Supporting interactive summarization for explainable exploratory search

Jing Li

Helsinki June 5, 2020

UNIVERSITY OF HELSINKI
Department of Computer Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Studieprogram — Study Programme
Faculty of Science	Master's Programme in Computer Science

Tekijä — Författare — Author
Jing Li

Työn nimi — Arbetets titel — Title

Supporting interactive summarization for explainable exploratory search

Ohjaajat — Handledare — Supervisors
Dorota Glowacka and Alan Medlar

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's thesis	June 5, 2020	43 pages + 4 appendices

Tiivistelmä — Referat — Abstract

Exploratory search is characterised by user uncertainty with respect to search domain and information seeking goals. This uncertainty can negatively impact users' abilities to assess the quality of search results, causing them to scroll through more documents than necessary and struggle to give consistent relevance feedback. As users' information needs are assumed to be highly dynamic and expected to evolve over time, successful searches can be indistinguishable from those that have drifted erroneously away from their original search intent. Indeed, given their lack of domain knowledge, searchers may be slow, or even unable, to recognise when search results have become skewed towards another topic.

With these issues in mind, we designed and implemented an interactive search system which integrated a keyword summaries algorithm, Exploratory Search Captions (ESC) to support users in exploratory search. This thesis investigated into the usefulness of ESC in terms of user experience, user behaviour and also explored impact of design decision in terms of user satisfaction.

We evaluated the ESC system with a user study in the context of exploratory search of scientific literature in Computer Science. According to the user study results, participants almost unanimously preferred the retrieval system that incorporated ESC; and the presence of captions dramatically impacts user behaviour: users issue more queries, investigate fewer documents per query, but see more documents overall. We demonstrated the usefulness of ESC, the improved usability of ESC system, and the positive impact of our design decisions.

ACM Computing Classification System (CCS):
Information systems → Information retrieval → Information retrieval → query processing → Query suggestion
Information systems → Information retrieval → Users and interactive retrieval → Search interfaces
Human-centered computing → Human computer interaction (HCI) → HCI design and evaluation methods → User studies

Avainsanat — Nyckelord — Keywords
layout, summary, list of references

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Thesis for the Networking and Services study track

Contents

1 Introduction

Search can be broadly divided into two categories: lookup and exploratory [1]. Lookup search is where users have precise search goals, such as question-answering and fact-finding. These search tasks are well-understood in terms of user behaviour and information retrieval (IR) system design. Exploratory search, on the other hand, is characterised by user uncertainty with respect to search domain and information seeking goals. Users performing exploratory search are trying to learn about a specific topic or otherwise acquire knowledge [1]. As such, exploratory search tasks tend to have open-ended search goals, where it can be difficult for users to initiate searches and assess the relevance of search results. Exploratory search is, therefore, considered to be challenging [2] and requires IR systems to provide additional support to help users articulate their information needs and understand the search results that are presented to them.

In exploratory search, document relevance is highly subjective. Users engaged in knowledge acquisition tasks will, by definition, lack the domain knowledge necessary to assess the importance of search results. In this context, users' search goals are ill-defined because there are many potential paths to obtain information and no objective criteria for when the task is complete. User behaviour studies have shown that users performing exploratory search behave differently from those performing lookup search: they scroll through more documents [3] and provide less consistent relevance feedback [4]. This can be frustrating for users, either because of time wasted inspecting lower ranked search results or because the accumulation of low quality feedback leads to query drift [5], degrading the quality of all search results. Unfortunately, it is unclear how to assess whether these behaviours are counterproductive – even during successful exploratory search, users' information needs are assumed to be highly dynamic and expected to evolve over time. IR systems should provide users with additional information summarising the current search session, helping users to assess whether search results are on-topic and to avoid unintentional digressions.

In typical exploratory search scenarios, users get a list of results produced by relevance ranking algorithms of the search engine after entering a search query. To gain knowledge in the search domain that users are not familiar with, users would browse through the result list, read and learn during the iterative search sessions. However, the whole search-and-learn process can be slow, leading to reduced user engagement. Traditional studies of IR systems primarily concentrate on developing

new algorithms to improve predictive accuracy and recommendations, while recent research has emphasised on the noticeable effect of visualisation and user interaction on boosting the performance of the IR system through improving user engagement in the search process [6].

To solve the above mentioned problems and to support users in investigating, exploring and acquiring new knowledge in exploratory search, in this thesis, we proposed and designed an interactive visualisation system with keyword summarising support. We focus on the integration of an existing search engine and ESC for exploratory search. Exploratory Search Captions (ESC) is a keyword summarization algorithm of the search engine results page that provides users with succinct, keyword-based descriptions of search results. In the context of exploratory search, ESC helps users to identify whether they have deviated from their intended search or as inspiration for follow-on search queries. We want to investigate whether and to what extent ESC supports users search activity and the impact of design decisions from user's perspective. This thesis aims to answer the following research questions (RQ):

- RQ1: Does ESC improve user experience during exploratory search?
- RQ2: Are there any behavioural differences between users of ESC and the baseline system?
- RQ3: What is the impact of design decisions on user satisfaction?

To investigate the above research questions, the basic rationale behind our method is to review previous related studies, find optimal way of design for integration and evaluate ESC system in practical scenarios through user study. User data would be collected to combine with survey results, such as usability survey, for further investigation. The user study is conducted in laboratory settings that situated participants in an essay writing scenario based on topics they were unfamiliar with. In this thesis, we focus on academic literature from Computer Science, but this approach can be expanded to other scenarios. The rest of this thesis is structured as follows:

- Section 2 will introduce basics of information retrieval and exploratory search, necessary background knowledge and related work in the field of exploratory search interfaces and behaviours.

- Section 3 will detail the mechanisms within ESC as a summarization algorithm, including the autoencoder model used by ESC, model training, data augmentation, caption generation and ensembling.

- Section 4 will describe the system architecture and interface design. It introduces the original baseline system which is an existing search engine developed by the research group, present the design decisions we made about caption system integrating ESC, the experimental approach we adopted and indicators of information search behaviours.

- Section 5 will illustrate the evaluation of our designed caption system and the performance of ESC algorithm. We compared the baseline system and the caption system. We record every step in the evaluation, from design, measurements to execution.

- Section 6 will analyse and show the results of the data collected in the user study. It covers the usability of the system to the user perception and behaviours.

- Section 7 will discuss the user study results and answer the three research questions in this thesis. We also summarised the contribution of the thesis and the future work in brief.

2 Background and related work

In this section, we introduce background literature, ranging from information retrieval to the adopted usability questionnaires for evaluation. We also identify related work on summarising and interactive interfaces, and studies of exploratory search behaviour.

2.1 Information retrieval

Information retrieval (IR) has been characterised in a variety of aspects: from a depiction of its goals, to comparably abstract models of its elements and processes. Though some of the characterisations have been in disagreement with each others, they have some commonalities. Generally, an IR system is regarded to perform the role of "leading the user to those documents that will best enable him/her to satisfy

his/her need for information"[7]. More commonly, information retrieval refers to how information is found, and the goal is identified as "for the user to obtain information from the knowledge resource which helps her/him in problem manangmant" [8].

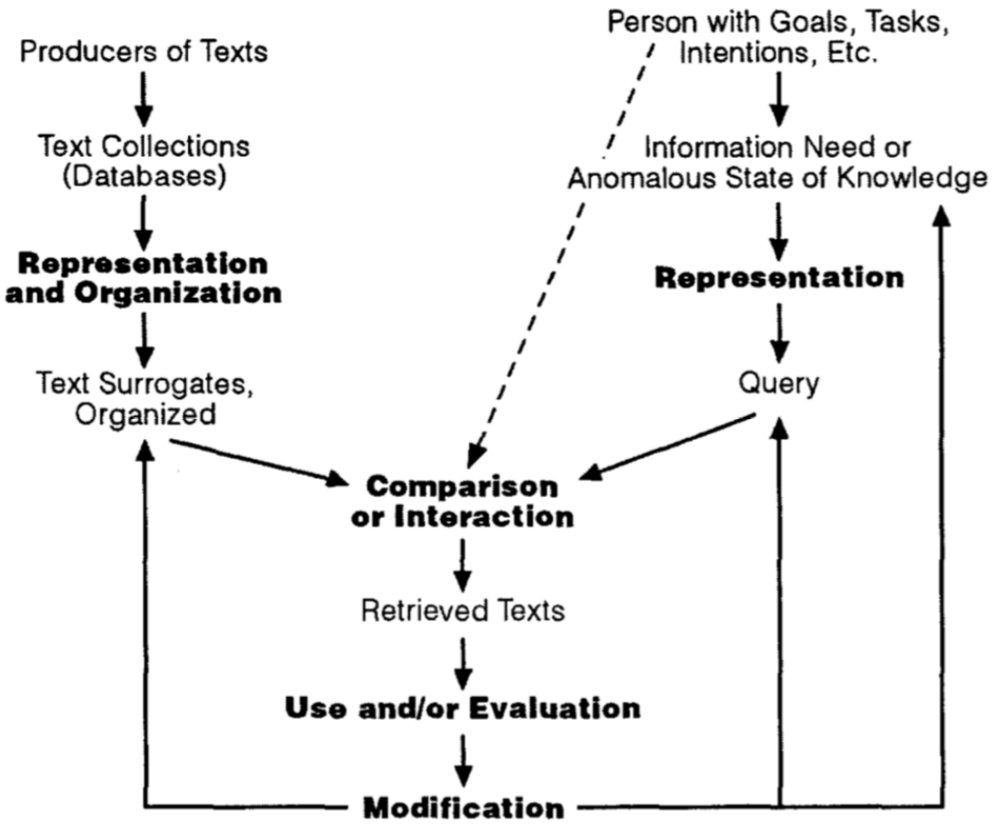

Figure 1: A general model of information retrieval based on Belkin 1992 [9]

Figure 1 presents models that illustrate such goals of IR. It is a general model revealing elemental objects and processes in IR, developed by Belkin [9]. When a user with some goals or tasks finds that the goals cannot be achieved because of insufficient knowledge or resources, then the user is prompted to participate actively in information seeking behaviours – for example, submitting queries into an IR system. The issued query, which should be formulated in a language that can be comprehended by the IR system, is a depiction of users' information need. This is presented on the right side of the figure. The other side of the figure pays attention to the information resources that the users would ultimately reach: from the producer of text, to the organisation of text representation into collections of text surrogates. Then, the two sides of the model merge into the comparison of query and text

surrogates, which results in potential relevant retrieved text. The retrieved texts are then used or evaluated and forwarded to the modification phase, such as modification of query. The process of query modification by means of user evaluation is referred to as *relevance feedback* in information retrieval.

In IR, both phrase and boolean queries have a long history. Early boolean IR systems permitted users to define their information needs using a set of boolean, such as AND, OR and NOT. But boolean systems have various drawbacks, e.g., hard for searchers to formulate good search queries. Additionally, the application of phrase in text representation has been studied since the early research into information retrieval. For example, Salton (1968) [10] depicted a diversity of experiments with the SMART system which uses phrases. The experimental outcomes received with phrases, however, were very blended, vary from small enhancements to reductions in effectiveness. They do not support the hypothesis that proper use of phrase would improve the performance of text representation.

