Information vs Interaction – examining different interaction models over consistent metadata

Kingsley Hughes-Morgan
Department of Computer Science,
Swansea University
Swansea, SA2 8PP, UK
kingsleyhm@googlemail.com

Max L. Wilson
Mixed Reality Lab
University of Nottingham
Nottingham, NG8 1BB, UK
max.wilson@nottingham.ac.uk

ABSTRACT

In the quest to develop better and more useful search systems, many novel search user interface features have been developed, such as relevance feedback, clusters, tag clouds, facets, and so on. Yet all of these novel 'interactions' have required novel forms of 'information', or metadata, to make them work. Consequently, we do not know whether users have been benefiting from better interaction or simply richer forms of metadata, or both. In this research, we aimed to show that better interaction can be provided, regardless of whether we have access to, or the ability to generate, richer forms of metadata. Using only search engine query suggestions as a consistent form of metadata, we built interface conditions for three common interaction models for search: query suggestions (our baseline), hierarchical browsing, and faceted filtering. Our results showed that, despite interacting with the same underlying metadata, users experienced significant performance gains with different forms of interaction. These findings have implications for search user interface designers, who are often working with fixed metadata or within limited budgets. Our future work will focus on complementing these findings by recreating the same interaction with different forms of metadata, such that we can then compare the performance gain separately provided by both information and interaction.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering and Information filtering. H.5.2 [**User Interfaces**]: Interaction styles.

Keywords

Information, Interaction, Representation.

1. INTRODUCTION

Over time, research has provided many exciting new user interface designs to help searchers find the information they need, including: relevance feedback [20], clustering [7], faceted searching and browsing [28], and many more [12, 31]. Yet each of these approaches has required new metadata or new algorithms to produce metadata. Consequently, it is hard to say how they have improved support for searchers: whether they have indeed provided a better way to interact with search results, or whether they simply provide better metadata to the searchers. Arguably, therefore, we do not know if striving to provide better search interfaces, means simply providing better metadata, or providing powerful interaction; or whether they are separable or fundamentally tied.

Although we may want to provide users with rich user experiences with, for example, faceted browsing, designers and developers also may not have access to the rich forms of faceted metadata necessary to do so. In this situation, research into search user interface design begins to fail us – if we are limited to query suggestions, are we limited in the forms of interaction we can provide? Or alternatively – if we spend money and strive to present query suggestions with an exciting interaction, will it provide any benefit to users? In fact, is money better spent on interaction, or on information?

The aim of our on-going research is to fundamentally separate information and interaction, such that they can be controlled independently of each other in order to examine their individual contributions to searchers. In this paper, we report on the first of an on-going series of studies, which has taken a single form of metadata, search engine query suggestions, and created three search user interfaces that allow searchers to interact with them in different ways. The aim of this first study, therefore, is to prove that searchers can benefit from improved interaction, when limited to a certain type of information or metadata.

In the following sections, we first provide some historical context into search user interface design, as well as background on how novel search interface features have been studied. We describe the three interface conditions built for our user study, and how they represent different models of interaction. We then present the structure of our study before describing the results and discussing implications for the future.

IIiX'12, Nijmegen, The Netherlands.
Copyright 2012 ACM 978-1-4503-1282-0/2012/08 … $15.00.

2. RELATED WORK

Over the last few decades, there have been many studies of Interactive Information Retrieval (IIR) and Search User Interfaces, which have recently been catalogued in larger surveys and books. Hearst provided an extensive text on Search User Interfaces [12], Morville and Callender provided an industry perspective on Search Patterns [23], while White and Roth focused on designing interfaces to support exploratory forms of search [29]. More recently, Wilson provided both a brief overview of early search user interfaces, and a framework for understanding search user interface features [31]. Wilson's framework suggests that different interface innovations can be considered as: 1) Input, 2) Control, 3) Informational, or 4) Personalisable, noting that strong innovations make contributions in three or four areas. A keyword search box is a powerful interface feature and a good example, which is primarily used for query *input*, but queries can be modified to refine (*control*) a search. Typically, keyword search boxes continue to display the current search, thus making them *informational* too. If a keyword search box has an auto-suggest feature, then previously used queries can be suggested too (*personalisable*). The primary focus of this paper, and much prior IIR work, is on *control* interactions, which help refine queries to get better results.

2.1 Examples of IIR Control features

Many IIR developments have been suggested and studied over time. One early proposal for improving IIR was Relevance Feedback [26]. By asking users to select relevant results from a results list, additional example query terms could be automatically extracted, and applied to return more relevant results. Harman showed that the average relevance of results could be increased when users chose the best additional terms to use [11]. In line with the theme of this paper, Koenemann and Belkin showed specifically that searchers performed better at search tasks, when given control over terms used from relevance feedback [20]. Despite these advances, relevance feedback has not become popular in search user interfaces. Studies indicate that providing relevance feedback (identifying relevant results) is too much of a distraction mid-task, creating unnecessary cognitive load [1, 2], and taking too much time from the main search task [3].

Below, three particular case studies relevant to this paper are described: Interactive Query Expansion (particularly query suggestions), Hierarchical Clustering, and Faceted Filtering. Each of these approaches have provided significant support for searchers and improved search user interfaces, but all required specific algorithms or more advanced metadata. Consequently, although we know these approaches can be successful, we do not know whether the benefits come from their information or interaction.

