Profile-Based Summarisation for Web Site Navigation

Azhar Alhindi\textsuperscript{1}, University of Essex
Udo Kruschwitz\textsuperscript{2}, University of Essex
Chris Fox\textsuperscript{3}, University of Essex
M-Dyaa Albakour\textsuperscript{4}, University of Glasgow

Information systems that utilise contextual information have the potential of helping a user identify relevant information more quickly and more accurately than systems that work the same for all users and contexts. Contextual information comes in a variety of flavours, often derived from records of past interactions between a user and the information system. It can be individual or group based. We are focusing on the latter, harnessing the search behaviour of cohorts of users, turning it into a domain model which can then be used to assist other users of the same cohort. More specifically, we aim to explore how such a domain model is best utilised for profile-biased summarisation of documents in a navigation scenario where such summaries can be displayed as hover text as a user moves the mouse over a link. The main motivation is to help a user find relevant document(s) more quickly. Given the fact that the web in general has been studied extensively already, we focus our attention on web sites and similar document collections. Such collections can be notoriously difficult to search or explore. The process of acquiring the domain model is not a research interest here; we simply adopt a biologically inspired method that resembles the idea of ant colony optimisation. This has been shown to work well in a variety of application areas. The model can be built in a continuous learning cycle that exploits search patterns as recorded in typical query log files. Our research explores different summarisation techniques, some of which use the domain model and some that do not. We perform task-based evaluations of these different techniques — and hence of the impact of the domain model and profile-biased summarisation — in the context of web-site navigation.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Profiling, Summarisation, Task-Based Evaluation, User Studies

Additional Key Words and Phrases: Term association networks, group profiling, single-document summarisation (SDS), multi-document summarisation (MDS), browsing, navigation, log analysis

ACM Reference Format:
Azhar Alhindi, Udo Kruschwitz, Chris Fox, and M-Dyaa Albakour 2014. Profile-Based Summarisation for Web Site Navigation. ACM Trans. Inf. Syst. 0, 0, Article 0 (August 2014), 40 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

\textsuperscript{1}ahalhi@essex.ac.uk
\textsuperscript{2}udo@essex.ac.uk
\textsuperscript{3}foxcj@essex.ac.uk
\textsuperscript{4}m-dyaa.albakour@glasgow.ac.uk

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2014 ACM 1046-8188/2014/08-ART0 $15.00
DOI: http://dx.doi.org/10.1145/0000000.0000000

ACM Transactions on Information Systems, Vol. 0, No. 0, Article 0, Publication date: August 2014.
1. MOTIVATION

“The easier access to information becomes, the greater become our expectations for ubiquitous access in all kinds of situations.” [Marchionini and White 2009]

The continuous growth of document collections on local web sites makes it desirable to develop techniques that assist users not just in the search process but also in navigating the collection. It can be surprisingly difficult to track down a specific document or a specific piece of information on an intranet or a university web site even if the information is there, it is just difficult to find. There are a number of reasons for this problem, one is that such collections are different to the web in many respects [Hawking 2011]. For example, there is much less redundancy than on the web in general; only a single document might exist in a large collection which satisfies a specific user need. Another reason is the mismatch of terminology between what a searcher is after and what the documents are about, sometimes referred to as the “vocabulary gap” [Smyth 2007]. Using a university site example, a new first-year student might not actually know where to register, how to find the accommodation office, how to obtain a parking permit, and whether to enrol for a “module”, a “course” or a “unit”. In such cases, it is possible that the local search engine will not be of much help either [Hawking 2011].

Harnessing contextual information offers great opportunities to address the above-mentioned problems. Contextual search and recommendation refers to a diverse set of techniques that are all aimed at moving away from a one-size-fits-all approach and making a system more effective by incorporating contextual information derived from a wide range of variables, such as content, geographical, interaction and social variables [Melucci 2012] or simply the users’ search histories [Smyth et al. 2005]. While many contextual systems attempt to personalise a system for individual users, contextual approaches can also be group based; contextualisation should not be equated to personalisation [Ruthven 2011]. Group-based information appears to be a promising route for a community of users with common concerns. Such communities are formed of individuals — e.g. employees of a company or members of a university — that, over time, collectively acquire knowledge about a resource such as a local web site. The idea is to tap in to this knowledge, and facilitate the sharing of search and navigation experiences among community members [Smyth 2007]. We can therefore characterise our context fairly generally as the environment in which a cohort of users operates in. This bears some resemblance with the idea of “trait-based groups” as people who “may be highly likely to repeat or augment tasks already accomplished by other group members, have interests in the same queries and results as other group members.” [Teevan et al. 2009]. The idea is that learning from one user should benefit future users with similar information needs, an idea that our work shares with other approaches of assisting users in navigating a collection [Kantor et al. 2000; Wexelblat and Maes 1999, for example].

All this suggests that utilising the accumulated search histories could be beneficial. But it does not tell us (i) how to build usable knowledge structures, or (ii) how to apply them. To address the first point, we note that query logs have emerged as a very valuable resource that can be mined to derive useful knowledge, such as query sug-

---

5 When we talk about “local web sites” we like to refer to document collections that represent the intranet or the web presence of a university, a company or some other organisation. To simplify matters and to make the approaches more generalizable we ignore the internal organisational structures that are common features of such collections. This also allows us to treat other collections, such as digital libraries, in the same way.

6 We will use the terms “cohort”, “community”, “group” and “population” interchangeably. In addition to that we also use “domain model”, “model” and “profile” as synonymous terms in this paper. Finally, we make the simplifying assumption of referring to “browsing” and “navigation” as the same thing; in fact any reference to “search” in our studies is to be interpreted as searching for information in a navigation context.
gestions. In fact, search trails and associated click-through information are used in many modern search engines to help in the search process [White and Huang 2010]. Our approach utilises a profile acquired from log data that contains information relating to previous search trails involving a local search engine. We do not propose a new paradigm for the profile. Instead, we adopt a state-of-the-art approach from the literature which can easily be applied to typical query logs. Many methods can be applied to build different types of profiles. Here we adopt a term-association network in which the nodes represent queries and links between queries reflect (unspecified) relations between these nodes. We acquire our profile by applying an ant colony optimisation (ACO) analogy. This offers the additional benefit over more static approaches by embodying a temporal context; it learns from user interactions with an information system but is also able to forget. For example, a concept “timetable” in a model derived from a university web site will, for example, reflect a stronger association to a concept “teaching timetable” at the start of the academic year and “exam timetable” towards the end of the year. In summary, we turn the search history of a local web site into a profile that represents the search interests and interactions of the population of searchers. This is different to individual user profiles.