However, a great amount of regular IR system users look at rank retrieval in IR systems. IR systems rank retrieved documents by giving each document a numeric score as an estimated rate of usefulness for users' query [11]. Various types of retrieval models have been proposed and studied for this procedure in IR, the three most frequently applied models are:

- Vector space model. Text is interpreted by a vertex of *terms*, which are regular phrases and words.
- Probabilistic model. Because genuine probabilities are not accessible to an IR system, probabilistic IR models calculate the probability of relevance of documents on a issued query.
- Inference network model. Retrieval of documents is demonstrated as inference procedure.

Nevertheless, there are occasions that users lack the knowledge or contextual understanding to formulate queries in a "correct" way that could be well interpreted by the system. For instance, how to find classic music to personal interest in the case that the user has little knowledge about Beethoven or Berlioz? The answer is we normally submit a provisional query and start from that point, investigating the remaining information, specifically looking for and inactively getting signs

about what the following stages are. Researches on supporting these sorts of situation is known as "exploratory search" [12]. Since exploratory search often requires additional support from IR systems [13], this thesis aims to help support users in exploratory search process in IR systems.

2.2 Exploratory search

Marchionini (2006) [1] divided search activities into two vast categories: lookup and exploratory search. He depicted exploratory search and lookup search activities as overlapping clouds considering that users may engage in multiple search types at one time. Exploratory search tasks emerge from circumstances where searchers "lack the knowledge or contextual awareness to formulate queries or navigate complex information spaces," and the search tasks naturally have ambiguity, vagueness and are dynamic. For instance, possible situations are that we need to discover something from a domain where we have a general intrigue however not explicit information [12]. In 2009, exploratory search was further defined as an information-seeking problem context that is open-ended, persistent, and multifaceted, and processes that are opportunistic, iterative, and multi-tactical [2]. An exploratory task is motivated by the searcher's desire to broaden his or her knowledge of a topic, i.e. to foster learning or investigation [14]. It has vaguely structured information needs and it can be difficult to phrase and often includes multiple aspects or a number of concepts [14].

Figure 2 illustrates that search activities defined by Marchionini, where exploratory search can be mainly divided into learn and investigate activities, each containing several sub-activities. Whereas lookup tasks return distinct formatted objects which are considered as answers to fact finding or questions, exploratory search tasks, on the contrary, do not have specific or well-structured answers. Searching to learn is a typical exploratory search activity as more learning material flourishes online. In our study, we choose the learning search scenario.

More frequently, exploratory search activities are also discussed with reference to their essential attributes. In a review of past research, a series of tasks attributes associated with exploratory search are distinguished [15]:

- Exploratory search tasks are general. For example, task descriptions can be very vague and ill-defined.
- Exploratory search tasks are open-ended rather than focused, often targeted

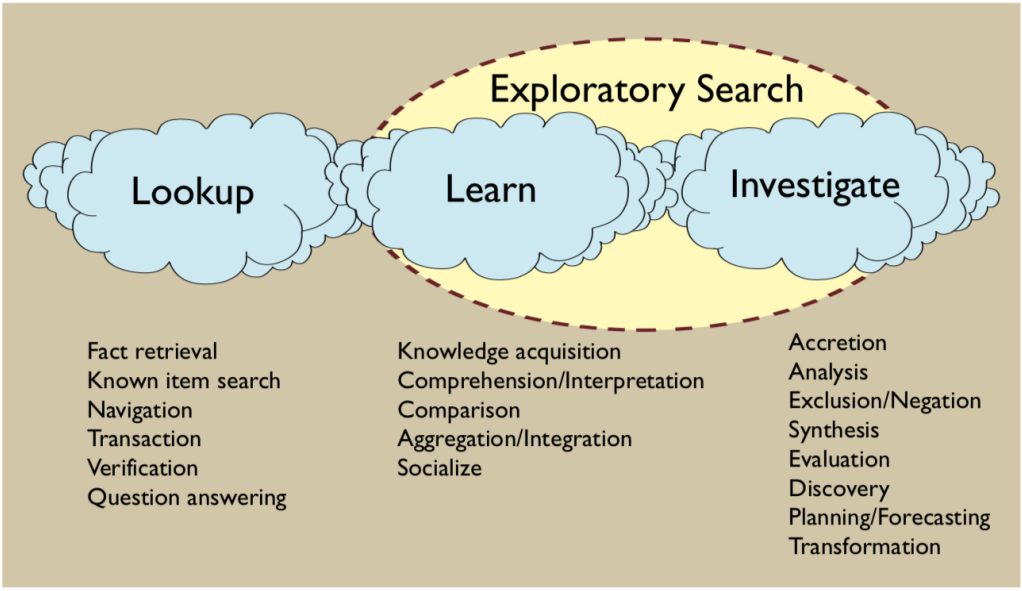

Figure 2: Search Activities according to Marchionini 2006 [1]

at several objects or documents. For instance, an earlier study by Marchionini [16] noted open-ended information tasks and questions, depicted as questions "for which there was no specific answer".

- Exploratory search tasks contain uncertainty as well as motivation derived from ambiguously defined problems.
- Exploratory search tasks have a dynamic information need, dynamic process and they often develop with time. The changes may come from the evolving need and understanding of searchers as their knowledge broadens during the search sessions.
- Exploratory search tasks are multifaceted and the procedure may be complicated. Depicting a search task as multifaceted implies that it "incorporates various viewpoints or various ideas".
- Exploratory search tasks commonly come with context information or cognitive behaviours, such as decision making, organizing and inspecting search results and, most frequently, sensemaking.

2.3 Relevance feedback

Relevance feedback asks information seekers to make relevance judgments about search results and then execute a revised query based on those judgments [1]. In relevance feedback, users do not describe in detail what their information needs are, instead, they simply mark relevant objects. Typically, users would mark the returned objects or documents that they consider relevant, then the system would give improved recommendations to users, such as potential query terms, etc [17]. Relevance feedback appears to be a natural approach to support searchers in reformulating their queries in a heuristic manner. It has been shown that relevance feedback can be especially effective for content-based image retrieval scenario [18], and for improving retrieval accuracy and performance [19][17].

In spite of the benefits, however, studies showed users' reluctance to provide explicit feedback. For example, in an exploratory searching experiment in the area of relevance feedback [20], responses indicate a tendency that although the searchers understand how to use relevance feedback, they lean toward finding relevance feedback not very useful during the search sessions. On the other hand, marking relevance feedback all the time can be cognitively demanding, but not giving feedback may lead to imprecise judgement in systems. For example, users may stop giving feedback when they are happy with search results and switch to a browsing behaviour, but then some systems interpret this to mean everything is not relevant. User perception and experience of relevance feedback can also be affected by the framing of the relevance feedback mechanism, e.g. users prefer giving positive relevance feedback as opposed to negative relevance feedback [21].

2.4 Types of systems for exploratory search

Over the last two decades, there have been many specialised interfaces created to support different aspects of exploratory search. These include summary visualisation to enable faster relevance feedback [22, 23], automatic adaptation based on search task [24], interactive visualizations that allow users to direct their search [25, 26], interactive exploration of faceted data sets [27] and interfaces to support query formulation [13].

2.4.1 Faceted search

Faceted search provides an iterative way of refining the search results by facets [28]. Faceted search systems helps searchers to find what they are looking for, not only by keywords related to their information need, but metadata specified by searchers query refinement. Faceted search utilizes organized metadata to give a diagram of results and consolidate interactive categories into result lists. This helps users limit the result range and explore the results without reformulating the queries [29].

Facet is defined as "a set of meaningful labels organized in such a way as to reflect the concepts relevant to a domain"[30].Faceted search constructs information and uses filters to support users investigate a set of returned items [31]. For example, online accommodation platform, such as Airbnb, allows users to narrow down their selection by adding limits regarding facets such as the price, district and the type of an accommodations. On the basis of a faceted taxonomy, faceted search enables users to choose facets and facet terms in an iterative manner which aims to narrow down the search results and prevent the contextual problems or pitfalls through the exploitation of global characters or features [31].

When the query is too broad or the number of results is excessively large, exploratory search may not have a good performance since it will be hard to decide the relevance ranking of the numerous results. The faceted method bunches the results into a wide assortment of categories as indicated by metadata, such as keywords. As a result, users can only concentrate on the results to their needs, saving time from browsing other irrelevant ones, which also adequately eases the data over-burden issue [28].

2.4.2 Summary exploratory search interfaces

During the exploratory search session, sometimes the search results are too long for users to browse easily, or the ranking of the search results seems less reliable. To support users in such scenarios, Mika (2005) [32] developed a search result categorisation algorithm and a filtering interface: the documents are processed into several categories, and a list of categories are displayed in the left sidebar on the result page. During a search session, users can filter the results by clicking on a category. Researchers conducted a longitudinal study with 16 participants who used the prototype system for two months and analysed their interaction data. The results confirmed the usefulness of result categorization in real settings, especially when the ranking of search results fails – users feel that the ranking does not contain use-

ful search results and therefore fail to fulfil their information needs. However, the factors affecting the usefulness remain complicated, the users' behaviour was not controlled.

While Mika aims to improve the result evaluation phase, Kules et al. (2008) [33] dug into the effectiveness of result visualization for users mainly in the phase of discovery and the search process. A study of existing information systems is given as examples that have been demonstrated to effectively improve the user searching experience.

Matejka et al. (2012) [22] introduced the Citeology System which provides summary visualisation in a combined view of heatmap and line chart to enable faster relevance feedback. It looks into the relationships between publications in different years through their citations. However, the information in the system is generally presented textually and there is no way to look at multiple generations of ancestors or descendants or get a view of the overall connection network of the corpus [22]. Since the heatmap visualization focuses on the big picture, the lack of zooming capability in the applet and the simple text-based formated data decrease the practicality.

Although these summarising interfaces give an aid to users to varying extents and aspects, they fail to deliver the summarised information in a intuitive way and so inexperienced searchers can hardly make good use of these professional interfaces. Our goal is to design an intuitive and effective interface which is friendly for searchers at all levels: from novice to skilled searcher.

2.4.3 Interactive recommendation interfaces

In most recommender systems, users have little knowledge of how the recommender engines work, their derived distrust of search results as well as the uncertainty in the exploratory tasks have therefore adversely affected the usability of the search systems. To engage users in the search process more positively and place them in continuous control, researchers have devised highly interactive user interfaces, and many studies have been done to improve the user engagement, controllability, transparency, etc. The main advantage of interactive visualization is that a multi-dimensional representation allows the user to more easily see multiple aspects of data while being in control when exploring information.

There have been various attempts to engage users in the feedback loop through

interactive network visualisations. Combined with machine learning algorithms, Chau et al. (2011) [34] developed a system named Apolo that engages the user in bottom-up sensemaking to gradually build up an understanding over time by starting small. They argue that just finding and filtering or even reading items is not sufficient, so they endeavour to provide support with multi-group sensemaking of network data. Moreover, users can incrementally and interactively build up a personalized visualization of the data in the system. Though both quantitative and qualitative results are positive in the usability evaluation, the system does not take into account the evolving needs of users over time, such as the groupings created by users through merging or subdivision. Zhao et al. (2013) [27] designed an interactive interface in the form of node-link graph, focusing on faceted data sets. It allows users to discover the implicit and explicit relations among the large-scale data and also manages their queries in a multi-focus way.