Interactive Query Expansion (IQE), in the way most search engines or search user interfaces now present them as Query Suggestions, followed on from the non-uptake of relevance feedback. Typically, query suggestions are extended versions of the submitted query, taking either terms from the most relevant results, or popular similar queries from other users. Kelly et al suggested that participants preferred query suggestions, rather than simply suggesting additional *terms*, and that searchers found more results, despite not actually improving in task performance [19]. Kelly further compared automatically generated and user-generated suggestions, finding that user-generated were better for query, rather than term, suggestions. Ruthven discovered, however, that human searchers were less likely to pick good query suggestions than automatic systems [25]. Further, in studying different representations of query suggestions, Diriye et al discovered that query suggestion features in search user interfaces can actually hinder or slow searchers during simpler tasks, while supporting participants in more exploratory tasks [8]. Ultimately, although better than alternatives like term suggestions, research indicates that query suggestions have limitations in their support for searchers.

Hierarchical Clustering is an approach that clusters results into smaller groups that refer to different topics. Early work on an approach called Scatter/Gather [7] split results up into groups, labelling them for their unique concepts. Despite previous methods suffering from speed problems, Scatter/Gather ran in linear time and subsequent work used pre-processing to reduce the approach to constant time [6]. Empirical work by Hearst and Pedersen further showed that showing results in groups according to their cluster improved retrieval performance in tasks [16]. Further, Pirolli et al. showed that Scatter/Gather helped searchers to understand the structure of the collection they were browsing [24]. More recently, Hierarchical Clustering [4] techniques are typically provided as a tree structure (e.g. clusty.com), where users can interactively browse the results of the query. As opposed to query or term suggestions (or interactive query expansion in general), hierarchical clustering features typically do not re-issue new queries, but simply display portions of the result set and leave the original query intact.

Overall, Hierarchical Clustering techniques typically rely on pre-processing result sets with specific algorithms, in order to produce new richer metadata that can be used to explore the result set. Thus, in studies, the benefits of hierarchical clustering could be explained by both the interaction (e.g. Hearst and Pedersen [16]), or the presentation of better metadata (e.g. Pirolli et al [24]). Much more could be said about approaches to hierarchical clustering, but it is beyond the scope of this paper to do so. Readers can refer to Carpineto et al for a more detailed literature survey [4].

Faceted Filtering is another IIR approach that utilizes rich metadata to provide a new way to filter and explore data. Faceted metadata is made up of several orthogonal dimensions in the metadata that each relate to the dataset of results [13]. Searchers may then use these different dimensions, such as price, size and colour, to reduce the information space during search [22]. The most familiar experience with faceted browsing online was embodied by Hearst's original work on the Flamenco Browser [34]. Hearst's later work noted that a carefully crafted coherent set of faceted metadata (information) was more important than interaction [14], where automatic processing (like hierarchical clustering) can produce overly large or confusing facets. Faceted metadata is thus extremely rich, where Wilson and schraefel, like Pirolli et al with hierarchical clustering, suggested that the structure of facets can help to communicate information and previous browsing decisions [32, 33]. Despite attempts to provide faceted filters over large heterogeneous corpora like the web [21], faceted browsing is typically provided over more homogeneous corpora like online retail collections or digital libraries. Services like Amazon and eBay, for example, require searchers to reduce results to a relatively homogeneous collection before providing faceted filters.

Faceted browsing can be used as an interactive technique to submit more refined queries like query suggestions, or to reduce result sets for queries like hierarchical clustering (often called faceted search or faceted filtering) [28]. Ultimately, the important feature of faceted interaction is that searchers can apply combinations of orthogonal metadata, in any combination, to reduce the number of results that they can see. Like Hierarchical metadata, however, the benefits of faceted interaction can be mixed between the novel interaction and the rich metadata, while requiring specific high-performance algorithms. Clarkson et al compared a range of approaches to designing the interaction with faceted browsing [5], while Tunkelang provides a large review of faceted search literature [28].

2.2 Interaction coupled with metadata

Overall, each of the IIR *Control* features discussed above provide a new interaction, whilst providing different forms of metadata, and requiring special algorithms to work. Of the approaches, faceted filtering is both powerful and increasingly common online. Query suggestions are perhaps the most frequently provided IIR feature, but the simplest of the interaction models. Consequently, those building search systems might wish to provide rich faceted filtering experiences, but be limited in the technology and metadata available; or the budget to acquire better algorithms or produce better metadata. In this paper we seek to ask: given a specific form of metadata, can we recreate more advanced IIR interface features such that searchers can still experience their benefits?

We do know that search user interface design changes alone can provide significantly better support for searchers. In a very simply form, Drori and Alon showed that searchers performed tasks faster, when results were labelled with the category that they are related to [9]. Dziadosz and Chandrasekar [10] showed that the addition of result thumbnails helped users to identify more relevant results. Subsequent work by Teevan et al, however, showed that, while thumbnails were good at supporting re-finding, abstracted images built from logos, titles, and colour-schemes were equally good at supporting searching and re-finding [27]. Early efforts like Hearst's TileBars [15] augmented results with a visualization of how segments of each result were relevant to submitted query terms. Although we don't see TileBars frequently in practice, White et al [30] showed that search result snippets (the text below search results) were more effective if they showed sentences that specifically included the query terms.

