach for result diversification. In *WWW*.

Järvelin, K., and Kekäläinen, J. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.* 20:422–446.

Jones, R., and Klinkner, K. 2008. Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In *CIKM*.

Lee, U.; Liu, Z.; and Cho, J. 2005. Automatic identification of user goals in web search. In *WWW*, 391–401.

Little, G.; Chilton, L.; Goldman, M.; and Miller, R. 2010. Turkit: Human computation algorithms on mechanical turk. In *UIST*.

Parameswaran, A.; Sarma, A. D.; Garcia-Molina, H.; Polyzotis, N.; and Widom, J. 2011. Human-assisted graph search: It's okay to ask questions. Technical report.

Rose, D. E., and Levinson, D. 2004. Understanding user goals in web search. In *WWW*, 13–19.

Santos, R.; Macdonald, C.; and Ounis, I. 2010. Exploiting query reformulations for web search result diversification. In *WWW*.

Zhai, C.; Cohen, W.; and Lafferty, J. 2003. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *SIGIR*.