Using crowd-sourcing for query classification and analysis

Keywords: query, web search, crowd sourcing, classification

1 Introduction

In order to gain a better understanding of users and their intent behind web searching activities, first steps involve the analysis of the query submitted by the user and correct categorical classification of the query as an input for further analyses. Natural Language Processing (NLP) is an area with many inaccuracies for problem areas in Word Sense Disambiguation (WSD) and terms detected that are beyond the scope of the verification source[16]. Other difficulties include named entity recognition where capitalisation and language anomalies cause issue. These classic information retrieval issues are all relevant to query classification and this proposal will outline another approach to classification that uses the intelligence and judgement of people to counter the aforementioned challenges. There is clearly potential to resolve tasks with crowd-sourcing that require complex judgement and have many input factors for consideration, but with web search queries, the subjectivity is less prominent depending on the query type.

The work outlined here describes experimentation with the use of Amazon Mechanical Turk (MTurk) as a classification tool. 1200 queries have been carefully selected and an experiment will be set up in MTurk to classify those queries into categories derived from various sources. Once a given query has been classified (according to a minimum inter-ranked consistency for that query), the query will subsequently be subjected to classification into sub-categories of its confirmed, higher level category, eventually resulting in its removal from the system when its classification is definitive.

This is a subset of work contributing to the study in personalisation of web search activities and is a preliminary stage of understanding user intent that is required for further advancement into evaluating current personalisation techniques and developing new or modified means of best delivering search results based on various personalised factors.

2 Research Problem

The explicit nature of user evaluation means that the resource must be carefully considered to ensure quality of evaluations so the various costs involved such as time, effort and finances are not misused. Therefore, perhaps the most challenging problem with using the wisdom of crowds to classify information is in fact in measuring the level of wisdom delivered and the legitimacy of the results obtained from users.

In this work, the validity of MTurk workers' judgements will be assessed in two ways. Firstly the inter-rater agreement will be calculated so that the variation among participants can be assessed. Secondly, ground truths are to be undertaken, where a number of colleagues of the author separately classified the queries. This also provides a level of inter-rater consistency among the groundtruthers which can be compared with that of the MTurk participants. Finally, the experimental questions were constructed so as to identify consistently poor clicking from individual participants.

The classification has been trialled with a smaller set of queries and the preliminary results show that there is a distinct agreement within high level categories of queries that indicates that MTurk is promising for query classification at lower levels should measures be added to handle more complex queries. There was still some consistency with difficult classifications despite the reduction in agreement overall.

2.1 Related Work

A significant amount of research involving crowd-sourcing for various tasks has been conducted, from web-centric applications such as attempting to measure the highly subjective issue of search to result relevance[1] to annotating images[14], the latter of which is in direct contrast to implicit image labelling methods explored in previous work[2]. Pure web search query classification according to a tiered query taxonomy has not yet seen the application of crowd intelligence as a resource, despite other similar classification systems involving queries of some manner being studied. Some examples include the relevance of news queries to articles[12], the disambiguation of ambiguous terms (similar to queries) [1] and development of high quality training data for NLP using crowd-sourcing for manual verification[6]. Kittur et al. [11] evaluated the reliability of MTurk with
subjectivity of relevance to Wikipedia articles, using ground-truthing as a baseline to measure MTurk Workers. This study will take a similar approach as has been done in previous work completed by the author[17]. There is consistent agreement in related work that MTurk is a faster, cheaper alternative to specialist methods and promisingly more accurate than reliance on automated systems requiring complex processing to achieve a similar results to that which humans can provide. Each of the works nevertheless underline the importance of imposing restrictions on MTurk workers to reduce spam and obtain quality ratings, but also to ensure the workers understand the task at hand so the relevance of the instructions provided is relative to the outputs received.

3 Methodology

The work ahead proposes that queries of real-world nature be manually classified utilising crowd-sourcing as the means in order to evaluate the viability and accuracy of such classification methods against a ground-truth and in comparison with automated query classification seen in recent research by Kathuria et al. [10]. In fact, the manual classification may be used as training for automated methods, or a hybrid system may be devised with one complementing or refining the other approach for best practice.

3.1 Analysis

The key questions involved are in determining whether the classifications made are of suitable quality. Therefore, analysis will include statistical measures to inspect the quality of judgements from MTurk workers, and infer whether the judgements are reliable and can be considered truthful and of significant quality such that the assertion can be made that crowd-sourcing is in fact a viable resource. Fleiss Kappa[5] will provide the inter-rater agreement measures based on the categorical judgements made by users.