Interactive recommendation interfaces that allow users to direct their search is another important branch of research in the area of interactive visualizations [25, 35]. Users can directly manipulate document features keywords to indicate their interests and reinforcement learning is used for user modeling by allowing the system to trade off between exploration and exploitation [25]. Verbert et al. (2013, [6]) proposed a combined approach of both interactive visualization and traditional ranked result list recommendation to improve users' trust in the results offered by the black-box recommender engines. The approach considered personalized relevance and social relevance prospects and enabled users to explore interrelated bookmarks by other users using a cluster map visualization.

In addition to the above approaches, Medlar et al. (2016) [13] presented a system that supports query formulation in exploratory search for scientific literature. It applies an interactive alluvial diagram showing what topics there are in different years and how they are related to one another in the phase of discovery and query formulation.

These types of interfaces can greatly aid users in conducting exploratory search, however, they fail to account for the evolving nature of user knowledge in exploratory search tasks and the associated drift from the initial search query. Our proposed method will improve the search session transparency to the user by summarising the content of documents linked on the screen.

2.5 Studies of exploratory search behaviour

Understanding user behaviour in exploratory search can also inform interface design and search results presentation. There are a number of models of user behaviour that specifically address these issues. For example, Information Foraging Theory [36] is a model of exploration that predicts information seekers' decisions from the expectation of information gain. This model was recently applied to predict the subjective information needs of users conducting exploratory search [37].

Hassan et al. (2014, [38]) proposed a model that aims to disambiguate search exploration from struggling with search. Through their analysis, detectable behavioural differences can be seen between struggling and exploring, and use these insights to develop machine-learned models capable of accurately distinguishing between the two situations[38]. As user factors have an effect on the user behaviour and search results of exploratory search, Jiaxin et al. (2018) [39] dug into users' domain expertise, investigated and compared the effects of domain knowledge level across three different domains. Their work maps domain expertise level on users' interaction and search outcomes.

In general, previous studies of exploratory search behaviour point to its dynamic nature with narrowing and broadening of search queries throughout a search session as well as changes in click behaviour and dwell time as search progresses [40]. Our proposed method will further facilitate the capture of gradual changes in users' search intents throughout a search session, which could lead to further improvements in user modelling and search results presentation.

2.6 Usability questionnaire frameworks

Usability is not a common quality that exists in any genuine or absolute sense. Maybe it tends to be best summarised just like a general nature of the fittingness to a motivation behind a specific artefact [41]. It can only be measured under a certain context, hence there is no absolute metric for usability. The measurement of usability is to test the usability of the artefact in the context where the artefact is used. This, in exchange, means that the measures of usability is dependent on the context, or even defined by the context as well. In spite of this, there is a demand for broad generic measures which can be applied to reflect the usability of an artefact over a wide scope of contexts and settings. A few generic aspects of usability are proposed by ISO 9241-11 and state that the measures should include:

- effectiveness
- efficiency
- satisfaction

Although it is feasible to discuss features of usability in a general sense, due to these varying classes of usability, it is quite often the situation that it will be hard to compare two or more systems having various purposes and which work in different manners [42]. The exact measures to be applied inside each of these classes can hold significant differences. For instance, metrics of effectiveness are, with no doubt, strongly related to the sorts of assignments that are conducted with the system; a metric of effectiveness of a word refinement system may be the quantity of letters composed, and whether the composed letters have misspellings.

This section introduces two highly cited usability assessment frameworks.

2.6.1 SUS

The System Usability Scale, commonly referred as SUS, is developed under the requirements of general measure in 1996 by John Brooke [41] as a "quick and dirty" questionnaire scale. SUS is a straightforward *Likert* scale with 10 items. It enables a fast and effective assessment of people's perception of the usability of a computer system which they have worked with [42].

To select the items for a Likert scale, the technique used in SUS development is to identify and select extreme expressions as item statement. Instead of general expressions which earn loads of agreement between respondents, extreme statement can arouse strong opinions in respondents on the agreement or disagreement towards the artefact. For example, if the subject is to investigate people's attitude to crimes, respondents would have a more clear answer when they are asked to answer statements such as "things like this", than general statements such as "It is not good" [41].

The construction of SUS starts with a pool of 50 collected items and then items prompting the highest extreme feedback were chosen. The coverage of the 10 chosen statements includes multiple system facets, for example, the demand for support, integration, and intricacy. The score in SUS ranges from 0 to 100. From project administrator to IT expert, it offers a simple score that can be handily comprehended

by the broad scope of people who are ordinarily engaged with the improvement of artefacts, such as product and service [43].

Studies have confirmed that SUS is an important assessment tool, being rigorous and dependable. It mirrored a pressing need in the area of usability for a tool that could efficiently gather users' subjective responses about an artefact's usability [43]. So we choose to use SUS to evaluate the usability of our proposed system.

2.6.2 ResQue

A recommender system actively gives suggestions of items to the interest or needs of the users depend on their behaviour or their expressed preferences. As recommender technologies are getting generally acknowledged, it is of utmost significance to have the capability to describe user experiences and users' subjective preferences of the technology. In response to the needs, Recommender systems' Quality of user experience(ResQue) is a unifying evaluation framework that aims for assessing the quality of the suggested items in a recommender system, as well as usability, handiness, interface and interaction qualities, etc [44].

As an evaluation questionnaire framework, ResQue comprises of a collection of constructs. It starts from defining the construct domain, and then comes to sample question generation and phrasing. The potential question items are divided into four construct level considering four aspects:

- Users' perception of system qualities. For example, the quality of suggested items, the interaction quality to deal with user preferences, the quality of interface to introduce and clarify results, and the quality of providing data adequacy.
- Users' beliefs of the qualities, comprising perceived usability, control and transparency [45].
- Users' subjective attitude, containing confidence, trust, satisfaction, etc.
- Users' behavioural intentions, containing their agreement, willingness of purchase, return and tendency to introduce the system to others.

The full model consists of 15 constructs and 43 questions. ResQue can be used as a complete or a modified version of an evaluation survey, 43 and 14 questions respectively, that can help practitioners and specialists to investigate the appropriation of

recommender systems. The model can be used to evaluate various kinds of recommender frameworks [45]. In our experimental design, we selected 12 questions from the database as a modified ResQue questionnaire for the evaluation of our proposed approach.

3 Exploratory Search Caption

In this section, we introduce the autoencoder model used by ESC for representation learning of ranked search results. We described how the model is trained, including how data augmentation is performed, and describe how it is used to generate captions and ensembling.

ESC generates search captions using a combination of semantic and lexical information. Semantic information comes from a sequence-to-sequence autoencoder, which we used to learn a distributed representation for ranked documents. In this distributed representation, semantically similar sets of ranked documents will be proximate to one another in vector space. By encoding the current search engine results page of documents into vector space, ESC can find nearby examples of ranked documents that are associated with known queries. This approach is analogous to finding semantically-related terms in word embeddings [46]. In situations where it is difficult to derive semantic information, such as when there is no coherent theme present in the search results, ESC falls back to a simpler, lexical method. ESC automatically detects when the autoencoder makes weaker, less coherent predictions using a logistic regression model, which is used to determine the proportion of captions to be presented from each method.

3.1 Autoencoder model

Autoencoders are unsupervised learning methods composed of two neural networks: an encoder network and a decoder network [47]. The encoder network transforms unlabelled training examples into a vector representation and the decoder network reconstructs the original input from this learned representation. The autoencoder learns a function, $h_{W,b}(x) \approx x$, of the weights, W, and biases, b, of the neural network such that the mean squared error is minimized between the input data, x, and its reconstruction, x', in the output layer: $\mathcal{L}(x, x') = \sum (x - h_{W,b}(x))^2$.

As shown in Figure 3, our sequence-to-sequence autoencoder model consists of an

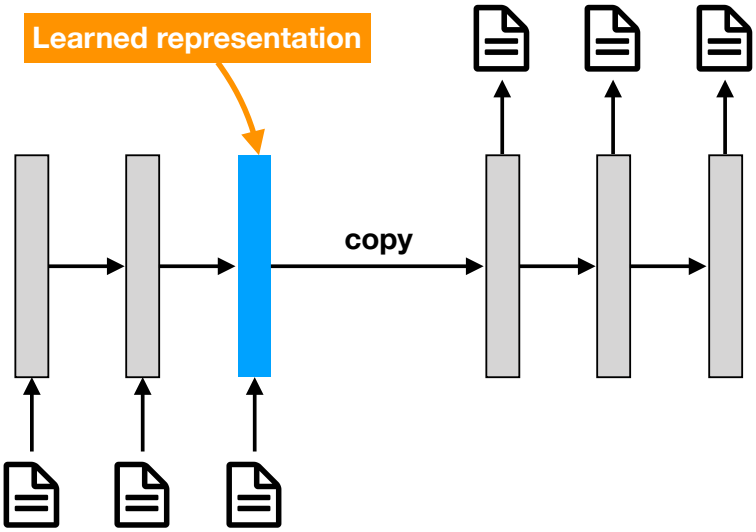

Figure 3: Sequence-to-sequence autoencoder model. The output of the encoder network (left) is the input to the decoder network (right). The learned representation (blue) is used to find semantically similar search results pages.

encoder LSTM and a decoder LSTM. In our case, we want to create a learned representation of ranked search results, where each search result is modelled as a time-step in the LSTM network.

3.2 Data and model training

In order to generate the ranked search results to train the autoencoder, we used:

- A corpus of documents from a compiled collection of $\sim$170,000 Computer Science articles from the arXiv preprint server (`www.arxiv.org`, downloaded November 2018). Each document was represented by concatenating the article title and abstract.
- A method to generate queries and a retrieval algorithm. We preprocessed the text of each document and detected common phrases by extracting bigrams, which were allowed to skip over stop words (making "eye of the beholder" a valid bigram).

As the number of query phrases was relatively limited for training the autoencoder, we performed data augmentation to increase the diversity of the training set. We used 95% of the data for training and the remaining 5% for validation to avoid

overfitting. We used Adam [48] with a learning rate of 0.001 to train our autoencoder model using a vector length of 300 (i.e. the length of the learned representation in Figure 3).

3.3 Caption generation and ensembling

The autoencoder forms the basis of a nearest-neighbour-based approach for caption generation. The autoencoder projects ranked search results into a vector space containing the learned representations of search results associated with known queries, the queries from the nearest-neighbours are used as candidates for captions. We display the top-10 captions that would be unique after stemming, to avoid redundancies such as "neural network" and "neural networks".

When presented with search results not found in the training set, the autoencoder outputs novel vectors whose positions in vector space reflect the contents of those search results. For example, if presented with documents related to "information retrieval" and "neural networks", then these terms should be among the top ranking captions. Indeed, summing the vectors from our caption database associated with the queries "information retrieval" and "neural networks" achieves exactly this: the nearest neighbours include "neural network", "information retrieval", "neural", "information retrieval systems" and "feedforward network". Ideally, when the autoencoder is presented with a mixture of documents from two separate queries, the resulting vector representation should approximate vector arithmetic in the same manner.

The autoencoder's precision varies as a function of the average positive pointwise mutual information (PPMI) between all pairs of captions. Performance is worse when average caption PPMI is lower, making caption PPMI a predictor of caption quality. When ESC is predicted to have lower performance, we can use lexical information to improve results.