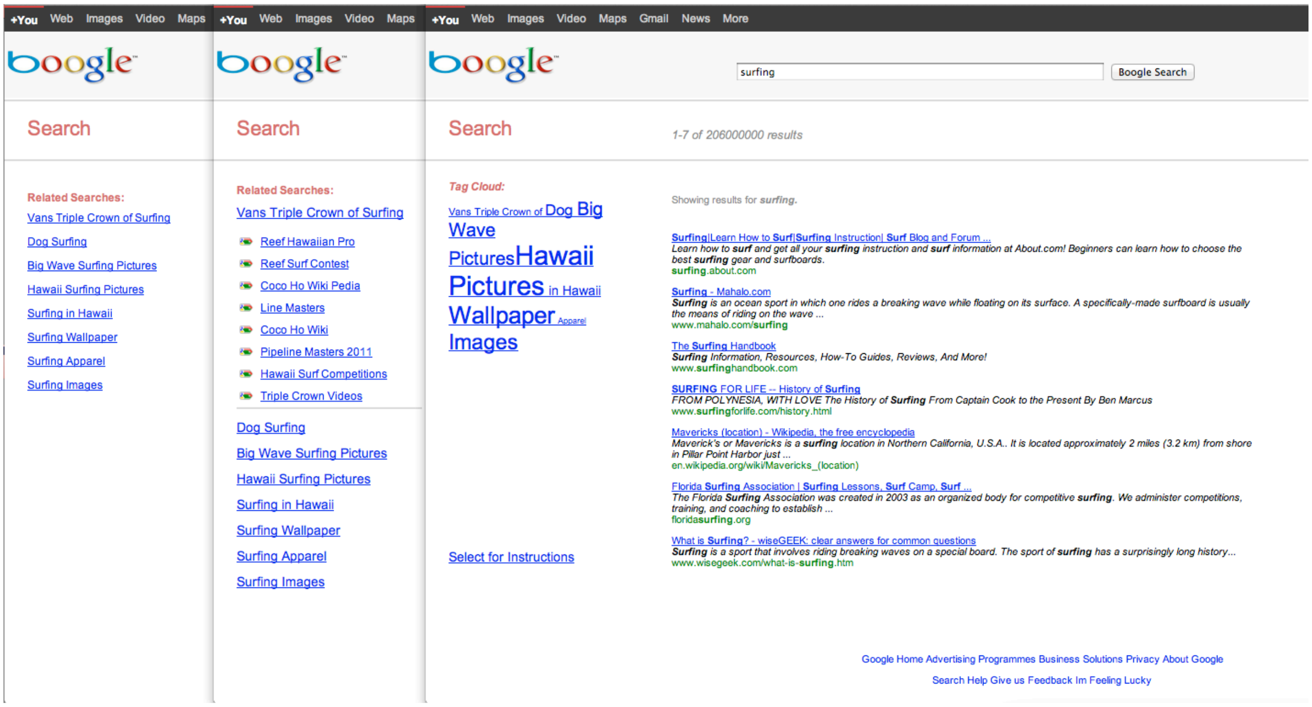
More inline with the aims of this paper, some studies have compared representing the same metadata in different forms. Joho et al, for example, compared presenting term suggestions as a linear list and as a hierarchy [17]. Their results favoured hierarchical, but the two approaches used separate algorithms, and the interaction was in cascading menus rather than in the form that we typically experience query suggestions. Further, as mentioned above, Kelly compared query and term suggestions separately, favouring query suggestions, but both of these approaches provided the same form of interaction: re-issuing queries [19]. Below we describe a comparative study of different interaction models, using a single form of metadata, produced by the same algorithm each time. To our knowledge, no prior work exists that has specifically compared different forms of interaction, while controlling and removing metadata as a potentially confounding variable.

3. RESEARCH DESIGN

Our on-going research is motivated by understanding whether better support for searchers, or perhaps how much support, comes from a) better interaction, or b) better information and metadata. Consequently, our research questions include:

- RQ1: Can we support searchers better with different interaction if we are limited by available metadata?
- RQ2: How much performance gain can searchers get by changing interaction alone?
- RQ3: How important is improved metadata for providing better search?
- RQ4: How much should companies invest in metadata versus search user interface design to best support their searchers?

This paper describes a study that is focused on RQ1 and RQ2, where our future work will work towards answering RQ3 and RQ4. This particular study focuses on the following hypothesis:



Surfing|Learn How to Surf|Surfing Instruction| Surf Blog and Forum ...
Learn how to surf and get all your surfing instruction and surf information at About.com! Beginners can learn how to choose the best surfing gear and surfboards.
surfing.about.com

Surfing - Mahalo.com
Surfing is an ocean sport in which one rides a breaking wave while floating on its surface. A specifically-made surfboard is usually the means of riding on the wave ...
www.mahalo.com/surfing

The Surfing Handbook
Surfing Information, Resources, How-To Guides, Reviews, And More!
www.surfinghandbook.com

SURFING FOR LIFE -- History of Surfing
FROM POLYNESIA, WITH LOVE The History of Surfing From Captain Cook to the Present By Ben Marcus
www.surfingforlife.com/history.html

Mavericks (location) - Wikipedia, the free encyclopedia
Maverick's or Mavericks is a surfing location in Northern California, U.S.A.. It is located approximately 2 miles (3.2 km) from shore in Pillar Point Harbor just ...
en.wikipedia.org/wiki/Mavericks_(location)

Florida Surfing Association | Surfing Lessons, Surf Camp, Surf ...
The Florida Surfing Association was created in 2003 as an organized body for competitive surfing. We administer competitions, training, and coaching to establish ...
floridasurfing.org

What is Surfing? - wiseGEEK: clear answers for common questions
Surfing is a sport that involves riding breaking waves on a special board. The sport of surfing has a surprisingly long history...
www.wisegeek.com/what-is-surfing.htm

Figure 1: The three interaction conditions in the study. UIQ on the left presents query suggestions in their common form. UIC in the middle presents secondary query suggestions with an interaction model based on hierarchical clustering. UIF on the right, which includes the whole view of the Google UI recreation, provides terms, or facets, that can be applied to or removed from the search in any combination to 'filter' the results.