That leaves us with the question as to how to apply the model/profile. Interactive search support has attracted a lot of attention, e.g. [White et al. 2002; Dumais et al. 2001; Paek et al. 2004], and cohort modeling has shown to be effective for Web search [Yan et al. 2014], but not much has been reported to support users in a navigation context. Our aim is to explore the benefit of profile-based summarisation for navigation, providing “tool tips” that give web-site users a summary of a document as the user hovers the mouse over a link. This allows them to assess whether it is worth following the link or not (as illustrated in Figure 1). One of the main objectives is to cut down the number of steps and the time taken to get from the user’s entry page to the desired document.

Summarisation appears relevant to the navigation problem as it helps present salient information to a user by condensing a document’s content and extracting the
most relevant facts or topics included in them [Lloret and Palomar 2012]. Query-biased summarisation has been shown to be highly effective in a search context [White et al. 2003, for example]. We hypothesise that it is also useful in web-site navigation. Both single-document summarisation (SDS) and multi-document summarisation (MDS) have been investigated extensively and independently within the natural language processing (NLP) and information retrieval (IR) communities [Wan 2010]. SDS aims to produce a concise and fluent summary of a single document, whereas MDS summarises multiple related documents. The two tasks are very closely related in both task definition and solution method. The major difference between the tasks of SDS and MDS is the greater amount of redundancy when we start with multi-documents [Jurafsky and Martin 2009, Ch. 23, p. 831]. Both approaches are useful and commonly applied [Nenkova and McKeown 2011]. For example, given a cluster of news articles, a multi-document summary can be used to help users understand the whole cluster, a single summary for each article can be used to help users to know the content of the specific article. We are interested in both SDS and MDS.

The research questions we aim to answer with our work are as follows:

1. Can web site navigation benefit from the automated summarisation of results?
2. Will a domain model/profile capturing the search behaviour of a group of users be beneficial for the summarisation process?
3. Will such methods result in measurable (quantifiable) benefits such as shorter sessions, fewer interactions etc?

In this paper we report on a number of experiments and studies aimed at addressing our research questions. The core of the experimental work consists of task-based evaluations in a web-site-navigation context. We use a specific university web site to conduct our studies, but would argue that the methods are applicable to a wide range of web sites and intranets. The caveat of such a study is that it is limited to a single web site and the findings may or may not be transferable to other document collections, see also Kruschwitz et al. [2013]. Despite this limitation, we argue that our results could provide insights and serve as a baseline for future studies on different web sites. The major bottleneck in conducting research into using any form of query logs is the difficulty in getting hold of realistic and large-scale log data, both of which we managed to address in this study.

The structure of the paper is as follows. We will start by discussing some related work in Section 2. We then describe the construction of the log-based profile in Section 3 and give a brief description of our data sets in Section 4. We outline how we apply the acquired model to construct profile-based summaries in Section 5. An initial pilot study, to assess the potential that profile-based summarisation might have for web-site navigation, is presented in Section 6, followed by the two main task-based evaluations in Section 7. Section 8 concludes.

2. RELATED WORK

Information retrieval is a daily activity for most people. When searchers have well-defined needs in mind, the retrieval of documents following a one-shot query might be sufficient, but when they are seeking information for complex mental activities — such as for decision making or learning — retrieval is necessary but may not be sufficient [Marchionini and White 2009]. In this case, tools and support services that assist users in managing, analysing, and sharing sets of retrieved information may be helpful. The problems addressed in this paper touch on a number of different research areas. We will discuss each of them briefly, in turn.
2.1. Search and Navigation

Browsing and searching are the two predominant interaction modes for locating information. These can be characterised as searching by navigation and searching by query [Jul and Furnas 1997; Furnas 1997].

Searching by query is accomplished by submitting a search query — typically a list of keywords or terms — to a search engine. The search engine then returns ranked links to pages that match the query. Search-by-query might return inappropriate results due to, for example, polysemy and synonymy but it is very common. It is able to identify the pages that contain relevant terms quickly.

In contrast, searching by navigation, or “browsing”, is useful when knowledge of the required information is more vague, and when the user is unable to formulate a conventional search query due to a lack of knowledge about the appropriate terminology to use, or when a user is simply exploring a web site or information resource. In this scenario, providing contextual support may help bridge the “vocabulary gap”, and reduce the frustration of navigating web sites that can arise when using the static, general-purpose link structure that has been set in place by the web site administrators and content editors [Karim et al. 2009]; and it might also help those users that have no specific knowledge of the content of a collection [Joachims et al. 1997]. We also note that use of contextual information for navigation has not received the same attention as contextual information for search, which has been explored extensively (at least in the case of general web search). We do however know that employing users’ interaction histories in a browsing context has the potential to significantly reduce a user’s effort in finding the right information [Wexelblat and Maes 1999].

With search-by-query on the web, typically the user’s query is entered into a text box. The search engine then produces one or more pages of ranked links to documents. Each link is typically accompanied by a query-biased summary of the link’s destination. The results may include links to different kinds of information, such as documents in various formats, videos, images, news articles and encyclopedia entries.

A variety of improvements for web search have been suggested and adopted. These include, for example, query-term highlighting within the results [Wilson 2011], and query suggestions based on previous user interactions [Jones et al. 2006]. These improvements can lead to a more interactive search process. Olston and Chi [2003] build an intelligent system which combines the strengths of searching and browsing in a single interface. This guides users towards search results by highlighting relevant hyperlinks on the web pages that they are browsing. Another approach to combine the two interaction modes is proposed by Freyne et al. [2007] in an attempt to harness and harvest community wisdom by incorporating social search and social browsing. White et al. [2007] enhance web search by suggesting links to web sites frequently visited by other users with similar information needs — in addition to the regular search results. This exploits the searching and browsing behaviour of previous users. Our work inherits ideas from these approaches. But instead of proposing links or queries, we aim to help a web-site user by applying text summarisation to hyperlinked documents in order to assist their navigation.