Finally, we use a given logistic model for ensembling by using the probability to define the proportion of captions to use from the autoencoder, with the remainder taken from a lexical method. The logistic model used mean keyword pair PPMI to predict whether the captions had a coherent theme, and the response variable was defined as a precision > 0.6 and that this threshold was determined through simulation. We use the probability output by the logistic regression model to define the proportion of keywords to use from ESC, with the remainder made up of the keywords with the highest Okapi BM25 term weights. The intent is to take advan-

tage of the autoencoder in scenarios where it performs well, but fall back to a lexical method when we detect there is no coherent theme to the search results.

4 System design

The goal of this chapter is to investigate how to integrate keyword summaries algorithm into an existing search engine effectively and in a user-friendly way. In this thesis, we focus on the academic practice – scientific literature search, with a baseline system which we refer to as the baseline system. We start by introducing the system and then describe our practice on the interface design of the caption system incorporating ESC. We explore the design solutions for the system to integrate ESC in terms of interface and practicality. Next, we identify the importance of taking an experimental approach and the matching information search indicators.

4.1 System architecture and interface design

In order to give a context on what the system is like, we first give an overview of the baseline system before going deeper into the design of our improved caption system.

4.1.1 Baseline search engine

We used a literature search system which has a typical layout and functions as many other search engines – a navigation bar at the top with a search box, and a list of result articles with title and abstract. Figure 4 depicts the interface.

As we want to enable experimental features in the system, like logging, we can not use Google scholar although it is the widely used interface for literature search. So we choose to use a baseline system with high flexibility, as a case study and integrated the ESC into the system. In many aspects, the interface is very close to Google Scholar: every results block contains the document title, authors, date of publication, venue and abstract. In the baseline system, a search is initiated by typing a query into the search box at the top of the page, which results in a list of articles appearing on-screen. Users can click on any article on the screen and continue to explore the details if needed.

Moreover, as Google scholar does not allow free access to its data, we use a free digital library, arXiv, and all Computer Science papers from the arXiv data set as

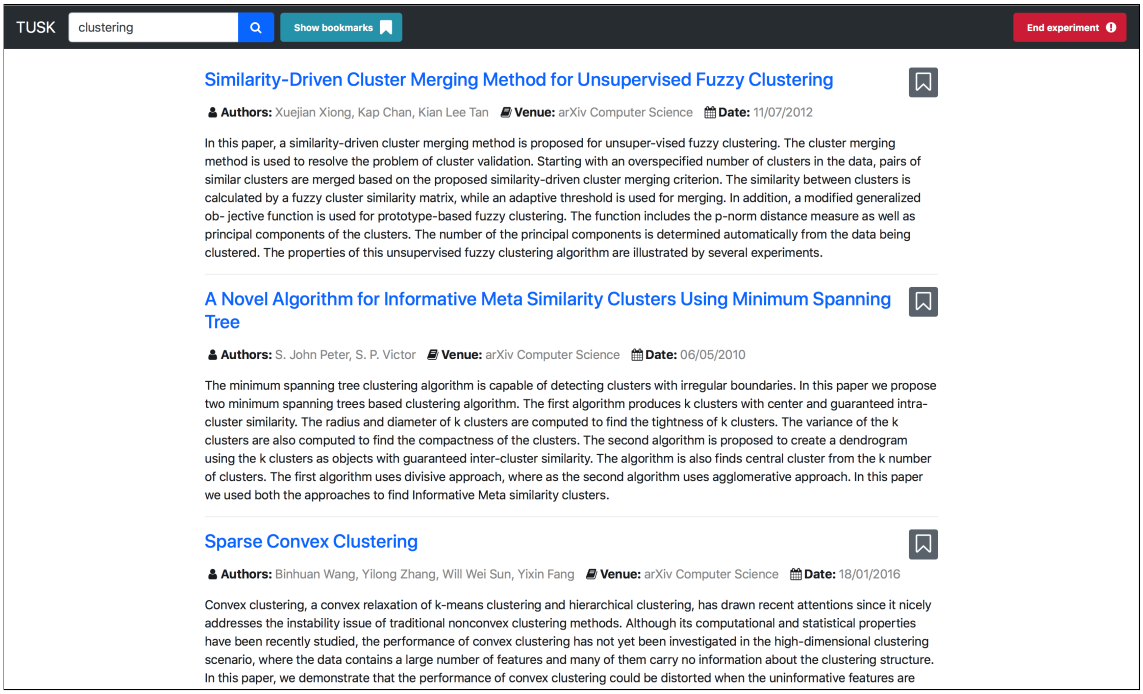

Figure 4: Screenshot of the baseline system interface. This image shows only part of articles that are displayed as outcome. Users can perform infinite scroll down to explore more articles.

our document corpus. arXiv is also one of the most accepted open access digital libraries in the domain of computer science.

In exploratory search tasks, users are predicted to investigate more results [2]. However, users have the tendency not to move after the first search result page despite the fact that they are interested in exploring more articles [49]. Thus, to investigate users' behaviour without forcing them to click for "next page", our baseline interface implements infinite scroll, fetching additional search results simultaneously on-demand as the user scrolls down the page.

In addition, the baseline system has a bookmark function where searchers can mark the articles of their interest by clicking on the *mark* icon at the top-right of a result block (Figure 5). To review the bookmarked articles, searchers click on the teal button next to the search box.

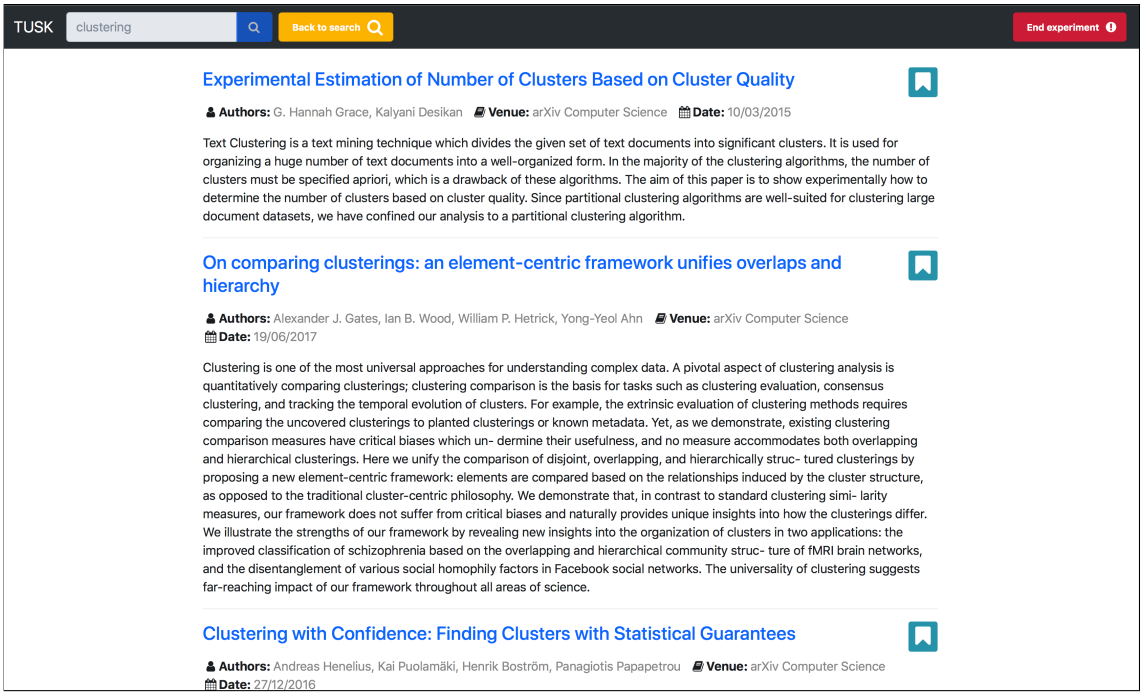

Figure 5: Screenshot of the bookmark list in the baseline system interface. This image shows only part of bookmarked articles that are displayed as outcome. Users can go back to result page by clicking on the yellow button.

4.1.2 Interface and system design with keyword summaries

For open-ended exploratory search tasks, it is of vital importance to provide tools that support users with concentration through organizing search results in a meaningful way and also enhancing their learning processes. Studies [6] have shown positive effects brought by visual elements and strengthened interaction that can largely increase users' engagement in the search iterations, boosting the performance of IR systems.

Considering the needs from the user perspective, we derived a few user interface principles for the caption system:

- Must be intuitive to understand even for first-time users.
- Must not be ambiguous to use or understand.
- Must follow the style of the original interface, not breaking the harmony of the entire system.

- Must be visually distinguished in a clear manner, not blending with other interface elements.
- Must not distracting for users to perform the search sessions, e.g., looking into the result list.
- Must deliver the complete support provided by ESC with as little disturbance as possible.

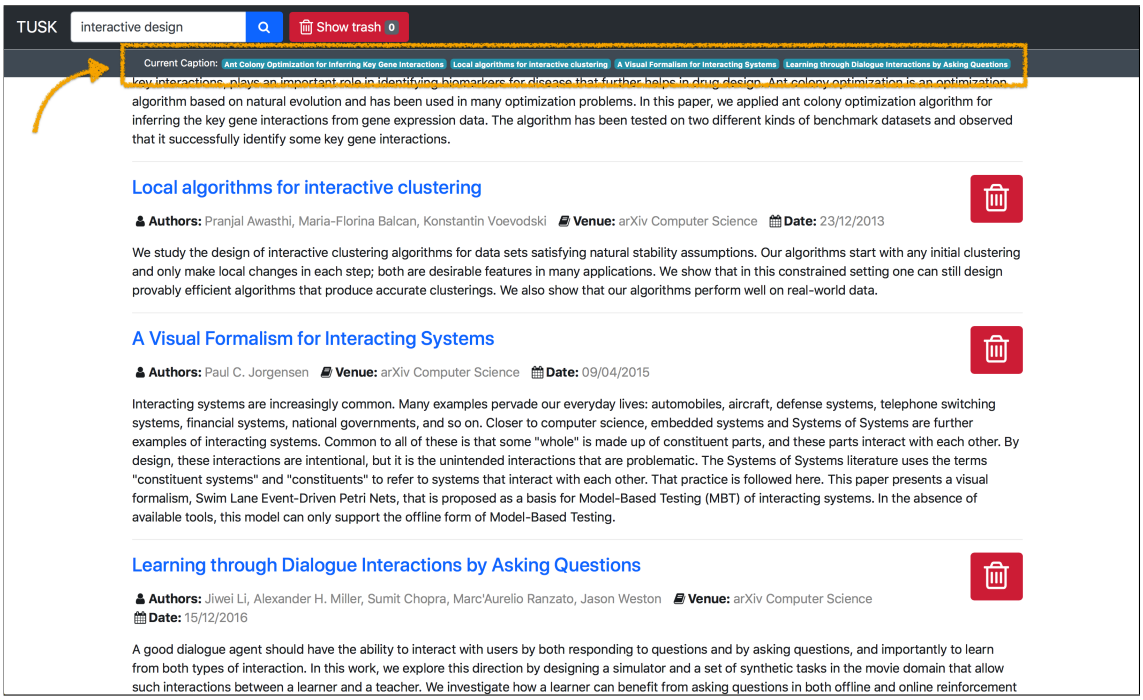

Figure 6: First version of ESC incorporated system: horizontal caption bar. Captions would change as the user scrolls through the result list.