- H1: Searchers will be more efficient with more powerful interaction, using the same metadata, when completing search tasks.
- H2: Searchers will enjoy more powerful interaction, despite using the same metadata.
- H3: Searchers will use query recommendations more when they are presented differently.

In order to accept or reject these hypotheses, we designed a 3x2 repeated-measures study using two independent variables: 1) form of interaction, and 2) type of task. There were 3 forms of interaction, described below, covering standard query suggestions, hierarchical clustering, and faceted filtering. There were 2 types of task: simple and exploratory. Below, we describe these factors in more detail, beginning with the three search interfaces.

3.1 Form of Interaction

For this study, we built three search user interfaces (Figure 1) that closely resembled Google, but using the freely available Bing Search API[1]. Bing's API was chosen because it was a) free to use, b) easy to process on the server-side[2], and c) less limited in terms of number of API calls than the alternatives. We chose for all three user interfaces to resemble Google, as it was most likely to be familiar to the majority of study participants.

The three user interfaces varied only in the form of IIR interaction to the left of the results, described in turn further below. Otherwise, all three interfaces allowed searchers to search the web as normal, submitting queries and clicking on results. The alternative forms of search, such as image, maps, and YouTube, were disabled. Elements like spelling corrections were also implemented, as was the inclusion of information like number of results and time taken. Finally, however, based upon the results of a pilot study, a design decision was taken to remove paging in order to encourage the use of query refinements. Although paging was relatively infrequent, removing this feature did encourage additional use of refinements, without noticeably affecting user opinion of the design. In fact, some pilot participants did not even notice that paging was missing. This was the only design diversion away from the typical Google search experience.

UIQ: Query Suggestions. Using query suggestions in their most natural form of interaction, UIQ presented query suggestions from the Bing API as a list down the left hand side of the results page. As per the standard interaction provided by search engines, selecting a query suggestion simply issued an entirely new query, presenting new results and a new set of query suggestions to go with them. In

[1] http://www.bing.com/toolbox/bingdeveloper/

[2] Google's API uses javascript, which means that the data manipulation is restricted to client-side processing.

Google, these are typically found at the end of the search results, and in Bing they are typically found to the left of the search results. Ultimately, however, UIQ was our baseline condition and simulated the typical behaviour of query suggestions.

UIC: Hierarchical Clustering. Our second user interface provided a browsing experience similar to hierarchical clustering interfaces like Clusty.com. In terms of interaction, clustering interfaces use hierarchical clustering techniques to automatically generate a tree-like structure of entities and sub-entities that can be found in the results set. The searcher can then filter all the search results retrieved by the system by either top-level or sub-level entities in the hierarchy. When selecting an entity in the hierarchy, the results are filtered, and any sub-entities are shown in the hierarchy. At all times, the searchers original query is left in the search box, indicating that the results have been filtered rather than the system submitting a new query.

To recreate the hierarchical clustering experience, standard query suggestions were retrieved from the Bing API for the current query. These were used as the top-level entities in the hierarchy. For each query suggestion, UIC then asked for subsequent query suggestions, which were represented as the sub-level entities in the hierarchy. To create the same sensation of simply filtering and browsing through the results, as opposed to reissuing queries, the searchers original query was left in the search box. As well as highlighting the item that had been selected in the hierarchy, UIC also used Google's standard terminology to say 'Showing results for [selected item in hierarchy]'. Consequently, although the system was technically issuing more specific queries underneath, the experience appeared to participants as choosing to display different sub-clusters of the initial results returned by the query.

UIF: Faceted filtering. In faceted filtering systems, searchers can take any of the items of metadata made available to them, and apply them in combination in order to filter the results. Thus the user is able to flexibly combine, add, or remove any number of keyword filters in order to describe what they are looking for and narrow their results. Like with clustering, systems typically maintain any search query as a constant in the search box, and then apply the selected keywords to filter the results to portions of the overall result set.

Once again, for UIF, we restricted ourselves to using just the Bing API query suggestions, but aimed to create a search feature that allowed searchers to apply multiple suggested terms in combination. Without carefully constructed metadata we were unable to create a set of distinct facets, such as sets of prices, colours, brands, and so on, which are commonly seen in online retail stores. Instead, we extracted terms from the query suggestions to display separately as additional query terms that could be applied in any combination to the query. Consequently, we chose an output that appeared much like a tag cloud, such that it would appear in a form that would be familiar for many users. The tag cloud was displayed in a common style, with popular terms displayed in a larger font. Overall, however, the tag cloud provided the same interaction model as items in facets: users were able to 'turn on' and 'turn off' any term in the list as a filter, where 'on' terms were highlighted using background. This is a different interaction model to providing term suggestions, which would issue a refined query, provide new results, and new term suggestions, similar to our baseline condition. Like UIC, however, the faceted filters remain constant until the user changes their query, which was left in the search box. The initial query and filters were displayed together using Google's phrasing as: 'Showing results for [query + selected terms]'. Again, this combination made the experience appear as if searchers were applying filters to the results returned by the original query, but in reality, the system was still issuing refined queries to the Bing API.

3.2 Type of task

Two standard types of user study task were used in the study: 1) a simple lookup task and 2) an exploratory task. All six tasks are shown in Table 1.

The simple lookup tasks had a fixed answer, but the chosen task description was presented in such a way that the most likely query would not find the answer without subsequent queries or refinements. This approach was chosen to intrinsically encourage participants to use the IIR features on the left of each user interface condition.