In order to fully satisfy users’ search needs, a search engine should take the user past the result list, and aid them as they browse towards their information goal [Coyle and Smyth 2007]. The act of browsing can assist users in formulating a more focused goal; quite often users’ information needs may be ill-defined at query time. The SearchGuide browsing assistant [Coyle and Smyth 2007] helps users both in navigating to pages that are linked from the selected page, and in browsing through the contents of a selected result. The content of the “anchor-text” associated with links, and the surrounding text, often provides insufficient information for users to make reliable deci-
sions about whether to open a linked page. But users are often entirely dependent on this information when browsing hypertextual documents returned by a regular information retrieval search engine [Jones and Li 2008]. As a consequence, users may be obliged to follow a link in order to determine whether it leads to useful information, and then return to the previous page when it does not. With this in mind, the question is whether there is some scope for providing assistance with browsing by summarising the documents that can be found by following a given hyperlink. There may be a trade-off between offering such help and the additional reading effort by a user. It is also unclear how long such a summary needs to be. But by way of support, the use of automatically generated summaries has been shown to allow users gauge document relevance more effectively than by using a standard ranked title/snippet approach in the case of web search [White et al. 2003]. Overall search time can be significantly reduced, as short summaries can be read more quickly than full pages of text [Chen and Dumais 2000].

There are many different ways of improving general web search and navigation. Carbonaro [2010] presents a summarisation process through exploring concept-based search retrieval for enabling user-friendly and intelligent content exploration. Jones and Li [2008] provide topical feedback for link selection in hypertext browsing and a term cloud preview of the contents of each linked page. However, there appears to have been little progress in making navigating a web site more adaptive to users’ needs (without requiring the users to explicitly express what they are after [Joachims et al. 1997, for example]). Web sites and intranets can be difficult to navigate, with a static and possibly idiosyncratic organisation [Karim et al. 2009; Berendt and Spiliopoulou 2000]. In addition, search can be difficult on such sites [Hawking 2011].

A common approach to add assistance to a web site is to use an overlay window or hover text, essentially adding a “layer” on top of an existing site. This can be used for presenting search results [White et al. 2002; Dumais et al. 2001, for example], or for navigation by introducing links and suggestions to commonly visited pages, taking advantage of the collective search and navigation effort of other users [Karim et al. 2009; Saad and Kruschwitz 2011]. We will adopt the idea of hover text by creating a pop-up “tool-tip” that contains a summary of the target of a link whenever users hover their mouse cursor over that link. The summary is generated from the document, or documents, that can be reached by following the link. It uses information such as the title of the document, and relevant nodes in the profile. The actual content of the web site itself is left unchanged. 7

2.2. Text Summarisation

Text summarisation has been an active research area for more than 50 years [Luhn 1958; Carbonaro 2010]. The increasing availability of large volumes of digital information calls for automatic text summarisation. The aim is to provide an extract or precise of one or more texts that contains the most important or salient points [Nenkova and McKeown 2011]. Hovy and Lin [1998] divide summarisation approaches into two main groups based on their output type: abstractive summaries and extractive summaries. Here, we focus on extractive summarisation. 8

As we have already mentioned, a summary can be generated for multiple, related source documents — for example, a cluster of news stories on the same topic — and

7There may be legal and moral issues when wrapping other people’s content inside an interface that, in some sense, changes the “content” and appearance of a web site. But we believe the beneficiaries of our work will normally be web site managers, as with other work that aims at supporting the designers and owners of a web site [Chi et al. 2000, for example].

8A possible alternative for our work would be to use a term cloud view, as in Jones and Li [2008].
that this is known as multi-document summarisation (MDS) [Mani and Maybury 1999], or, in the case of single-document summarisation (SDS), summaries may be generated for individual documents. We are interested in exploring both types.

MDS has drawn attention in recent years, as it can be used to summarise an entire corpus, and enable the user to obtain an overview of it [Yeloglu et al. 2011]. There are various approaches to extractive MDS. For example, common topics can be identified through clustering. This can then be followed by selecting one sentence to represent each (sub)cluster [McKeown et al. 1999]. Alternatively, a composite sentence from each cluster can make up the summary [Barzilay et al. 1999]. Another approach is first to create single document summaries. These summaries can then be grouped in clusters, and representative passages from the clusters extracted [Stein et al. 2000]. Other approaches are described in Radev et al. [2004; Lin and Hovy 2002; Armano et al. [2012; Stokes et al. [2007].

In recent years, there has been interest in topic-focused MDS, e.g. [Wan 2009]. There are many approaches to topic-focused MDS [Daumé III and Marcu 2006; Li et al. 2008, for example]. They adapt traditional summarisation methods so that the information conveyed in the summary is biased towards the given topic.

Summarisation techniques have also been used in information retrieval to summarise search results. Query-biased summarisation applied to web search was found to be useful and effective and users preferred it over a standard search engine presentation format [White et al. 2002; White et al. 2003]. Summaries in a search context tend to be extractive, but this is not always the case [Berger and Mittal 2000, for example].

Our approach can be seen as topic-focused summarisation where the topic is a representation of a profile. The potential of personalised summarisation over generic/traditional summaries has already been demonstrated [Díaz and Gervás 2007, for example]. Summarisation of web documents is typically based on an individual profile (if any) and the query rather than a full profile [Wang et al. 2007; Park 2008, for example]. In our case, we are interested in using the profile of a cohort of users, rather than of individual users.

Evaluating the quality and consistency of a generated summary has proven to be a difficult problem [Fiszman et al. 2009, for example]. The main problem is that there is no clear notion of what constitutes a good summary. Two classes of metrics have been developed: form metrics and content metrics. Form metrics focus on grammaticality, overall text coherence, and organisation. They are usually measured on a point scale [Brandow et al. 1995]. Content metrics are more difficult to measure. Typically, system output is compared sentence by sentence, or unit by unit, to one or more human-generated ideal summaries. As with information retrieval, the percentage of information presented in the system’s summary (precision) and the percentage of important information omitted from the summary (recall) can be assessed. Pyramid evaluation has emerged as a standard method for manual evaluation of summaries [Nenkova and Passonneau 2004]. However, manual evaluation is expensive. Therefore competitions such as the Document Understanding Conferences (DUC)\(^9\) and, more recently, the Text Analysis Conference series (TAC)\(^10\) have made extensive use of automatic evaluation metrics. Automatic evaluation metrics such as those of the ROUGE family [Lin 2004] have been shown to correlate well with human evaluations for content match in text summarisation.

In our work we evaluate the quality of summaries using human ratings and (rather implicitly) as part of task-based evaluations.