With these principles in mind, the first version of caption system situates captions in a horizontal list under the navigation bar (Figure 6). The motivation for the visualisation was to offer users a simple overview of the articles they are browsing in the ongoing search session. As the user scrolls down, the displayed articles would change and accordingly the captions would change. We initially thought that the caption bar below the search box works as a typical way to balance the performance and perception of dynamic keyword summaries. However, the weakness of this layout was identified quickly after our pilot test: on the one hand, users find it hard to understand the relations between captions because horizontal list can barely depict the ranking relationship among the captions; on the other hand, the fixed width of

the caption bar limits the number of caption displayed especially when the length of the captions is uncertain. In addition, plain text captions are not visible enough to be perceived by users – they may not even realize there is a supportive function like this throughout the search session. Visual animations of caption may be more disturbing than beneficial as a results of the centering position. It is also hard to add more instructions or interactions to these captions since the space is really limited.

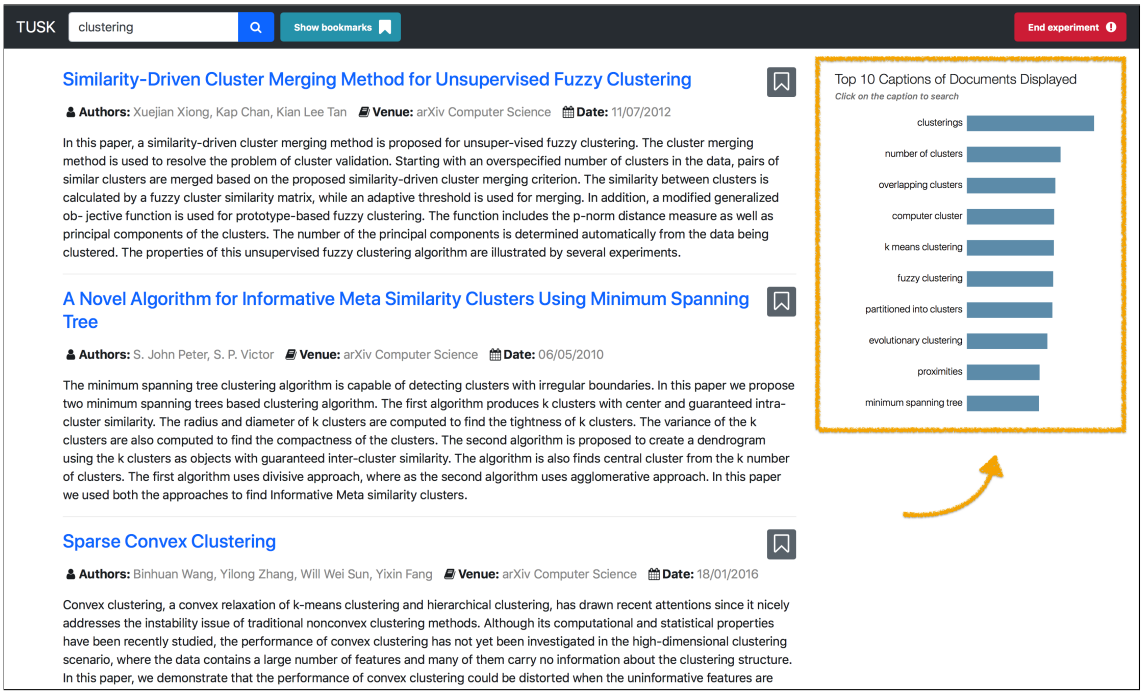

Figure 7: Second version of ESC incorporated system: vertical caption sidebar. Top 10 captions are displayed by default and it changes as the user scroll through the result list. The link to the demo video: `https://youtu.be/vn8_Vli6tzo`

These findings prompted us to transform the captions into a dynamic bar graph in the second version of design. We replaced the horizontal caption bar with a vertical sidebar section to present captions, situated at the right side of the screen(Figure 7). The coloured bar length is proportional to our confidence, and can directly demonstrate the top-down ranking of the captions for users to obtain the big-picture information at a simple glance. The ranking is dynamic as the users scroll up or down the result list – a caption would move up if its relevance probability increases as well as its ranking. When the captions apply an animation to reflect the changes, the ranking movement of each caption can be easily perceived by users. We also added a tool tip to provide faster navigation, when users hover on the caption bar.

Furthermore, to dig into a certain caption, users can start a new search session using the caption as the search query by simply click on it, either on text or the bar. The core advantage of this interactive visual design is to provide users with a multi-dimension view that allows them to comprehend different parts of data more effectively while staying in control when discovering new information.

4.2 Experimental approach

We designed and conducted an experiment to evaluate the caption system incorporating ESC. A good experiment should include practical tasks, questionnaires, and subsequent interviews that provide researchers with sufficient collection of data while having control over other elements that have an impact on search behaviour [3].

We control over three factors that could influence user behaviour in the search process:

- domain knowledge
- search expertise
- subjective task difficulty

Past studies [50] found that the domain expertise of students brought forth better search performance on the condition that the participants had adequate information-seeking experience. In our settings, participants with good search skills perform tasks on topic domains they are unfamiliar with. The tasks are set to be in the scenario of academic information seeking, for the reason that a major purpose of exploratory search is new knowledge acquisition, which is especially significant in the academic environment [15]. We chose the machine learning (ML) domain to produce all the tasks on the grounds that there is a decent inclusion of ML courses at the university, which makes it simpler to recruit participants with similar knowledge of the subjects. Furthermore, numerous ML articles are freely accessible in our selected data set, and ML experts are easier to find for task assessment in the later stage of the experiment.

Despite being specific and involving precise requirements in participant recruitment, the exploratory tasks are general, open-ended and ill-structured in a sense that we do not specify how participants should find a solution or how much information

is needed for completing the tasks. It means that participants, as searchers, will independently interpret the tasks, their queries, interactions, and their answers to the tasks – while retaining experimental control to a certain extent.

4.3 Indicators of information search behaviours

After analysing previous work, we identified the most appropriate indicators of information search behaviour for our study. We focus on behaviour that could be collected easily without interrupting users performance and meanwhile would help discover the insights throughout the experiment.

We select two behaviours related to query.

- Number of queries issued per session: In exploratory search tasks, to formulate queries, users spent more time to familiarise themselves with the topic [2]. To capture this behaviour, we select number of queries the user issues per search session.
- Number of documents inspected per query: Users benefit from looking through multiple documents from time to time, as the inspected documents enables them to grab a better understanding about the search domain and the topic, and particularly, give the users extra hints on search query formulation. [51]. We included the measure because the relevance of documents reflects the quality of formulated query.

User interactions have been an important measure, from which we selected two behaviours:

- Maximum scroll depth: The maximum number of documents from the result list thats users were exposed to through scrolling. In exploratory search tasks, users tend to explore more documents [1] but the scroll behaviour is hardly inspected. We create index for each document in the result list and then calculate the number of documents the user is exposed to by infinite scrolling.
- Bookmark list: The total number of documents the user bookmarked in the search process. Bookmarks consist of pages or documents that the user feels helpful, thus clearly flagging that those pages are considered relevant [52].

We also selected two behaviours concerned with time:

- Duration of dwelling: the time users spend reading the clicked documents. Dwell time is utilised as an indicator of document relevance when participant information and task type are known beforehand [53].
- Task completion time: Total amount of time user spend on a search task, from the moment submitting the first query till the *end* button is clicked.

5 User study design

We conducted a user study to evaluate how well the system incorporating ESC (caption system) works compared to the system where users do not have support from captions (baseline system). The purpose of this study is to gather information on user interactions and search behaviours in both systems, so as to investigate how effective and useful the ESC algorithm can be in supporting users in exploratory search. We also concentrated on users' understanding of the generated captions and their engagement in search iteration and users' perceptions of captions in context of the search interface.

5.1 Participants

To recruit participants we posted advertisements in the Computer Science (CS) department mailing list of University of Helsinki. We selected university students with at least 3 years of study in Computer Science or Data Science, including postgraduate students and PhD students. Because they are active users of exploratory and scientific search as a major part of their work [54]. Nineteen participants took part in the study. Eight of them (42%) were female and 11 were male. The mean age of the participants was 30 years (min. age = 23 and max. = 48 years). All participants were computer science students: 8 MSc students and 11 PhD students. They have studied at the university level for at least 4 years, whereas the median number of years was 7 (min.year = 4, max. = 29).

To ensure the exploratory search skill within the participants is at a moderate level, they were asked to fill out a background questionnaire before the study. According to their answers to the questionnaire, Google Scholar is the most frequently used tool with 18 participants; ACM and arXiv are also popular among the participants; while 9 participants selected *Just Google* as one of their search methods. The medium frequency of performing scientific literature search was 4.3, whereas

the lower boundary was 3 (ratings are given on a 5-point Likert scale as 1[*never*] to 5[*a great deal*]). All the participants were experienced users of exploratory literature search tools.

5.2 Design

We designed a within subject study, where every participant performed one search tasks in each system: the caption system and the baseline system. To avoid the order effect, for each participant, the sequence of the topics and the search systems was determined by coin tossing.

We compared two systems: one incorporating ESC and a baseline without captions. Thus, any differences in observed user behaviour are generally attributable to the presence of captions. Aside from the presence or absence of captions, both systems were identical. A search is initiated by typing a query into the search box at the top of the page. In the caption system, a list of 10 captions summarising the search results currently displayed on the screen is shown in the right-hand margin. The captions are ordered according to their relative importance, which is quantified by the length of the bar next to each caption (shorter bars indicated greater uncertainty). As users scroll down the page, the captions change to reflect the contents of the documents currently visible to the user, i.e. the order/importance of the currently displayed captions changes or some of the displayed captions are swapped for new ones.

5.3 Tasks

5.3.1 Task Design

In the main tasks, users were instructed to write a short essay draft on a given topic. The task descriptions followed the template:

> *"You are going to start a new research project on the following topic: X. You would like to learn as much information as possible about this topic, e.g. applications, problems, specific algorithms. Write your answer as an essay draft or in bullet points, and bookmark at least 10 relevant articles that you could use as a reference in writing the article."*

The search task with each system was limited to 30 minutes with a short break in-

between the two tasks (average task duration was 29.5 and 28.4 minutes for caption and baseline systems, respectively).

The task end time was logged when the participants thought the task was finished and clicked the red "End Experiment" button on the top-right corner of the result page. The users were provided with pen and paper as well as a text editor for note taking and could select the method they felt most comfortable with. All the participants opted for the text editor.

5.3.2 Familiarity Questionnaire

To ensure that the domain knowledge of topics for each task, we provided a questionnaire to subjectively rate the familiarity with a list of topics which is carefully selected for the tasks. Before the experiment, users rated their familiarity, on a 4-point Likert scale, with the following topics: robotic surgery, political bias online, sports analytics, gender recognition, cancer diagnosis, autonomous driving evaluation, gender bias in natural language processing and music recommendation. The topics were selected beforehand to ensure appropriate coverage in the data set used in the experiment. To ensure that users would engage in exploratory search, only topics with low levels of familiarity for a given user were considered, i.e. topics marked as 1 or 2 on the Likert scale. If a user indicated only two topics that they are not highly familiar with, then those two topics were randomly assigned to the two systems. If users indicated low familiarity with more than two topics, then they were asked to select their two preferred topics, which were then randomly assigned to the two systems. Users who did not indicate low familiarity with at least two topics were excluded from the study.