Table 1: Tasks set to participants in the study. S = Simple, E = Exploratory

ID	S/E	Task Description
1	S	What is the population of Ohio?
2	E	Find an appropriate review of "Harry Potter and the Deathly Hallows". - Compare the rating with the previous film.
3	S	Find the first state of America.
4	E	Deduce the main problems that Steve Jobs incurred with regards to his health.
5	S	What is the iPad 3's proposed processor name?
6	E	Explore information related to Apple's next iPhone, the iPhone 5. - Note the expected release date. There could well be multiple rumours.

The exploratory search tasks were chosen to be tasks with multiple sub-problems, such that searchers would have to perform a series of searches or refinements to combine answers from several websites. The tasks, therefore, resembled a collection-style task, without there being specific dependencies between the sub-elements. There was also no fixed answer to these tasks, where users could choose answers subjectively.

3.3 Procedure

Participants were first provided with their legal rights and with sufficient detail to give informed consent. Participants then filled out a demographic questionnaire before beginning to perform the tasks with each user interface. Participants were provided with an initial launch page, which associated their details with a unique ID for the study and provided links to the three conditions, where UI (link) ordering was counter-balanced in a latin square rotation. Before beginning the two tasks with a given condition, the UI was introduced to the participants with a quick demo of the interaction available. Participants then performed one simple and one exploratory task. Ten minutes was allocated per UI to complete the two tasks, although most did not need all this time. After completing both tasks with a UI, participants filled in a quick survey gathering some immediate feedback on the experience. After completing tasks in all three conditions, participation concluded with a final short survey and a debriefing discussion of the different interactions, usually lasting a further 10 minutes. Participants were given a £10 Amazon voucher in appreciation of their time and contribution.

3.4 Measures

All tasks were timed, and all queries, refinements and page-views were logged into a series of back-end databases. Each initial search along with subsequent query refinements were recorded, including the pages that were visited from the results, hence obtaining the answers to the tasks. In the post-condition surveys, likert-scale questions were asked including rating elements such as: ‘the ease of use’ and ‘satisfaction after task’. The closing survey questionnaire provided subjective retrospective insight into ‘which UI provided the quickest correct answer?’, ‘which UI provided the most enjoyment of searching?’ and ‘which UI design was most appealing?’.

3.5 Participants

18 participants were recruited from across a British university, from a range of academic, student, and technical roles. Participants aged between 16-55 (mean: 28), with mixed educational backgrounds; 5 were undergraduates, 9 were educated up to college level, and 4 educated to either a master’s or doctoral level. 8 participants were male, and 10 female. All participants indicated that they searched the web at least daily, with the vast majority using the web for work and social media. 16 participants recalled that they had used some form of query refinement whilst searching.

4. RESULTS

In the study, participants were given two types of tasks: ‘simple’ and ‘exploratory’. Overall, there was no significant difference in the amount of time spent on these two types of tasks (mean S=176s, E=179s). These similar times were most likely observed because the three simple tasks had a hard to find answer, in order to encourage use of the different refinement interactions in the study. However, participants submitted significantly more queries (mean S=1.75 queries per task, E=2.33, t(53)=2.3751, p<0.05) and visited significantly more links (mean S=1.65 pages visited per task, E=2.09, t(53)=3.456, p<0.005) in the exploratory tasks. Participants did not use significantly more refinements (mean S=2.42 refinements per task, E=2.45) in exploratory tasks. We conclude that both types of tasks encouraged participants to use the three different interactions for query refinements, but the multi-part exploratory tasks required participants to search for more sub-topics and thus visit more pages. Below we discuss the individual differences between the three interface conditions, first with the logged interactions and second by subjective feedback.

4.1 Queries, Refinements, and Page Visits

The results below analyse the three interface conditions for both the simple and exploratory tasks separately, where we see more significant differences in the exploratory tasks (Table 3) than for simple tasks (Table 2). For each measurement in these two tables, an ANOVA was performed.

4.1.1 Simple tasks with a single answer

For simple tasks, there was a significant difference in the number of queries submitted (F(51,2)=4.899, p<0.05). A post-hoc TukeyHSD revealed that participants submitted significantly more queries in the UIC (p<0.05) and UIF (p=0.05) conditions, but showed no difference between UIC and UIF. Conversely, however, we saw comparable use of refinements and similar numbers of page visits between the three interface conditions. These results indicate that for simple tasks, or tasks with a fixed answer, the different interaction models did not create a significant effect on *refinement* behaviour. Participants did, however, perform significantly faster in the hierarchical clustering UIC condition (p<0.005, F(51,2)=6.53), where a post-hoc TukeyHSD showed that UIF and UIQ were not significantly different from each other.

Table 2: log data for simple tasks (*=significant)

Mean (std)	**UIQ**	**UIC**	**UIF**
Queries (#) *	1.22 (0.55)	2.11 (1.13)	1.94 (0.93)
Refinements (#)	2.44 (0.70)	2.5 (1.95)	2.33 (1.19)
Page visits (#)	1.94 (1.11)	1.61 (0.69)	1.39 (0.61)
Time (s) *	189 (3.15)	154 (2.57)	184 (3.07)

It is not clear exactly why participants submitted significantly more queries in the UIC and UIF conditions, but the results indicate that participants were fastest when interacting with a hierarchy. Although we didn’t reach statistical significance, there was a downward trend to visiting fewer links (to get the right answer) when using the faceted approach (the p value was ~0.1). Informally, therefore, we might consider that the same number of refinements, but applied in combination in UIF, led to pages with the right answer in fewer page views. Next, we consider exploratory tasks.