\(^9\)http://duc.nist.gov/
\(^10\)http://www.nist.gov/tac/
2.3. Profiles

In order to capture a user’s or a user group’s interests, a model must first be built [Teevan and Dumais 2011]. There are a number of common methods of structuring such models [Gauch et al. 2007]. Models can be built from queries that users submit to search the collection by building query flow graphs [Deng et al. 2009; Boldi et al. 2009, for example], from anchor text [Kraft and Zien 2004], from mining term association rules [Fonseca et al. 2003], or by extracting term relations from documents [Kruschwitz 2005; Sanderson and Croft 1999, for example]. They can aim at modelling individual user’s interests [Teevan et al. 2010] or cohorts of users [Yan et al. 2014]. Models can be explicit, where users input topics of interest, or implicit, where those interests are inferred from their actions [Teevan and Dumais 2011]. However, explicit models have several drawbacks such as the time it takes to build them and their static nature. The types of implicit data used to construct profiles can vary [Teevan and Dumais 2011], e.g., the analysis of log records, has been shown to be good at approximating explicit feedback, and query log analysis has developed into a very active research area [Jansen et al. 2009; Silvestri 2010]. It has been widely recognised that query log files represent a good source for capturing implicit user feedback. This feedback can then be exploited to build knowledge structures that can assist in interactive search, for example to derive query substitutions [Jones et al. 2006], or to extract meaningful knowledge [Baeza-Yates and Tiberi 2007]. There are obvious drawbacks when using log data. These include the fact that they can be noisy and patchy in their coverage. Both of these issues are particularly prevalent when it comes to records that represent the “long tail”. Ways to mitigate against these issues include the use of robust methods to build models, to focus on frequent queries, or to “back off” to generic knowledge sources such as Wikipedia if necessary.\footnote{We do not make use of external knowledge sources due to the problem of adapting them to domain-specific applications [Clark et al. 2012].}

Clark et al. [2012] provide a broad overview of methods to turn document collections as well as query logs into structured knowledge. As pointed out earlier, we are interested in profiles representing a population of web site users instead of building individual profiles (nor are we addressing “collaborative search” in which several people contribute to a shared task [Ringel Morris et al. 2008, for example]). To achieve this we adopt an approach that uses an ant colony optimisation analogy of building adaptive community profiles, a biologically inspired model applied to query logs. This has been shown to be effective for generating query suggestions in intranets (such as a local web site, where suitable knowledge structures are typically not readily available) and is easy to replicate [Albakour et al. 2011]. This approach builds up a network of query terms reflecting the accumulated search knowledge of a cohort of users which can be continuously updated. We will focus on a rather broad cohort by only building a single profile, and not distinguishing between individual user groups accessing the web site. It is however a simple step from the approach we take to building profiles that make finer-grained distinctions based on the login credentials of a user as long as that information is present in the log files (e.g. models that reflect students, or administrative staff, or first-year undergraduate Biology students, etc). While the profile we use is built using logs relating to queries submitted to a local search engine, we assume that such a profile is also relevant for web site navigation.

2.4. Key Differentiators

There has been much prior work on assisting users in finding information in a hyperlinked environment. Our work is novel in targeting a cohort of users and combining summarisation techniques and profiling, using query logs, to assist a user of the co-
hort in finding information when navigating a web site. While summarisation has been used for search, there is very little work in assisting a user in navigation and browsing without altering the actual content of the web site itself. We identified a single web site to conduct our studies. We would argue that the site chosen is a good example of the type of site where these methods can be applied. We hope that our work will serve as a benchmark for future studies.

3. BUILDING A LOG-BASED PROFILE

Query log analysis has become one of the most promising research areas for the automatic derivation of knowledge structures for search assistance [Jansen et al. 2009; Silvestri 2010, for example]. This motivates our use of log data in generating cohort-personalised summaries to assist in navigation-based search.

Different methods for exploiting query logs to derive new query modification suggestions, for web-site search, have been explored by Kruschwitz et al. [2013]. The methods can be classed as either adaptive, e.g. ant colony optimisation, or non-adaptive, such as query flow graphs. Adaptive methods are able to learn over time in a continuous learning cycle to build and adapt a domain model representing queries and query suggestions. Changing the frequency of the update is then likely to result in a different model. In contrast, non-adaptive methods take the entire log and derive suggestions from the aggregated data; building such models incrementally would not result in a different model. Kruschwitz et al. [2013] demonstrate that the log-based methods (even without incorporating click-through information) outperform non-log based approaches for query suggestion quality. Of the methods considered, ACO performs well overall and is simple to implement. It is particularly appealing due to its adaptive nature and its capability to capture some temporal context.\(^\text{12}\) For these reasons, we adopted ACO to construct a domain model for our experiments — although the details of the specific model are not under investigation in the work described here.

A major bottleneck in conducting research into query logs is the difficulty in obtaining realistic, large-scale log data. In our experiments we use the logs of a local web-site search engine. These have been collected over a three-year period.