3.1.1 Trap questions

The use of seeded trap queries to detect and therefore mitigate spam will be utilised in a standard format covering three stages. Firstly, selected queries with correct answers will be deployed in the mix. Human ground-truthing by nominated experts will provide these answers, with such queries only being selected for monitoring if full agreement between the experts exists. Users consistently incorrectly classifying these queries will be regarded as spamming, particularly where such queries are replicated in the examples and instruction set provided prior to commencement. Secondly, by providing descriptions of the various types of queries and their sub-categories, a mix of both accurate and false descriptions can be provided against the category name to determine whether the user has any intuition, or correctly understands the classification required. This would be implemented in a pre-stage and only users with a high success rate for classification would be permitted to continue. Finally, workers that consistently provide incorrect answers to the trap questions will be stopped from commencing or resuming classifications. An example may be the following question:

Where would a navigational query lead the user to from a search?
a) A website
b) Wikipedia
c) Another search engine
d) Disneyland

3.2 Stage One

Prior to the submission of queries to MTurk, the caveats of such a system require careful consideration to ensure the data received is free from pollutants that would affect the credibility of the obtained data. Initial spam prevention measures will include the parameters available from the MTurk interface, which include country restrictions (prior experience found restricting to first world countries, such as the US reduced spam significantly) and high rater reputation (>95%).

- 1200 queries will be classified at the highest level initially (Informational/Transactional/Navigational), described in Section [7]
- $0.01 per classification, equating to approximately $6.00/hr

3.2.1 Query Selection

Challenges behind the research of user intent in web search include the sourcing of data. Where web logs are the obvious source of such information, privacy concerns[7] and major search organisations willing to provide such data as a result are sporadic at best. Queries for such research need to be real world, generated from real
users and preferably from a natural environment where users are unaware such information is being collected. This study will use queries from the 2006 AOL data set \[13\] due to its continued use in research publications and will see limited use of unique and isolated queries only. A sampled selection of queries from this data-set raises the question of whether the queries selected are representative of the overall data-set with regards to popularity, frequency and categories. As a result, this study will employ the following methods to ensure the queries selected are sufficiently meet the requirements.

**Initial Random Sampling** 300 queries taken at random from a table of 10,154,411 unique queries, which are then arranged into 20 buckets consisting of 15 queries each.

**Random Sampling, Evenly Distributed by Raw Frequency** Another 300 queries are taken from the table of unique queries and ordered by their raw frequency, producing a power-law distribution whereby a small number of very popular queries dominate the relatively unpopular queries (80-20 rule). This query pool is to be divided into 20 equal sized buckets of 500,000 queries whereby 15 queries are randomly selected from each bucket. The aim is to create a series of buckets of samples equally distributed in terms of raw frequency.

**Random Sampling, Evenly Distributed by Unique Use** The approach described above captures the raw frequency of the queries but could lead to some inconsistencies. A user submitting the same uncommon query repeatedly would bias the query into a high ranking of the highest frequency table. If we use unique query selection, each users’ query submission would count just ones in the query selection process. Another 300 queries from a table of unique AOL queries ordered by ‘unique use’ frequency and cut into 20 equal sized buckets of 500000 queries, with 15 queries randomly selected from each bucket.

**Random Sampling, Evenly Distributed by Cumulative Unique Use** By taking 300 queries from a table of unique queries ordered by ‘unique use’ frequency, dividing them into 20 buckets with 15 selected from each bucket the series of samples are equally distributed in terms of cumulative frequency form unique users. This involves splitting the query pool up using equal slices of the area under a graphed Frequency-Rank curve, as shown in Figure 1.

![Figure 1: Graph showing the division of the query pool according to frequency and rank](image)

### 3.3 Stage Two

Once the high level classifications have been confirmed based on the level of confidence that is dismissive of classification by chance, the subcategories of queries devised by Jansen et al. \[9\], described in Section 7 will be used for further and more specific classification of the query set with similar restrictions and calculations used in the first stage. Some important considerations and outcomes that are expected to arise in the second stage are as follows.

- Queries may not belong to a sub-category at all, or the user may have trouble making this classification. This must be considered in revising the instructions provided.
- Some queries may present in more than one sub-category, where the higher level classification was not formed with absolute majority.
- Discovery of new sub-categories might be possible for queries that fail to fall under any sub-category (an option may be implemented to allow the user to skip the query due to difficulty in classification).
3.4 Stage Three

The final stage will be devoted to analysis of the statistical measures applied in the first two stages and development of strategies to proceed with further work and where the classification of queries can be applied to benefit web search tasks. Some possible extensions to this work is expected to be:

- Further studies into reliability of crowd-sourcing for classification, spam prevention etc.
- Assessing whether it is possible to relax MTurk restrictions based on advanced spam prevention measures devised and if this has any effect on results.
- Application of query classification to web search personalisation
- Modification of query classification system to perform classification of any number of categories of any type, utilising the advanced intellect of users to categorise difficult inputs for any automated system.

4 Preliminary Results

Another line of our work\[18\] makes use of a limited number of queries (40) that were already classified in the literature\[3, 8, 9, 15\]. To test the classification method proposed here, this small set of queries were subjected to the MTurk query classification method at the top level only (N-I-T) to confirm the already-known classifications, enabling us to validate whether the classifying procedure agreed sufficiently with accepted classifications elsewhere. Queries were submitted in alphabetical order in order to shuffle the categorical classification of the queries.