5.4 Measurements

For every task performed, we logged the following details: search topic for the task, system used (caption system or baseline), the length of note-taking on the given paper. Both system logged task starting and ending time as well as the user interaction data: the number of clicks, maximum scroll depth, bookmarked documents, all queries used in the search, duration of results page browsing, and the time spent on reading the clicked documents. Specifically, for the system with captions, we logged the interactive data with captions: captions that users clicked on as a new search query.

We also collected qualitative feedback on the search systems through semi-structured interviews at the end of the experiment. As an indicator of the task performance, an expert assessor from the machine learning domain rated, on a 5-point scale, the quality of the answers returned using both systems for the search tasks. The assessor conducted the review in a blind manner, with no awareness of which system was used in the task.

5.5 Procedure

All the studies took place in a controlled laboratory, with a 15-inch Macbook Pro 2016. First, we explained the purpose and procedure of the study to the participants. We informed the participants that the purpose was to "test our search engine for scientific literature." Therefore, we instructed the participants to perform the tasks as they would normally do using the search systems we provide. Further, we explained that the search tools used all Computer Science papers from the arXiv data set as our document corpus and search systems basically work the same as most-commonly used literature search tools. Figure 8 shows every step the participants were involved in this study.

The experiment comprised the following steps:

1. **Consent form** We asked the participants to sign a form of consent to take part in research.

2. **Short tutorial** We showed the participants a 1.5 minute video tutorial on how the systems work. This was followed by a 15 minute practice session to allow the users to familiarize themselves with the two systems at their own pace. In the practice session the participants were instructed to perform searches related to their own interests.

3. **Pre-experiment questionnaire** Prior to the study, we provided a pre-experiment questionnaire for the participant to fill in, comprising the background questionnaire and the topic familiarity rating questionnaire. This is to capture the basic information and exploratory search skills of the participants and topics of low levels of familiarity for later use in the tasks.

4. **Sequence of systems to be used** We flipped a coin to decide the order of topics and systems for the two tasks so that the noise brought by sequence was eliminated.

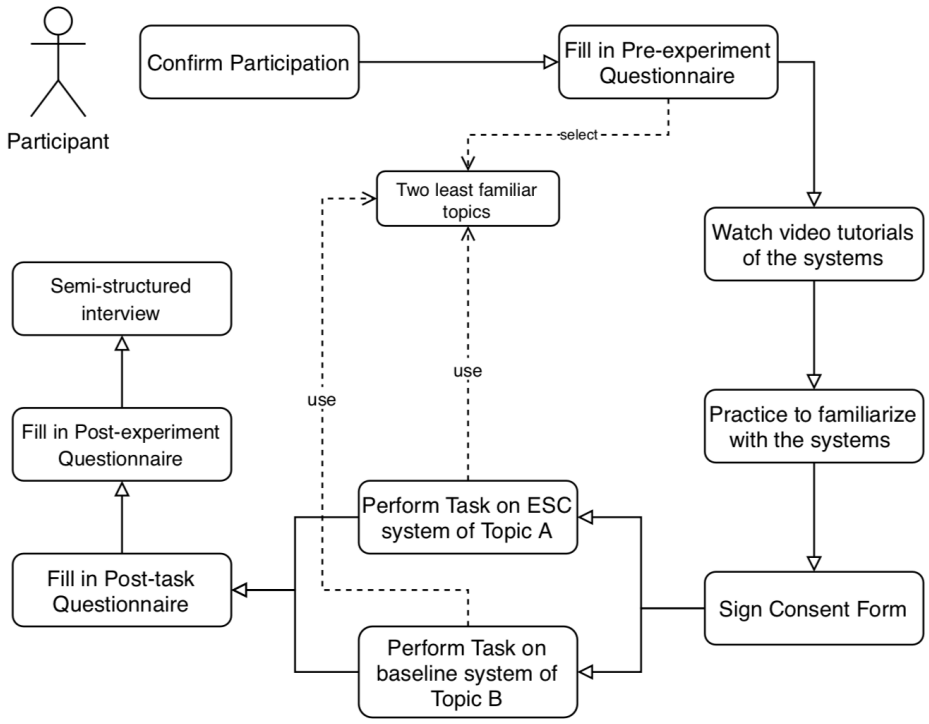

Figure 8: All steps of the study from the view of a participant

5. **Performing task** We told them that each task has a 30-minute time limit, and we put a clock on the table so the participants could keep track of the time themselves. To keep the whole process as natural as possible, we did not ask the participants to think aloud or keep on reminding them the time spent.

   - A written task description was presented to participants to read through until they understood it.
   - Participants typed in the query, clicked the search button and proceeded to the search result page. Meanwhile, their start time was logged.
   - Participants were instructed to write their answers in a text editor as well as a blank A4 paper for each task to take notes, if necessary.
   - While they were performing the task, we quietly observed their behaviour and made notes of anything special, which we may discuss with them during the interview.

6. **Post-task questionnaire** After each task, the participants completed the SUS[41] questionnaire with 10 questions (Table 6.2.1) and a modified version of the ResQue questionnaire[44] with 12 questions (Table 6.2.2).

7. **Post-experiment questionnaire** After both tasks were completed, the participants completed a post-experiment questionnaire, which consists of 19 questions (Table 1 about the utility of captions and the user perception of the two systems in exploratory search scenarios.

8. **Interview** We conducted a semi-structured interview about their search experience with the search systems. The aim of the interview was to better understand the user behaviour and explore the insights of participants regarding the two systems, especially the caption system.

Each study lasted approximately 90 minutes in total and we compensated each participant with a 15 euro bookshop gift card.

6 Results

We examined the results using both quantitative and qualitative feedback that we collected during the study, as well as the expert assessment scores of participants' answers to the tasks. All the participants completed all the tasks successfully. The 19 participants performed 38 search tasks in total (19 exploratory tasks $\times$ 2 systems = 38). No data points were excluded.

6.1 User Perception

To explore if and how the integration between the original system and ESC is from the perspective of user perception, we recorded the explicit preference and the ratings users gave in post-experiment questionnaire. We also analysed qualitative feedback the participants gave in the semi-structured interview.

6.1.1 User rating

During the post-experiment questionnaire, users stated that they preferred the system with captions to the baseline almost unanimously (18/19, $p = 7.6 \times 10^{-5}$, binomial test). Table 1 shows the mean score users gave for exploratory tasks (ratings are given on a 5 point Likert scale). A majority of users felt that the presence of captions reassured them that search results were relevant to their search goals (16/19,

$p = 0.004$). While we were concerned that the dynamic nature of the caption component would annoy users engaged in exploratory search, none of the participants thought the captions were distracting (0/19, $p = 3.8 \times 10^{-6}$), although a minority found the caption animation distracting (3/19, $p = 0.004$).

Average Score	p	Question
I = 18, B =1	7.6e-05	1. Which system did you prefer to use?
3.9	0.063	2. I found it easier to perform the search with captions
3.9	0.063	3. I found it easier to write the essay draft with captions
4.1	0.0007	4. The labels of the caption interface are clear
3.5	0.359	5. The bars of the caption interface are clear
3.2	1.0	6. The captions should be an optional function
4.1	0.0007	7. The captions enhanced my search session
4.1	0.0007	8. The captions were related to my search results
4.2	0.004	9. The captions reassured me that my search results were relevant to my search goals
3.8	0.063	10. The captions provided a good summary of my search results
3.6	1.0	11. The captions provided good followup queries
1.9	3.8e-06	12. The captions were distracting
2.4	0.004	13. The caption animations were distracting
3.4	0.647	14. The system's confidence in each caption was clearly indicated
4.0	0.0007	15. The system was better with the captions than without
2.0	3.8e-06	16. There too many captions shown

Table 1: Post-experiment score question averages with p-values. Question number 1 was binary. Questions 2 – 16 were scored on a 5-point Likert scale from 1 (disagree) to 5 (agree).

6.1.2 Qualitative feedback

After the main tasks, we conducted a semi-structured interview with each user. Interviews further proved that the caption system exceeds the baseline system for exploratory tasks.

Almost all users reported that they preferred the caption interface to the baseline. Ten participants stated the caption system to be significantly better when switching between two tasks: "when no captions, it takes me more time to look at the actual papers" [Participant 9], "it feels disappointing with the system without captions." [P4]. Fifteen participants looked at the captions frequently during their search, while 4 participants were not really paying attention and 2 participants confirmed that they never clicked on the captions. As for the perception of the caption utility, 12 participants gave affirmative answers: "it summarizes the key information well" [P2], "mainly for the relevance of the documents and get a sense of what's in there for predicting new things to search for" [P19]. Other 7 participants stated that the captions were not really useful for their search mainly because they considered the captions unreliable: "it's not hardcore useful, but interesting" [P7], "the first caption is not what I want to focus on" [P13]. When asked about preference over the interface of the caption system, 15 participants expressed their satisfaction with the existing style: "the current one is peaceful" [P7], "nice and minimus" [P16].

The most often mentioned benefits of the caption interface were:

- a concise summary of the search results
- suggestion for followup queries
- help with search context
- help to find new and interesting documents faster

Users also provided some suggestions for future improvements, e.g. some of the captions provided by the system were too generic to form a basis of a good followup query, lack of "back button" functionality and indicating which papers are most correlated with which captions.

6.2 Usability analysis

The caption interface obtained higher scores in both SUS and ResQue questionnaires than the baseline.

6.2.1 SUS Analysis

In SUS, the overall score was 76.8 for the caption system and 71.2 for the baseline, as shown in Table 6.2.1. While the difference was not significant ($p = 0.136$, Wilcoxon

signed rank test), this shows that the usability of the system did not suffer from the added functionality provided by the caption component.

I	B	p	Question
3.4	2.7	0.0049	1. I think that I would like to use this system frequently
1.7	1.9	0.107	2. I found the system unnecessarily complex
4.3	3.8	0.107	3. I thought the system was easy to use
1.6	**1.5**	0.726	4. I think that I would need the support of a technical person to be able to use this system
3.8	3.4	0.159	5. I found the various functions in this system were well integrated
1.9	1.9	0.705	6. I thought there was too much inconsistency in this system
4.3	4.2	1.0	7. I would imagine that most people would learn to use this system very quickly
1.9	2.1	0.129	8. I found the system very cumbersome to use
3.6	3.3	0.144	9. I felt very confident using the system
1.6	**1.5**	0.705	10. I needed to learn a lot of things before I could get going with this system

Table 2: SUS score averages for the caption system with caption (I) and baseline (B) systems with p-values from Wilcoxon signed rank test. Questions used a 5-point Likert scale from 1 (disagree) to 5 (agree). Best score in each row is bold; higher is better for odd numbered questions and lower is better for even numbered questions.

6.2.2 ResQue Analysis

In ResQue, the caption system significantly outperformed the baseline averaging 83.2 versus 67.8 ($p = 0.001$, Wilcoxon signed rank test), which shows that participants found the captions useful. As can be seen in Table 6.2.2, the caption system significantly exceeded the baseline system in helping them find the ideal documents.