3.1. Query Logs

The query logs we are using have a fairly standard format; nothing special is assumed about the structure. Here is an extract from the actual (pre-processed) log files ("xxx" is a field separator):

```
... 1657406 xxx 02f4e527278fbe909aedb97cea27f8d9 xxx wed nov 17 11:41:43 gmt 2010 xxx 0 xxx 0 xxx 0 xxx\ library xxx library xxx library
... 1657409 xxx 02f4e527278fbe909aedb97cea27f8d9 xxx wed nov 17 11:41:58 gmt 2010 xxx 0 xxx 0 xxx 0 xxx\ albert salmon xxx albert salmon xxx albert salmon
... 1657412 xxx 02f4e527278fbe909aedb97cea27f8d9 xxx wed nov 17 11:42:23 gmt 2010 xxx 0 xxx 0 xxx 0 xxx\ albert salmon library xxx albert salmon library xxx albert salmon library
... 1657415 xxx 02f4e527278fbe909aedb97cea27f8d9 xxx wed nov 17 11:42:37 gmt 2010 xxx 0 xxx 0 xxx 0 xxx\ albert sloman library xxx albert sloman library xxx albert sloman library
...```

The logs record (i) a query identifier, (ii) a session identifier, (iii) the submission time, (iv) the submitted query, and some other additional information, see also Kruschwitz et al. [2013]. The above extract shows four interactions submitted within the same

\(^{12}\) Regarding the adaptive nature of ACO, previous studies applying ACO to query logs have demonstrated that the ACO paradigm can learn useful query suggestions and improve over time [Albakour et al. 2011; Kruschwitz et al. 2011, for example]
session. In a sequence of steps the user replaces the original query “library” by a new query “albert salmon”, then by “albert salmon library” and the final query “albert sloman library”.

As can be seen from the sample log entries, we do not identify individual users, nor do we associate IP addresses with sessions. This is to comply with data protection issues and to avoid potential privacy problems. But it is also consistent with the idea of treating all users as part of the same cohort, which fits with the aims of the current study, and is in line with alternative approaches to address privacy and security concerns when building personalised search systems, such as I-SPY, which are aimed at the needs of a community without the need to store individual search histories [Smyth et al. 2003; Freyne et al. 2004].

The second field in the log record is an automatically generated session identifier. Automatically identifying the boundaries of sessions is a difficult task [Göker and He 2000; Jansen et al. 2007]. One of the reasons is that a session can consist of a number of search goals and search missions [Jones and Klinkner 2008]. Identifying topically related chains in user query sessions has been studied extensively [Gayo-Avello 2009]. We use the default server timeout, i.e. a session expires after 30 minutes of inactivity. This method has been shown to give highly accurate session boundaries [Jansen et al. 2007]. To test the applicability of this approach to the current work we randomly sampled 50 sessions from the log files. Three of the authors independently assessed whether each of those sessions was concerned with a single topic. They then compared their judgements. There was agreement that all sessions were about a single topic. In addition, we found that there was no session longer than 30 minutes. However, given that sessions in our sample domain tend to be short — with only 1.53 queries per session on average [Kruschwitz et al. 2013] — we randomly sampled another set of 50 sessions containing at least two queries. Using the same manual assessment, no single session was identified that was clearly and unambiguously about more than one topic, although there were six sessions that potentially fell into that category (e.g. a query “study abroad” followed by “psychology”). Again there was no session longer than 30 minutes. We conclude that applying the standard timeout approach appears sensible in this study.

3.2. Ant Colony Optimisation Model

Ant Colony Optimisation (ACO) is a form of swarm intelligence technique. It is motivated by consideration of ant behaviour, with the laying and following of pheromone trails that can evaporate over time. It has been studied extensively in the context of solving problems in domains such as scheduling [Socha et al. 2003], classification [Martens et al. 2007] and telecommunications routing [Di Caro and Dorigo 1998].

In the current work, the ACO analogy is used to first populate and then adapt a directional graph similar to Query Flow Graphs. In this analogy the edges in the graph are weighted with the pheromone levels that the ants, in this case users, leave when they traverse the graph. This is a fully automated process entirely relying on implicit cues. This is different to the idea of AntWorld, which also aims at assisting users in navigating web sites but requires some explicit user judgements [Kantor et al. 2000].

In effect, the user traverses a portion of the graph by using query refinements (analogous to the ant’s journey). The weights of the edges on this route are reinforced (increasing the level of pheromone). Over time, all weights (pheromone levels) are reduced by introducing an evaporation factor. This notion of evaporation captures the reduced popularity of an edge that has not been used recently by reducing the weight of

---

13 The actual model acquisition process is not the focus of the paper; the description given here is largely reproduced from Kruschwitz et al. [2013].
untraversed edges over time. This then penalises inappropriate or less relevant query modifications. In this way, we expect outdated terms effectively to be removed from the model.

While ACO is an adaptive technique that can be continuously updated, for our experiments we will take a snapshot of the model; we do not continuously update it during the experiments. This helps to simplify the experimental design, and focus on assessing the impact of using cohort-personalised summarisation to guide navigation.

Let us assume that we update the pheromone levels on a daily basis. For the edge $q_i \rightarrow q_j$ the pheromone level $w_{ij}$ is updated using Equation 1.

$$w_{ij} = N \ast ((1 - \rho)w_{ij} + \Delta w_{ij})$$ (1)

where:

1. $N$ is a normalisation factor, as all pheromone trail levels are normalised to sum to 1;
2. $\rho$ is an evaporation co-efficient factor;
3. $\Delta w_{ij}$ the amount of pheromone deposited at the end of the day for the edge $q_i \rightarrow q_j$.

The amount of pheromone deposited should correspond to ant moves on the graph. In our case, this can be the frequency of query refinements corresponding to the edge. Also the cost of ant moves can be taken into account when calculating the amount of pheromone deposited. Generally, it can be calculated using Equation 2 [Dorigo et al. 2006].

$$\Delta w_{ij} = \Sigma_k Q/C_k; \text{ For all ant moves on edge } q_i \rightarrow q_j$$ (2)

where:

(a) $Q$ is a constant,
(b) $C_k$ is the cost of ant $k$ journey when using the edge $q_i \rightarrow q_j$.

Previous work reports on experiments with a number of evaporation factors for Equation 1 and pheromone deposition calculation schemes in Equation 2 [Albakour et al. 2011; Albakour 2012]. Based on the finding of those studies, we used the best-performing combination of the parameters, in which the evaporation factor $\rho = 0.1$ and only immediate refinements in the sessions are considered to update the pheromone level on the graph (to be discussed below). In this work, the constant $Q$ in Equation 2 is chosen to be the average weight of all edges in the graph in the previous day. The cost $C_k$ is considered to be the distance between the queries in the session (for immediate refinements $C = 1$).

We build the edges on the graph by considering only subsequent queries in the same session (‘immediate refinements’). In other words, we use session boundaries, and treat each session as a transaction. For example, if we have the following query chain in a session: query$_k$, query$_l$, query$_m$, the edges considered are query$_k \rightarrow$ query$_l$, and query$_l \rightarrow$ query$_m$ as ant movements (the use of “immediate refinements” has been shown to outperform transitive relationships [Albakour 2012; Albakour et al. 2011]). At the end of each day, all edge weights are normalised to sum to 1, and the mean weight of all edges is then calculated. By normalising the weights, all weights (pheromone levels) of non-traversed edges are reduced over time.