<table>
<thead>
<tr>
<th>Navigational</th>
<th>Informational</th>
<th>Transactional</th>
</tr>
</thead>
<tbody>
<tr>
<td>% accuracy</td>
<td>0.74</td>
<td>0.97</td>
</tr>
<tr>
<td>Confidence Interval</td>
<td>95% of 0.5250 to 0.5915</td>
<td></td>
</tr>
<tr>
<td>Overall Kappa</td>
<td>0.5583, SE=0.0170</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Statistical evaluations of initial query set

The calculated Kappa across all categories is lower than expected with exception of the Informational category where 97\% accuracy indicates that the natural language nature and question orientated queries were simple to deduce as being of this type. Navigational queries received lower than expected classification accuracy (74\%) with users finding difficulty in distinguishing from searches of named entities and informational queries. Transactional queries were clearly the most difficult to classify (61\%) with considerable confusion made with the informational category, although work in Stage 2 outlined in Section 3.3 of this proposal will distinguish between the sub-category Informational-Transactional in future classifications.

The lower accuracies for some categories may be attributed to the fact that there are few categories available and it is expected that with future analysis of sub-category classifications that this will improve. Some problematic queries in the transactional category may have also had a significant impact to the Kappa rating. Revision of the instructions provided to users may also see an increase in classification accuracy as better examples may provide users with more knowledge of what is required, but at the same time, information overload may be detrimental. A split system with a concise cheat sheet for quick on screen memory refresh may suit better, and further examples available through hyperlinks, or even a search link.

5 The bigger picture

Experiments proposed in this document form part of a larger plan surrounding personalisation in web search and information retrieval. The placement of this work regarding query categorisation is one of the first steps towards a local personalisation engine that does not heavily depend on external search interfaces and engines, but aims to provide its own with a weighting on query analysis, regeneration and submission. Through local personalisation, the ability to control and filter user variables that are irrelevant can assist in enforcing and maintaining a level privacy as would be expected from query submissions, but is not always the case. We will also be able to compare collections of personalised query submissions and returned results with collections we have submitted over a 2 year period of categorical and common/uncommon queries to determine the difference in result sets and levels of personalisation being applied. Whilst initially observational, it can provide insight into any weaker areas of personalisation and those that can benefit from location personalisation on query submission. This proposal highlights most of the work to be undertaken in the first stage of the overall project, with an overview described below.
Stage One Through query analysis and categorisation the first logical personalisation approach is to determine from the query what type of web search request it is regarding. From this determination, query regeneration can be applied with the removal or addition of parameters that may force personalisation can occur.

Stage Two Implementation of the query management into a localised personalisation engine involves the development of the engine and pre-defined submissions of set queries obtained from previous work to provide initial measurements.

Stage Three User evaluation and determination of satisfaction in addition to concrete performance measures against non-personalised result sets is the final phase of the work.

6 Research Questions
The following research questions will guide the scope of this work.

- How different are personalised search results from ‘normal’ search results?
- Do personalised search results change at the same rate as ‘normal’ search?
- Are personalised results more stable?
- Are different query types more subject to change in personalised results?
- How similar are personalised search results across search engines, compared to ‘normal’ search results?
- Are users more satisfied with the personalised results from one search engine over others? If so why?
- Do users find instability in personalised search results to be frustrating?

The product of this work will also aid in answering the following questions when the results of previous studies by Ben Steichen of Trinity College Dublin are compared with the outputs resulting from this work. These questions are as follows.

- Are the lessons learned in personalisation as used in e-learning equally applicable in search personalisation?
- What is similar, and what is different? Why?

7 Query Categories
The highest level of query categories as used in preliminary work as devised by [3] are as follows.

**Informational** a query indicating the user is seeking to gain information or knowledge of a particular topic

**Navigational** a query indicating the user wants to navigate to a website, either directly, or through specific keywords

**Transactional** a query indicating the user wants to perform a transaction of sorts, which may be the acquisition of resources from, or via the web browser.

Additional sub-categorical classification will be conducted on the categories devised by [9]. The second level entails the following possible classifications.

**Informational (D,U,L,F,A)** Question related subcategories include whether a question is *directed* or *undirected*, a *list* of results is requested, whether the user wishes to *find* an actual product or service or obtain *advice* from a resource.

**Navigational (T,I)** Whether the query is intended to *direct* to a resource that is of a *transactional* nature to perform a transaction or *informational* in nature where the direct purpose is for the discovery of information.

**Transactional (O,D,R,I)** The user may wish to *obtain* a actual resource or object, *download* a virtual resource such as software, obtain a set of *results* to select from or *interact* with a particular resource on another website, such as playing a game.

A third level of possible query classification relates to the types of questions asked in queries, whether a resource should be obtained digitally or physically, and the nature of the transaction.
References