4.1.2 Exploratory tasks with multiple answers

For exploratory tasks, we saw significant differences in all four measures across the three interface conditions. Participants submitted a significantly different number of queries between the three conditions (F(51,2)=9.142, p<0.0005). A post-hoc TukeyHSD revealed that participants submitted significantly fewer queries in the UIC condition compared to UIF (p<0.05) and UIQ (p<0.0005), but that UIF and UIQ were not significantly different. Conversely, participants used significantly more refinements in the faceted UIF condition (F(51,2)=6.245, p<0.005). Again, a post-hoc TukeyHSD revealed significant differences between UIF and the two alternatives (both p<0.05), but no difference between UIC and UIQ. Together, these two sets of results indicate that participants behaved very differently in the three conditions, for exploratory tasks, using significantly fewer queries (with UIC) and significantly more refinements.

Table 3: log data for exploratory tasks (*=significant)

Mean (std)	UIQ	UIC	UIF
Queries (#) *	3.11 (1.49)	1.44 (0.51)	2.44 (1.29)
Refinements (#) *	2.17 (0.86)	1.78 (0.65)	3.39 (2.23)
Page visits (#) *	2.55 (1.04)	1.61 (0.69)	2.11 (0.75)
Time on task (s) *	190 (3.17)	169 (2.82)	177 (2.95)

In exploratory tasks, participants visited significantly more pages in the original condition (F(51,2)=5.615, p<0.01), where a post-hoc TukeyHSD saw only one key difference between UIQ and UIC. This finding may indicate that participants were able to find more relevant pages earlier in the task, thus needing to visit fewer pages. Time on task was also significantly different across conditions (F(51,2) = 5.501, p<0.01), with UIC being significantly faster than UIF (p<0.05) and UIF being significantly faster than UIQ (p<0.05). These results indicate that both experimental conditions provided significant performance gains for participants, with hierarchical clustering also helping them to be faster.

4.2 Subjective Responses

After performing both tasks with a UI condition, participants provided Likert-scale ratings (1-5, where 5 is best) for ease of use and satisfaction after task. Table 4 provides the median and mean scores for these subjective results, where a significant difference was seen in both measures using a Friedman test.

For ease of use, participants rated UIF as being significantly harder to use than the original suggestion UIQ condition (p<0.05) and the clustering UIC condition (p<0.005). Participants did not rate UIC as being significantly harder to use than the baseline UIQ condition.

Table 4: Likert responses immediately after taking part in each condition; 1-5, 5 is best. *=significant.

Median (mean)	UIQ	UIC	UIF
Ease of use *	4 (3.84)	4 (4.11)	3 (3.06)
Satisfaction *	4 (3.50)	4.5 (4.00)	3 (2.77)

For task satisfaction, participants rated UIC significantly higher than UIF (p<0.005 in a post-hoc analysis) and almost marginally higher than the baseline UIQ condition (~p=0.1). Further, participants were significantly more satisfied in UIQ than the faceted UIF condition (p<0.05).

The choices highlighted in Table 5 indicate that participants enjoyed and preferred the UIC clustering condition most, despite believing they performed fastest with the original baseline condition. This last point disagrees with actual timing data, indicating that, for participants, the familiar baseline felt faster. The faceted design was rarely chosen favourably in any of the subjective measures.

Table 5: Frequency of choice when reflecting on all three conditions at the end of the study.

Frequency of choice	UIQ	UIC	UIF
Quickest to correct answer	**11**	5	2
Most enjoyment during task	4	**11**	3
Most appealing design	5	**11**	2

5. DISCUSSION

Our study has provided evidence for both of our initial research questions (RQ1 and RQ2): that we can better support searchers by changing the interaction, when we have a fixed form of metadata. In simple tasks with a fixed but hard to find answer, participants submitted significantly more queries with the two modified interactions (UIC and UIF), found correct answers in fewer page visits with UIF, and were significantly faster with UIC. In exploratory tasks, participants submitted significantly fewer queries, needed significantly fewer refinements, and viewed fewer pages to find suitable answers with the UIC condition. Further, in exploratory tasks, participants used significantly more refinements in the faceted UIF condition. Finally, participants performed significantly faster in exploratory tasks with both experimental conditions, especially with the UIC condition.

Below we discuss how the data supported our hypotheses, how the findings relate to prior work in this area, and then our plans for further pursuing our research questions.

5.1 Our Hypotheses

In regards to H1, we expected participants would be more efficient with more powerful interaction. We are able to accept this hypothesis, in that participants were significantly faster in simple tasks, and performed more efficiently in exploratory tasks in all four measures. UIF, although more powerful than UIQ, was rarely significantly different from the baseline UIQ. This implies that simply being 'more powerful' is not sufficiently definitive, but that performance is affected by more factors, discussed below.

In regards to H2, we saw significant differences in subjective ratings for both ease of use and satisfaction after task. Notably, participants did not consider the hierarchical representation of query suggestions (UIC) to be significantly harder to use than the baseline. In fact, informally, UIC's mean (as a secondary indicator to the median) was a little higher than the baseline UIQ condition. Conversely, however, participants found UIF significantly harder to use. Further, participants were less satisfied in completing their task with UIF. Both of these indicate that participants did not find the faceted condition particularly intuitive. Together, combined with the preference data towards UIC, these results allow us to conditionally accept H2. The caveat to accepting H2 is that despite both UIC and UIF being more powerful, and both improving retrieval performance, only one significantly improved enjoyment and satisfaction.