To preprocess the log files, the logs are segmented into sessions. For each session, the queries are time-ordered. We process the logs by keeping only those sessions that contain more than one search query. To reduce noise, only sessions with ten or fewer queries are considered.\footnote{Longer sessions are potentially better to model individual searcher profiles [Bennett et al. 2012, for example], but for a group profile we assume shorter sessions to be the main building blocks of the model. In any case, only 0.31% of all our sessions are longer than ten queries.}

We perform case folding — i.e. all capital letters are trans-
formed into small letters — in order to normalise the query corpus. We also replace all punctuation marks such as colons, semicolons, and dashes by white space. We then use the processed log file to build the profile.

Figure 2 illustrates a snapshot of node “library” in the profile acquired using the three-year query logs (only the highest-weighted links are displayed). For comparison, Figure 3 shows what the corresponding part of the profile looks like after starting the construction at the same point of time, but running the model acquisition process on the first month of the query logs only.

3.3. Deriving related terms

If we were using the model as a tool to assist in search, then we could extract query suggestions from the graph by locating the original query as a node in the graph and then list all directly connected nodes ranked by edge weights. For the submitted query “library” in Figure 2, the list of possible query modification suggestions would then be “library opening times”, “catalogue”, “moodle”, “cmr” and “albert sloman”.

In a navigation context, we need to identify suitable terms that represent the document at hand and then perform the same corresponding steps. We will use the document title to do this. This will be the underlying principle of how the model can be used for profile-based summarisation. Alternatively, one might consider using the anchor text on which a user clicks or possibly the whole item containing the anchor text (exploring this idea is left as future work).

Fig. 2. Acquiring a profile from query logs.

4. DATASETS

A copy of the existing web site was created by crawling the main web site of the University of Essex with Nutch\textsuperscript{15}. We limited the crawl to a maximum depth of fifteen. The crawl for Essex was performed in November 2013. The resulting dataset contains 110,369 documents. We also simulated the Intranet search engine of the institution on

\footnote{\url{http://nutch.apache.org/}}
our machines using Apache Solr\(^{16}\) so that if a user decided to use search rather than navigation we would still be able to track their steps.

A total of more than 1.5 million queries, described in more detail elsewhere [Kruschwitz et al. 2013], were extracted from the logs of the existing site search engine over a period of three years (20 November 2007 till 19 November 2010) to build our model. No click-through information was exploited; this makes the techniques simpler and easier to replicate. In any case, click-through information was not available for much of the log-data collected for the model.

5. PROFILE-BASED SUMMARISATION

To generate a profile-based summary of a document (or a set of documents) we can, in the simplest case: (i) locate a suitable term as a node in the domain model (e.g. “library”) in Figure 2); (ii) extract all directly connected nodes, to generate a bag of words/terms that are related in the model;\(^{17}\) finally, (iii) a summary can then be generated by extracting sentences that are related to these terms. The hope is that a summary generated in this way will be helpful in a search and navigation task.

We propose to use a form of query-based summarisation in our experiments. In a search context, this is quite straight-forward: we can use the user’s query as our starting point in the model. But in a navigation context, there is no query.\(^{18}\) In this case, we assume instead that the title of a hyperlinked document can play the role of a “query”, and provide a suitable starting point in the domain model. This seems justified in a web site context that is virtually spam-free, and where content policies are in place. We are aware that such an assumptions do not always hold [Hawking and Zobel 2007, for example]. We then use those terms in the domain model that are related to the title in the “query-based” summarisation process. This approach is adopted for both SDS and MDS (as described in the following subsections).

If the resulting summary is empty, no summary will be displayed. An empty summary may be due to the nature of the document, for example there may be no text in the document, or the type of the document might not be supported. In the case of profile-based summarisation, there might also be no matching concept in the domain model.\(^{19}\)

\(^{16}\)http://lucene.apache.org/solr/
\(^{17}\)Intuitively at least, these terms are then related in a way that appears appropriate for a search task.
\(^{18}\)Although there might be a relevant query if the user is interacting with a search engine while navigating.
\(^{19}\)To test how common this might be we randomly sampled 50 HTML pages from the document collection and did not find a single case where there were no terms extracted from the title of the page matching a
5.1. Data Collection Preprocessing

Generating a summary for HTML documents involves a number of preprocessing steps [Stokes et al. 2007; Smucker 2011; Silva and Ribeiro 2003; Frakes 1992]. First, we perform some document cleaning, e.g., removing tags and tables. This is followed by a pipeline of NLP tools\textsuperscript{20}, namely sentence splitting, tokenisation, stop words elimination as well as stemming.

5.2. Profile-Based Single-Document Summarisation

Algorithm 1 illustrates generic pseudo-code of the profile-based SDS process.\textsuperscript{21} More specifically, in the experiments described here, we perform the following steps: First, the title is extracted, normalised and parsed identifying patterns used in terminological feedback extraction, namely nouns and noun phrases up to three words long [Justeson and Katz 1995]. Then, the longest possible phrases are extracted (e.g. if the set contains both “albert sloman” and “albert sloman library”, then the first term will be removed in this step). This results in a (possibly empty) set of terms. For each term, we check whether it is represented as a node in the profile. If it is, we extract all directly connected nodes (i.e. related “queries”) and construct the union of all these terms. For example, if the document title was The Library, then the only term considered is “library”. This then results in the collection of terms {“library”, “library opening times”, “catalogue”, “moodle”, “cmr”, “albert sloman”} being generated (Line 4 of Algorithm 1), as illustrated in Figure 2.\textsuperscript{22}

\begin{algorithm}
\caption{Profile-based single-document summarisation algorithm}
\begin{algorithmic}[1]
\REQUIRE document D, profile P\\\n\ENSURE a summary for a document\\\n\Begin\\\n\STATE T $\leftarrow$ GetDocumentTitle(D)\\\n\STATE /* Extract related terms from the profile / domain model */\\\n\STATE RelatedTerms $\leftarrow$ ExtractRelatedTerms(T,P)\\\n\IF{RelatedTerms $\neq$ NULL}\\\n\STATE Split D into sentences, $D = \{S_1, S_2, ..., S_n\}$\\\n\STATE /* Find similarity values between RelatedTerms and each sentence, and store them in array */\\\n\FOR{each $S_i \in D$}\\\n\STATE SM[i] $\leftarrow$ Similarity($S_i$, RelatedTerms)\\\n\END\FOR\\\n\STATE RankOrder SM[] descending\\\n\STATE SummarySet $\leftarrow$ ExtractTopSentences(SM[])\\\n\STATE Summary $\leftarrow$ SortSummaryAccordingToOriginalDoc(SummarySet)\\\n\Return Summary\\\n\End\\\n\end{algorithmic}