The post-experiment questionnaire showed that 18 out of 19 users preferred the caption system to the baseline and agreed that the system was better with the captions. Generally, users agreed that the captions provided a good summary of the search results (14 users), helped them to perform the search (14), write the essay (14), provided a good summary of the search results (14) and enhanced the search session (17).

I	B	p	Question
4.0	3.5	0.01	1. The documents recommended to me matched what I was searching for
3.7	3.1	0.0139	2. The system helped me discover new documents
3.6	2.8	0.1305	3. The documents recommended to me are diverse
3.6	2.8	0.0164	4. The system helped me find the ideal documents
4.2	4.1	0.7054	5. I became familiar with the system very quickly
3.8	2.8	0.0189	6. I found it easy to notice if the search results were not correct any more
3.9	3.2	0.011	7. I felt confident to modify my query
3.6	3.2	0.0522	8. Using the system to find what I like is easy
3.7	3.7	0.7192	9. I found it easy to re-find documents I had been recommended before
3.7	3.1	0.0079	10. The system gave me good suggestions
3.8	3.5	0.07	11. The system made me confident about the documents I bookmarked
3.6	3.2	0.0374	12. Overall, I am satisfied with the system

Table 3: ResQue score averages for caption (I) and baseline (B) systems with p-values from Wilcoxon signed rank test. Questions used a 5-point Likert scale from 1 (disagree) to 5 (agree). Best score in each row is boldface; higher is better.

6.3 User behaviour

To investigate whether there is a notable behavioural difference between the baseline and the caption system, and to find answer of RQ2, we analysed user behaviours in terms of: (1) task performance – expert assessment score, (2) the number of queries users issued – referred to as *interactions*, (3) the number of documents users inspected and bookmarked.

6.3.1 Task performance

Task performance was assessed in a blind manner by an expert assessor based on the written answers submitted by participants. The ratings were done on 5-point scale from 1 (bad) to 5 (good). The average task performance was 2.95 with the baseline and 3.37 with the caption interface ($p = 0.035$, Wilcoxon signed rank test).

6.3.2 Interactions

To understand the difference in average task performance between the two systems, we investigated how the presence of captions affected how users interacted with the system.

When using the caption system, users issued more queries per session (8.2 queries per session compared to 3.7 with the baseline ($p = 0.0006$, Wilcoxon signed rank test)), they inspected fewer documents per query (7.8 documents per query compared to 18.6 with the baseline ($p = 0.004$, Wilcoxon signed rank test)). But users were exposed to more documents in total when using the caption system (55.3 documents compared to 38.7 with the baseline ($p = 0.02$, Wilcoxon signed rank test)). There was no significant difference in the number of bookmarked documents, with 9.4 and 9.2 bookmarks for caption and baseline systems, respectively. This is almost certainly due to the fact that users were instructed to bookmark at least 10 documents as part of the essay writing task.

7 Conclusion and future work

This thesis aims to investigate the usefulness of ESC in terms of user experience, user behaviour and to identify an effective integration strategy for incorporating keyword summaries into an existing IR system.

We started with introducing information retrieval, and identifying related works and introduced the basics in the area of exploratory search: how summarised interface and interactive recommendation interface supports user understanding and system performance, what have been studied to assist exploration and user engagement in the search process. Next, we presented ESC, a method for generating captions from ranked search results combining semantic and lexical features. We then demonstrated our system design method for incorporating ESC. In this process, we first identified a few design principles for the integrated system; secondly, we went into interface design and changed from horizontal bar design to sidebar design after pilot test. Moreover, we detailed the importance of experimental approach and clarified the indicators to be used for information search behaviours. Finally, to evaluate the utility of ESC, we designed and conducted a user study in the scenario of scientific literature search. The results of the user study suggest the ESC system offers significant support in exploring articles and was positively reviewed by participants.

This work focuses on the academic scientific literature domain.

In the user study, participants almost unanimously preferred the retrieval system that incorporated ESC. We saw that the presence of captions dramatically impacts user behaviour: users issue more queries, investigate fewer documents per query, but see more documents overall. On average, participants reformulated their search queries over 2× more frequently when ESC was present and wrote higher quality essays incorporating more relevant details than users of the baseline system. In line with our expectations, users were reassured by the captions that search results matched their search intent and explored the data set more confidently than with the baseline system.

All three research questions are answered adequately in our thesis work. Based on a combination of quantitative and qualitative feedback collected in the user study, it can be concluded that:

- For RQ1, ESC reassures users and enhances their trust in the search results offered by the search engine and improves users' ability to explore articles according to their information need. Sixteen out of nineteen participants thought the captions reassured them about the relevance of the research results to their search goals, while no participant considered the captions distracting. In addition, the caption system significantly outperformed the baseline averaging 83.2 versus 67.8 according to ResQue results. One participant[P4] said: "it feels disappointing with the system without captions (baseline system)".

- For RQ2, Behavioural differences between users of ESC and the baseline system are noticeable. On a 5-point Likert scale, the average task performance was 3.37 with the caption system and 2.95 with the baseline. When using the caption system, users issued 8.2 queries per session compared to 3.7 with the baseline.

- For RQ3, the results also indicate that our design decisions in integration, including animation, layout and type of visualisation, facilitate a smooth user experience and helps intuitive user perception. Almost all users reported that they are in favour of the caption interface other than the baseline. Fifteen participants expressed their satisfaction with the style of caption system: " the current one is peaceful" [P7], " nice and minimus" [P16].

In future work, we want to understand whether similar approaches could be used to understand long term search goals; whether ESC could also be used in an effective

way with system other than academy literature search; whether the interface design strategy could be applicable to similar integration of IR system and add-on function.

Acknowledgements

I would like to express my deepest appreciation to my Supervisor Prof. Dorota Glowacka and Dr. Alan J. Medlar for their relentless support, valuable suggestion and extensive patience through the completion of this thesis.

References

1 G. Marchionini, "Exploratory search: from finding to understanding," *Communications of the ACM*, vol. 49, no. 4, pp. 41–46, 2006.

2 R. W. White and R. A. Roth, "Exploratory search: Beyond the query-response paradigm," *Synthesis lectures on information concepts, retrieval, and services*, vol. 1, no. 1, pp. 1–98, 2009.

3 K. Athukorala, D. Głowacka, G. Jacucci, A. Oulasvirta, and J. Vreeken, "Is exploratory search different? a comparison of information search behavior for exploratory and lookup tasks," *Journal of the Association for Information Science and Technology*, vol. 67, no. 11, pp. 2635–2651, 2016.

4 A. Medlar and D. Glowacka, "How consistent is relevance feedback in exploratory search?," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1615–1618, ACM, 2018.

5 M. Mitra, A. Singhal, and C. Buckley, "Improving automatic query expansion," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 206–214, ACM, 1998.

6 K. Verbert, D. Parra, P. Brusilovsky, and E. Duval, "Visualizing recommendations to support exploration, transparency and controllability," in *Proceedings of the 2013 international conference on Intelligent user interfaces*, pp. 351–362, 2013.

7 S. E. Robertson, "The methodology of information retrieval experiment," *Information retrieval experiment*, vol. 1, pp. 9–31, 1981.

8 N. J. Belkin, "Cognitive models and information transfer," *Social Science Information Studies*, vol. 4, no. 2-3, pp. 111–129, 1984.

9 N. J. Belkin and W. B. Croft, "Information filtering and information retrieval: Two sides of the same coin?," *Communications of the ACM*, vol. 35, no. 12, pp. 29–38, 1992.

10 G. Salton, *Automatic information organization and retrieval.* McGraw Hill Text, 1968.

11 A. Singhal *et al.*, "Modern information retrieval: A brief overview," *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 35–43, 2001.

12 R. W. White, B. Kules, S. M. Drucker, *et al.*, "Supporting exploratory search, introduction, special issue, communications of the acm," *Communications of the ACM*, vol. 49, no. 4, pp. 36–39, 2006.

13 A. Medlar, K. Ilves, P. Wang, W. Buntine, and D. Glowacka, "PULP: A system for exploratory search of scientific literature," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 1133–1136, ACM, 2016.

14 J. Kim, "Describing and predicting information-seeking behavior on the web," *Journal of the american society for information science and technology*, vol. 60, no. 4, pp. 679–693, 2009.

15 B. M. Wildemuth and L. Freund, "Assigning search tasks designed to elicit exploratory search behaviors," in *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, pp. 1–10, 2012.

16 G. Marchionini, "Information seeking in full-text end-user-oriented search systems: The roles of domain and search expertise," *Library & information science research*, vol. 15, no. 1, pp. 35–69, 1993.

17 D. Kelly and X. Fu, "Elicitation of term relevance feedback: an investigation of term source and context," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 453–460, 2006.

18 Z.-H. Zhou, K.-J. Chen, and H.-B. Dai, "Enhancing relevance feedback in image retrieval using unlabeled data," *ACM Transactions on Information Systems (TOIS)*, vol. 24, no. 2, pp. 219–244, 2006.

19 V. Lavrenko and W. B. Croft, "Relevance-based language models," in *ACM SIGIR Forum*, vol. 51, pp. 260–267, ACM New York, NY, USA, 2017.

20 N. J. Belkin, C. Cool, D. Kelly, S.-J. Lin, S. Park, J. Perez-Carballo, and C. Sikora, "Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval," *Information Processing & Management*, vol. 37, no. 3, pp. 403–434, 2001.

21 D. Tripathi, A. Medlar, and D. Glowacka, "How relevance feedback is framed affects user experience, but not behaviour," in *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pp. 307–311, 2019.

22 J. Matejka, T. Grossman, and G. Fitzmaurice, "Citeology: visualizing paper genealogy," in *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pp. 181–190, ACM, 2012.

23 P. Daee, J. Pyykkö, D. Glowacka, and S. Kaski, "Interactive intent modeling from multiple feedback domains," in *Proceedings of the 21st international conference on intelligent user interfaces*, pp. 71–75, 2016.

24 K. Athukorala, A. Medlar, A. Oulasvirta, G. Jacucci, and D. Glowacka, "Beyond relevance: Adapting exploration/exploitation in information retrieval," in *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pp. 359–369, 2016.

25 D. Glowacka, T. Ruotsalo, K. Konuyshkova, S. Kaski, G. Jacucci, *et al.*, "Directing exploratory search: Reinforcement learning from user interactions with keywords," in *Proceedings of the 2013 international conference on Intelligent user interfaces*, pp. 117–128, ACM, 2013.

26 T. Ruotsalo, J. Peltonen, M. J. Eugster, D. Głowacka, P. Floréen, P. Myllymäki, G. Jacucci, and S. Kaski, "Interactive intent modeling for exploratory search," *ACM Transactions on Information Systems (TOIS)*, vol. 36, no. 4, pp. 1–46, 2018.

27 J. Zhao, C. Collins, F. Chevalier, and R. Balakrishnan, "Interactive exploration of implicit and explicit relations in faceted datasets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2080–2089, 2013.

28 B. Zheng, W. Zhang, and X. F. B. Feng, "A survey of faceted search," *Journal of Web engineering*, vol. 12, no. 1&2, pp. 041–064, 2013.