Discussing the three conditions with participants provided some further insights into the results. Notably, some participants described the UIC condition as quite exciting compared to the baseline that they found familiar and mundane. Higher preference scores could have been the result of a novel experience, which should be investigated further in the future to remove it as a confounding variable. Additionally, some participants reported finding more clues to answers embedded in the query suggestion hierarchy, providing some evidence for Pirolli et al's hypothesis that hierarchical clusters help to communicate structure [24]. UIF, however, tended to split participants, with some using its capabilities quite effectively. Other participants reported UIF to be confusing, and many, including those who used it effectively, disliked UIF's visual design. Another possible conclusion about UIF is that interaction *is* tied to type of metadata or information, thus potentially contradicting our hypotheses. UIF, and its design, will be the focus of future investigation, discussed further below.

Finally, we were also able to partially accept H3, where participants used significantly different amounts of refinements in the three conditions, but only in exploratory-style tasks.

5.2 Relation to prior research findings

Some of the results above do match closely with the results of other studies. Kelly et al's findings indicated that query suggestions were better than term suggestions [19]. Of our conditions, UIF provided *terms* that could be applied in combination and was typically less effective than the other two conditions, which provided *query* suggestions. Similarly, although Joho et al used terms in their study, their results indicated that hierarchical representations were better than linear lists [17].

Like some prior studies (e.g. [17, 18]), results indicated that participants behaved very differently in different conditions. In exploratory tasks, for example, UIF led to fewer queries and significantly more refinements, while viewing fewer in less time than the baseline. Conversely, participants submitted significantly fewer queries, used less refinements, viewed significantly fewer links and in significantly less time than the baseline (Table 3). Unlike these prior studies, however, we did also see retrieval performance increase, with significantly different task times for both simple and exploratory tasks.

5.3 Limitations and Future Work

There are several areas where our study design and findings could be made more insightful, as well as helping us to answer all of our research questions. First, it was beyond the scope of this study to consider auto-suggestions that are often shown to searchers as they type. These dynamic suggestions are a common in search engines, but are typically designed to pre-empt the need to interactively filter results. This study focused on those interactively refining results after a query has been made. Also, although we focused on replicating a familiar web search experience, the study described above could be strengthened if performed over a dataset that includes relevance judgments. Using a TREC collection would allow precision and recall, and more advanced TREC-style measures, to be used. Such measures would allow us to examine, for example, whether these interaction styles helped users to view more relevant pages; or put another way, view fewer irrelevant pages. Our study asked participants to continue until they found an acceptable answer, which allowed us to use number of page visits as a measure. We were not able to tell how many of those, however, were relevant and helpful, but could with a controlled corpus.

Further, although this study has allowed us to show that changing the interaction style with a given metadata produced different searching behaviours, we would not be able to easily do the opposite: show that changing the metadata under the interaction has an effect. We leveraged a persistent style of metadata produced by the Bing API, but would struggle to produce powerful faceted metadata for the unbounded web. Again, developing faceted metadata for a TREC collection, would allow us to study the metadata associated with each of the interactions compared in this study. A series of larger studies would let us compare the three interactions against each of the three forms of metadata. The results of such a series of studies would then tell us separately the impact of interaction and metadata, and help us to answer the remaining research questions (RQ3 and RQ4). This study has provided evidence that embarking on this planned future work will provide valuable results.

Finally, although our faceted condition (UIF) recreated the interaction model of faceted filtering – being able to apply aspects from the query suggestions in any combination – we may be able to refine or improve this approach in the future. Visually, UIF appeared a bit like a tag cloud, but the interaction was different from what participants might expect from a tag cloud. An improved version of UIF might take both 1st and 2nd level suggestions (like UIC) and make them visually more similar to faceted browsers. We did not do this originally, as initial prototypes appeared very similar to UIC. When performing our future studies over a fixed collection, we will *have* to design a faceted interaction that is similar to those used in digital libraries or online retail sites, in order to provide a consistent faceted filtering experience over all three forms of metadata.

6. CONCLUSIONS

In this paper, we described an empirical user study that compared the interaction models of three notable advances in IIR user interfaces: query suggestions (as a baseline), hierarchical clustering, and faceted filtering. Where these IIR features have been studied many times in the past, their benefits are typically conflated because they provide both a) a new interaction, and b) newer, richer metadata. In this paper, we separated these two factors to specifically investigate interaction, while keeping the metadata constant. To do so, Bing's API query suggestions were used as a common form of metadata, and separate interface conditions were built to provide the three different interaction models.

Our results showed significant differences in searching behaviour and significant performance differences in all three interface conditions, for both exploratory multi-part tasks and simple but hard-to-find single-answer tasks. Of the three conditions, participants preferred, and often performed better using the UIC model of interaction. Against expectations, participants did not experience as many performance gains with the faceted model of interaction, providing mixed responses about its design. We conclude that, designed effectively, searchers can experience significant performance gains simply by improving the interaction over a given form of metadata.

For search system designers, who may often be limited by available metadata, systems and algorithms, or simply by budget, this paper has contributed notable findings. Our results conclude that it is worth investing time in developing the interaction model, even if the underlying system or data is fixed. Our future work will focus on performing similar studies that maintain an interaction model and use richer forms of metadata, perhaps providing support for Hearst's hypothesis about carefully constructed metadata. Ultimately, however, by rotating both the interaction and the information across these three search user interface features, our results will be able to plot the specific advantages brought separately by improving both interaction and information. Such findings would provide guidance to people developing new search systems, by beginning to quantify the benefits of investing time, money, and other resources on both information, and interaction.