\end{algorithm}

\textsuperscript{20}We use OpenNLP for some of these steps.

\textsuperscript{21}The method ExtractRelatedTerms simply extracts all terms from the profile that have a direct link to any of the terms extracted from the document title. More sophisticated methods such as random walk-based approaches are possible alternatives.

\textsuperscript{22}By ignoring the model weights and treating the terms as a bag of terms there is potentially a risk of query drift when expanding by many weakly connected nodes. We deliberately went for a simple summarisation step and leave it as future work to explore more sophisticated approaches.
A document is preprocessed and segmented into sentences. All sentences are rank-ordered according to the standard tfidf cosine similarity when compared to the terms extracted from the domain model. Finally, the candidate sentences (summary sentences) are sorted according to the order in the original document. Following DUC 2002 conventions, we generated summaries of at most 100-words [Lin and Hovy 2003].

5.3. Profile-Based Multi-Document Summarisation

For profile-based MDS we apply a similar sequence of steps as in Algorithm 1, but in this case the “document” that is summarised is generated from a collection of related documents. Normally, MDS is applied to a collection of documents that are related to each other. In our application, the collection of related documents is generated from a root hyperlinked document. This is done by extracting all outgoing links from this document, and retrieving the corresponding documents. These documents are then concatenated to create the “meta-document” that will be summarised. Following the extraction of candidate sentences for the summary, the sentences are ordered according to their similarity to relevant terms in the domain model. To avoid duplicate sentences appearing in a summary, we also apply a simplistic redundancy elimination step. This process is illustrated in Figure 4.

6. A PILOT STUDY ON PROFILE-BASED SUMMARISATION

We start by exploring the usefulness of the cohort-profile domain model in generating profile-based summaries. In this pilot study, we create summaries as if they were generated in a navigation context; more specifically, we generate “query-based” summaries using the title of the relevant page (see Section 5). We evaluate a range of profile-based techniques using both single-document and multi-document summarisation.

6.1. Experimental Setup

In order to generate our sample summaries, we use documents that correspond with frequently submitted queries, as commonly done in other studies [Paek et al. 2004, for example]. More specifically, we identified the ten most frequent queries in the logs (“timetable”, “courses”, “moodle”, “accommodation”, “library”, “fees”, “law”, “enrol”, “psychology”, and “graduation”), and then identified the top matching document for each query, as returned by the Google search engine. We then recruited human subjects to evaluate the summaries. The evaluators were in two groups, local users, and remote users, specifically MTurk workers.

(1) Local Users: Given the site-specific focus of this work, we recruited subjects to represent our target users; students of our institution. These represent typical users of the local search engine. To recruit a range of different types of student users, and to help avoid bias in the selection process, we sent an e-mail to the local university mailing list (a list that is used to send out “small ads” to students) and selected the first ten volunteers who replied. They were informed in advance that they would receive £10 for participating. Users were provided with hard copies of all of the original documents (the ones for which we generated summaries) and asked to read them before they rated the summaries on an evaluation form.

(2) Web Users: We also recruited ten subjects from an online workforce service, namely Mechanical Turk24 (MTurk). We gave MTurk workers an electronic version of the evaluation form, and links to the original documents. They were asked

---

23For example, by submitting the query “timetable site:essex.ac.uk” to http://www.google.com/.  
24https://www.mturk.com/
Fig. 4. Architecture of a profile-based multi-document summariser.

to read the documents first and then complete the form to provide us with their ratings. Subjects were paid $10 each.

For the local users, the evaluations typically took 30–40 minutes.

We used the following six methods to create the summaries, with the first three being treated as baselines:

(1) **Random SDS**: (“Baseline-1”) a random selection of sentences from the document (a baseline that is also employed by Yan et al. [2011]).

(2) **Centroid SDS**: (“Baseline-2”) a centroid-based approach to summarisation [Radev et al. 2004]. This algorithm takes into account the following parameters: centroid value, positional value, and first-sentence overlap, in order to generate a summary.

(3) **Centroid MDS (all documents)**: (“Baseline-3”) a centroid-based approach for summarising a collection of related documents. We retrieve all documents to which the document in hand has links, create a meta-document by concatenating all these documents, and then apply “Baseline-2” to this meta-document.
(4) **ACO title-based SDS**: ACO-based single-document summarisation, using the document title, as described in Section 5.2.

(5) **ACO title-based MDS (first five documents)**: ACO-based multi-document summarisation as described in Section 5.3 but only considering the first five outgoing links in the document.

(6) **ACO title-based MDS (all documents)**: ACO-based multi-document summarisation as described in Section 5.3, considering all outgoing links in the document.

Obviously, there would have been many other alternative summarisation approaches that could have been chosen. It could be argued that there should have been other query-dependent summarisation methods included in the methods under investigation. This is a valid argument, as anything we conclude from the experiment could be due to the specific query-dependent summarisation method rather than the actual profile. We decided to apply ACO for the reasons given in Section 3 and leave the exploration of alternative models for future studies.

We adopted an existing evaluation framework [Yan et al. 2011] to assess the quality of summarised documents, and allow a comparison between methods. Subjects were asked to assess the summaries (provided in a random order) on a five-point Likert scale\(^2\), where 5 was “excellent”; 4, “good”; 3, “average”; 2, “bad”; and 1, “terrible”.\(^2\)

We can illustrate the results of the different methods when applied to the University of Essex Accommodation homepage\(^2\):  

**Random.** Online applications are now open for all new undergraduate students for 2014-15. We offer a range of support services for students to help with everyday worries about your accommodation or for practical or personal issues. Where you live is a very important part of your life at university. Applications are now open for all returning undergraduate and postgraduate students for 2014-15. Applications have now re-opened, we apologise for any inconvenience caused. Construction work is underway near the residences and may cause some noise disruption to residents during 2013-14.

**Centroid SDS.** Where you live is a very important part of your life at university. Our accommodation provides the perfect base from which to start this new and exciting period in your life. Applications are now open for all new undergraduate students for 2014-15. Applications are now open for all returning undergraduate and postgraduate students for 2014-15. Applications have now re-opened, we apologise for any inconvenience caused. We are building exciting new facilities for our students at our Colchester campus. Construction work is underway near the residences and may cause some noise disruption to residents during 2013-14.

**Centroid MDS (all documents).