29 B. Kules, R. Capra, M. Banta, and T. Sierra, "What do exploratory searchers look at in a faceted search interface?," in *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pp. 313–322, 2009.

30 M. Hearst, "Design recommendations for hierarchical faceted search interfaces," in *ACM SIGIR workshop on faceted search*, pp. 1–5, Seattle, WA, 2006.

31 K.-P. Yee, K. Swearingen, K. Li, and M. Hearst, "Faceted metadata for image search and browsing," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 401–408, 2003.

32 M. Käki, "Findex: search result categories help users when document ranking fails," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 131–140, 2005.

33 W. Kules, M. L. Wilson, B. Shneiderman, *et al.*, "From keyword search to exploration: How result visualization aids discovery on the web," 2008.

34 D. H. Chau, A. Kittur, J. I. Hong, and C. Faloutsos, "Apolo: making sense of large network data by combining rich user interaction and machine learning," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 167–176, 2011.

35 A. Kangasrääsiö, D. Glowacka, and S. Kaski, "Improving controllability and predictability of interactive recommendation interfaces for exploratory search," in *Proceedings of the 20th international conference on intelligent user interfaces*, pp. 247–251, 2015.

36 P. Pirolli and S. Card, "Information foraging.," *Psychological review*, vol. 106, no. 4, p. 643, 1999.

37 K. Athukorala, A. Oulasvirta, D. Głowacka, J. Vreeken, and G. Jacucci, "Narrow or broad?: Estimating subjective specificity in exploratory search," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 819–828, ACM, 2014.

38 A. Hassan, R. W. White, S. T. Dumais, and Y.-M. Wang, "Struggling or exploring?: disambiguating long search sessions," in *Proceedings of the 7th ACM international conference on Web search and data mining*, pp. 53–62, ACM, 2014.

39 J. Mao, Y. Liu, N. Kando, M. Zhang, and S. Ma, "How does domain expertise affect usersâ search interaction and outcome in exploratory search?," *ACM Transactions on Information Systems (TOIS)*, vol. 36, no. 4, p. 42, 2018.

40 D. Choi, *A Study of Information Seeking Behavior: Investigating Exploratory Behavior in Physical & Online Spaces*. Rutgers The State University of New Jersey-New Brunswick, 2017.

41 J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

42 J. Brooke, "Sus: a retrospective," *Journal of usability studies*, vol. 8, no. 2, pp. 29–40, 2013.

43 A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the system usability scale," *Intl. Journal of Human–Computer Interaction*, vol. 24, no. 6, pp. 574–594, 2008.

44 P. Pu, L. Chen, and R. Hu, "A user-centric evaluation framework for recommender systems," in *Proceedings of the fifth ACM conference on Recommender systems*, pp. 157–164, 2011.

45 P. Pu, L. Chen, and R. Hu, "Evaluating recommender systems from the userâs perspective: survey of the state of the art," *User Modeling and User-Adapted Interaction*, vol. 22, no. 4-5, pp. 317–355, 2012.

46 T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

47 G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

48 D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

49 B. J. Jansen, D. L. Booth, and A. Spink, "Determining the informational, navigational, and transactional intent of web queries," *Information Processing & Management*, vol. 44, no. 3, pp. 1251–1266, 2008.

50 C. Hölscher and G. Strube, "Web search behavior of internet experts and newbies," *Computer networks*, vol. 33, no. 1-6, pp. 337–346, 2000.

51 A. Aula and K. Nordhausen, "Modeling successful performance in web searching," *Journal of the american society for information science and technology*, vol. 57, no. 12, pp. 1678–1693, 2006.

52 R. A. Palmquist and K.-S. Kim, "Cognitive style and on-line database search experience as predictors of web search performance," *Journal of the american society for information science*, vol. 51, no. 6, pp. 558–566, 2000.

53 J. Liu and N. J. Belkin, "Personalizing information retrieval for multi-session tasks: The roles of task stage and task type," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 26–33, 2010.

54 K. Athukorala, E. Hoggan, A. Lehtiö, T. Ruotsalo, and G. Jacucci, "Information-seeking behaviors of computer scientists: Challenges for electronic literature search tools," *Proceedings of the American Society for Information Science and Technology*, vol. 50, no. 1, pp. 1–11, 2013.

55 S. Chowdhury, F. Gibb, and M. Landoni, "Uncertainty in information seeking and retrieval: A study in an academic environment," *Information Processing & Management*, vol. 47, no. 2, pp. 157–175, 2011.

56 N. J. Belkin, "Some (what) grand challenges for information retrieval," in *ACM SIGIR Forum*, vol. 42, pp. 47–54, ACM New York, NY, USA, 2008.

57 K. S. Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: development and comparative experiments: Part 2," *Information processing & management*, vol. 36, no. 6, pp. 809–840, 2000.

58 A. Singhal *et al.*, "Modern information retrieval: A brief overview," *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 35–43, 2001.

Appendix 1. Script of interview

1. What do you like about the first system? How about the second?

p01: with caption, interface of the system is good p02: with caption. The captions are great. p03: no preference of the two system p04: with caption. The captions provide an instant analysis of the documents, they are vague but enough to go to a new information. The system without captions is bad, because there's no indication about the relevance(if the first document is the most related one) p05: with caption. It gives an idea of what the content is like in general. P06: no preference. p07: captions made me want to go forward, I really like the topic itself using the second system(without captions) P08: The system without captions: difficult to use, stucks a lot with it, can't go back, good to refine. The system with captions: much better, gives a direction to explore into information space p09: the system with caption makes it faster to see the information as a whole p10: it's easier to find alternatives using captions p11: nothing p12: the captions are related p13: I think the second topic is easier itself p14: the only benefit of the system is to check whether I got things wrong p15: the results are highly relevant. The system with captions provided more diverse results, especially about application p16: the extended version is better. When it works, it's giving you a very quick summary. If it's reliable. Sometimes it didn't seem perfect. P17: the first system is intuitive and easy to use. The second system is more useful, at some point. p18: the captions can help support problem solving. p19: cool to see the captions. But I don't know if it was just for the abstract or for the full paper. It would be nice if it was something more than the abstract. I don't know if it can provide some new context, some amazing features.

2. What do you dislike about the system with caption?

p01: no p02: no p03: it can not go back when I launched a new search by clickinging the caption p04: no p05: it can't go back, gets confused when goes to another search p06: no p07: Where the captions come from, what do they mean, clarify in a specific way p08: the speed of updating captions should be slower; more support and recommendation on the captions themselves would be good p09: captions are repeating, not related to the original query p10: no p11: captions are not very related p12: no keywords, no cited function p13: the first caption is not what I want to focus on(not useful) p14: the repetitiveness. The captions as a whole is not really a good summary, but better than without. p15: the caption works as summary, it's not a novel way to do, p16: maybe one thing that would be nice is,

if you highlight over the captions, it's good if it somehow provided a feedback of where it's deriving those captions from. For example, if you just highlight over the EEG thing, it shows like this article have/does, something like that. Show sort of a feedback of where that information is coming from. p17: The bars of the captions are tricky to understand when you scroll, the bars start to changing. But then I realized how it works. Don't know if the captions have an effect on the documents displayed p18: can not go back, the captions are sometimes too general, not concrete enough p19: I don't dislike it that much, but I don't want to be restricted to arXiv only.

3. How do you feel when you switch from one system to the other?

p01: the one with caption is better p02: nothing p03: no p04: disappointing with the system without captions. p05: way better p06: no p07: no p08: intelligent p09: when no captions, it takes me more time to look at the actual papers p10: downward p11: almost the same p12: A bit uncomfortable p13: nothing p14: nothing p15: nothing p16: it took just a second to get used to that screen. p17: it feels a little easier to for the task. p18: nothing p19: a little lost

5. How frequently do you look at the captions? About when? (e.g. after each scroll)

p01: After each scroll p02: Every 3,4 scroll p03: After each scroll p04: all the time p05: after each scroll p06: Glimpse at every scroll p07: a lot p08: Quite often. when get stuck/need help. p09: scanning to see if the titles are included in the captions p10: after scroll a few times p11: sometimes, not really paying attention p12: in the beginning very often. When the first three documents is not related, I explored a lot with the captions p13: not much. When the documents displayed didn't meet my expectation p14: sometimes. The width of the bars are similar, so I was just looking. p15: when the changes are big, I would take a look p16: It depends. If I'm just searching for information, I would just bounce back and forth between the two, fairly frequently, may every couple of seconds as I'm scrolling. When I notice there's a lot of good staff, then I would just be reading the abstracts. p17: every time when I want to dig into something, when I want to go into something more specific p18: quite a lot during the first few search p19: as I scroll down. When I start feeling lost

6. Did you click on the captions? When?

p01: Yes, when I found it interesting p02: Yes, when the caption didn't appear in any documents I saw p03: Yes, when I want to explore them p04: yes p05: yes.

But when I clicked on the captions, it's not connected to the original query which is confusing. So I didn't clicked it in the latter part. p06: no p07: yes, to see if it can be better, for browsing p08: yes, when relevant p09: yes, to see what's there p10: yes, when interesting p11: in the very beginning, out of curiosity p12: yes, to find new things p13: no p14: yes. When it's related to cognitive p15: yes, when the caption seems to be a synonym p16: yes, when there was a topic that I was hoping to explore more in depth, just to see what the results were. But I was not quite sure how they(the clicks) affected the results, it seemed something changed, but it didn't look like changed my search query. So I then didn't know if there was a way to go back. p17: yes, when interesting. For me, sometimes the captions feels like a summary of something, then I start looking at the captions, and when I found something relevant, I click on it. p18: yes for the first few search, it's hard to perform a second search so I dare not click on it p19: yes, a couple of times, not too much. Because I think the suggestion wasn't amazing

7. Did you find the captions useful for your search? About when?

p01: Yes, it's like an abstract of the abstract. Although sometimes the captions were repetitive p02: Yes, it summarizes the key information well p03: Not really, because I know so little about the topic, I'm not sure if the captions are correct and strongly related to the topic p04: yes, after 5,6 hits, when I'm about to grasp the content of the topic. p05: not really p06: no, because I find a really good review, so I basically focus on the review p07: it's not hardcore useful, but interesting p08: yes, it feels less tedious p09: quite soon. Because the time limit is short and I need to get things fast p10: very useful. After I can't find good articles any more p11: no p12: yes, to reduce the time spent. p13: maybe p14: not really p15: yes, but when I click on a caption, the new search is incomplete due to this caption. I have to manually add my original query. When the results are not interesting or have nothing new, I would click on captions. p16: yes. When they were longer. I think the shorter captions are just keywords, but longer ones to me is a topic. If it's more like a topic model thing, that can sometimes help. But sometimes they can be very very vague. p17: yes at some point, but sometimes it can be repetitive and confusing. p18: yes, when I'm not very familiar with the topic p19: yes, mainly for the relevance of the documents and get a sense of what's in there for predicting new things to search for

8. What preference do you have regarding the interface?

p01: The animation could be better, prefer the up-and-down style. (I explained the shortcomings of such animation) It should be attracting to users when big changes

happened, to remind them what has happened. p02: no p03: no p04: no p05: no p06: no p07: the current one is peaceful, I like it p08: no p09: no p10: no