7. REFERENCES

[1] Back, J. and Oppenheim, C., A model of cognitive load for IR: implications for user relevance feedback interaction. *Information Research*, *6*(2), 2001.

[2] Beaulieu, M. and Jones, S., Interactive searching and interface issues in the Okapi best match probabilistic retrieval system. *Interacting with Computers*, *10*(3), 237-248. 1998.

[3] Belkin, N.J., Cool, C., Kelly, D., Lin, S.J., Park, S.Y., Perez-Carballo, J. and Sikora, C., Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing & Management*, *37*(3), 403-434. 2001.

[4] Carpineto, C., Osiński, S., Romano, G. and Weiss, D., A survey of Web clustering engines. *ACM Computer Surveys*, *41*(3), 1-38. 2009.

[5] Clarkson, E.C., Navathe, S.B. and Foley, J.D., Generalized formal models for faceted user interfaces. In *Proc. JCDL 2009*, 125-134. 2009

[6] Cutting, D.R., Karger, D.R. and Pedersen, J.O., Constant interaction-time scatter/gather browsing of very large document collections. In *Proc. SIGIR 1993*, 126-134. 1993

[7] Cutting, D.R., Karger, D.R., Pedersen, J.O. and Tukey, J.W., Scatter/Gather: a cluster-based approach to browsing large document collections. In *Proc. SIGIR 1992*, 318-329. 1992

[8] Diriye, A., Blandford, A. and Tombros, A., A polyrepresentational approach to interactive query expansion. In *Proc. JCDL 2009*, 217-220. 2009

[9] Drori, O. and Alon, N., Using documents classification for displaying search results list. *Journal of Information Science*, *29*(2), 97-106. 2003.

[10] Dziadosz, S. and Chandrasekar, R., Do thumbnail previews help users make better relevance decisions about web search results? In *SIGIR'02*, 365-366. 2002

[11] Harman, D., Towards interactive query expansion. In *Proc. SIGIR 1988*, 321-331. 1988

[12] Hearst, M. Search User Interfaces. Cambridge University Press, 2009.

[13] Hearst, M., Elliot, A., English, J., Sinha, R., Swearingen, K. and Yee, P., Finding the flow in web site search. *Communications of the ACM*, *45*(9), 42-49. 2002.

[14] Hearst, M.A., Clustering versus faceted categories for information exploration. *Communications of the ACM*, *49*(4), 59-61. 2006.

[15] Hearst, M.A., TileBars: visualization of term distribution information in full text information access. In *Proc. CHI 1995*, 59-66. 1995

[16] Hearst, M.A. and Pedersen, J.O., Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *Proc. CHI 1996*, 76-84. 1996

[17] Joho, H., Coverson, C., Sanderson, M. and Beaulieu, M., Hierarchical presentation of expansion terms. In

Proceedings of the 2002 ACM symposium on Applied computing, ACM, Madrid, Spain, 645-649. 2002.

[18] Kelly, D. and Fu, X., Elicitation of term relevance feedback: an investigation of term source and context. In *Proc. SIGIR 2006*, 453-460. 2006

[19] Kelly, D., Gyllstrom, K. and Bailey, E.W., A comparison of query and term suggestion features for interactive searching. In *Proc. SIGIR 2009*, 371-378. 2009

[20] Koenemann, J. and Belkin, N.J., A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *Proc. CHI 1996*, 205-212. 1996

[21] Kules, B., Kustanowitz, J. and Shneiderman, B., Categorizing web search results into meaningful and stable categories using Fast-Feature techniques. In *Proc. JCDL 2006*, 210-219. 2006

[22] McGuffin, M.J. and schraefel, m.c., A comparison of hyperstructures: zzstructures, mSpaces, and polyarchies. In *Proc. Hypertext 2004*, 153-162. 2004

[23] Morville, P. and Callender, J. Search patterns: Design for Discovery. O'Reilly Media, Inc., 2010.

[24] Pirolli, P., Schank, P., Hearst, M. and Diehl, C., Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proc. CHI 1996*, 213-220. 1996

[25] Ruthven, I., Re-examining the potential effectiveness of interactive query expansion. In *Proc. SIGIR 2003*, 213-220. 2003

[26] Salton, G. and Buckley, C., Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, *41*(4), 288-297. 1990.

[27] Teevan, J., Cutrell, E., Fisher, D., Drucker, S.M., Ramos, G., Andre, P. and Hu, C., Visual snippets: summarizing web pages for search and revisitation. In *Proc. CHI 2009*, 2023-2032. 2009

[28] Tunkelang, D., Faceted Search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *1*(1), 1-80. 2009.

[29] White, R. and Roth, R. Exploratory Search: Beyond the Query-Response Paradigm. Morgan & Claypool, 2009.

[30] White, R.W., Ruthven, I. and Jose, J.M., Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In *Proc. SIGIR'02*, 57-64. 2002

[31] Wilson, M.L., Search User Interface Design. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *3*(3), 1-143. 2011.

[32] Wilson, M.L., André, P. and schraefel, m.c., Backward Highlighting: Enhancing Faceted Search. In *Proc. UIST 2008*, 235-238. 2008

[33] Wilson, M.L. and schraefel, m.c., The Importance of Conveying Inter-Facet Relationships for Making Sense of Unfamiliar Domains. In *The CHI2009 Workshop on Sensemaking*, Boston, MA, USA. 2009.

[34] Yee, K.-P., Swearingen, K., Li, K. and Hearst, M., Faceted metadata for image search and browsing. In *Proc. CHI 2003*, 401-408. 2003