** Where you live is a very important part of your life at university. Our accommodation provides the perfect base from which to start this new and exciting period in your life. If you have withdrawn temporarily from the University (intermitted) during the first term of the first year of study, and decide to return to the University the following academic year, you may re-apply for accommodation. Being an Essex student offers a unique experience of living and learning, sending you on an unforgettable journey that will shape your future.

**ACO title-based SDS.** Where you live is a very important part of your life at university. Online applications are now open for all new undergraduate students for 2014-15. Applications are now open for all returning undergraduate and postgraduate students for 2014-15. We are building exciting new facilities for our students at our Colchester campus. Keep up to date with the latest accommodation updates and offers from across our Colchester and Southend Campuses. We offer a range of support services for students to help with everyday worries about your accommodation or for practical or personal issues.

**ACO title-based MDS (first five documents).** All full-time, fully registered, postgraduate students studying at Colchester or Southend campuses are guaranteed single accommodation for their first year of study, provided they return their application and GBP250 prepayment deposit after applications open.

---

\(^2\)Strictly speaking, we employ Likert-type scales but refer to Likert scales throughout the study.

\(^2\)The total number of generated summaries for the ten documents was 59. This was because there was an empty summary for one document (ACO title-based SDS). The summary was empty because there were no matching sentences. We excluded this summary from the evaluation form and assigned it a rating of 1 (“terrible”). Hence, although each subject rated 59 summaries, in the follow-up analysis we have ratings for all 60 summaries.

\(^2\)http://www.essex.ac.uk/accommodation/
in April 2014. All full-time fully-registered undergraduate students studying at either our Colchester or Southend Campuses are entitled to single accommodation for the first year of study as long as they return their application by the deadline date, Friday 22 August 2014. If you have a sports bursary and want to apply for accommodation, please contact the Sports Centre direct.

**ACO title-based MDS (all documents).** All full-time, fully registered, postgraduate students studying at Colchester or Southend campuses are guaranteed single accommodation for their first year of study, provided they return their application and GBP250 prepayment deposit after applications open in April 2014. All full-time fully-registered undergraduate students studying at either our Colchester or Southend Campuses are entitled to single accommodation for the first year of study as long as they return their application by the deadline date, Friday 22 August 2014. This is normally available through the myEssex Applicant Portal.

### 6.2. Results and Discussion

In Table I, for each summarisation method, we report the average ratings obtained from each group of users (local users and web users). We visualise these results using the box plot in Figure 5. In both user groups, we can observe that all variations of the ACO-based algorithm outperform the other alternatives in achieving a higher average rating. We employ a non-parametric measure (Friedman) to assess the variance by ranks for the ratings across the different summarisation methods in each group of users. The Friedman test compares the mean ranks for each of our six summarisation methods against the mean ranks for all of the remaining five methods. In addition, we report the $z$-score for the rating ranks. The $z$-score represents how much the rating ranks of a method deviate from the rating ranks of all other methods, and its polarity indicates whether the difference is positive or negative and its value indicates significance. The $z$-scores for each method are also included in Table I to indicate levels of significance.

<table>
<thead>
<tr>
<th>System</th>
<th>Local Users</th>
<th>Web Users</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>z</td>
</tr>
<tr>
<td>Random</td>
<td>1.86</td>
<td>(-9.39)</td>
</tr>
<tr>
<td>Centroid SDS</td>
<td>2.51</td>
<td>(-5.44)</td>
</tr>
<tr>
<td>Centroid MDS (All)</td>
<td>3.04</td>
<td>(-0.89)</td>
</tr>
<tr>
<td>ACO title-based SDS</td>
<td>3.47</td>
<td>(2.69)</td>
</tr>
<tr>
<td>ACO title-based MDS (First 5)</td>
<td>4.52</td>
<td>(9.88)</td>
</tr>
<tr>
<td>ACO title-based MDS (All)</td>
<td>3.62</td>
<td>(3.15)</td>
</tr>
</tbody>
</table>

In the following, we detail the results for each group of users:

(1) **Local Users** (Friedman $\chi^2 = 33.7391$, $df = 5$, $p \ll 0.001$): The small $p$-value ($p \ll 0.001$) suggests that the choice of the summarisation method has an effect on the users’ rating. The post-hoc tests address the pairwise comparisons. Wilcoxon Signed Rank tests reveal that only five pairs of the methods do not show significant differences, namely Centroid SDS vs. Centroid MDS (All), Centroid SDS vs. ACO title-based SDS, Centroid MDS (All) vs. ACO title-based SDS, Centroid MDS (All) vs. ACO title-based MDS (All), and ACO title-based SDS vs. ACO title-based MDS (All). We note that Random and Centroid SDS are significantly worse than the average at a 95% confidence level. Centroid MDS (All) is worse than the average but not significantly at a 95% confidence level ($z > -1.96$). All ACO variations perform

 ACM Transactions on Information Systems, Vol. 0, No. 0, Article 0, Publication date: August 2014.
significantly better than the average at a 95% confidence level.\textsuperscript{28} Table II has more details on the post hoc tests.\textsuperscript{28}

(2) \textbf{Web Users} (Friedman $\chi^2 = 43.3871, df = 5, p < 0.001$): All differences are significant except Centroid SDS vs. Centroid MDS (All), and Centroid MDS (All) vs. ACO title-based SDS. Random, Centroid SDS and Centroid MDS (All) are significantly worse than the average at a 95% confidence level. In line with the results for Local Users, all ACO variations are significantly better than the average at a 95% confidence level.\textsuperscript{29} Table III gives detailed post hoc test results.

Figure 5 nicely illustrates the “missing summary”, as under ACO title-based SDS we see an outlier for both Local and Web Users representing the document for which no summary was generated and which we therefore penalised in order to have a fair comparison. The actually generated summaries for this method were judged much more positively.

\begin{table}[h]
\centering
\caption{Local users: $p$-values of Wilcoxon Signed Rank post-hoc pairwise tests.}
\begin{tabular}{cccccc}
\hline
 & Random & Centroid SDS & Centroid MDS (All) & ACO SDS & ACO MDS (First 5) & ACO MDS (All) \\
\hline
Random & 0.0127 & 0.0092 & 0.0108 & 0.0019 & 0.0059 & \\
Centroid SDS & 0.0526 & 0.0664 & 0.0019 & 0.0365 & \\
Centroid MDS (All) & 0.2616 & 0.0059 & 0.1055 & \\
ACO SDS & 0.0107 & 0.7211 & \\
ACO MDS (First 5) & 0.0140 & \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Web users: $p$-values of Wilcoxon Signed Rank post-hoc pairwise tests.}
\begin{tabular}{cccccc}
\hline
 & Random & Centroid SDS & Centroid MDS (All) & ACO SDS & ACO MDS (First 5) & ACO MDS (All) \\
\hline
Random & 0.0142 & 0.0059 & 0.0059 & 0.0059 & 0.0059 & \\
Centroid SDS & 0.0502 & 0.0488 & 0.0059 & 0.0059 & 0.0059 & \\
Centroid MDS (All) & 0.1522 & 0.0019 & 0.0089 & \\
ACO SDS & 0.0092 & 0.0091 & \\
ACO MDS (First 5) & 0.0142 & \\
\hline
\end{tabular}
\end{table}

We also compare the ratings of the two user samples. Mann-Whitney U tests between each pair of corresponding methods applied to the two user groups does not result in any significant difference. We can conclude that general web users’ assessments of the quality of summaries is consistent with the assessments of local users. This is interesting, given the domain-specific nature of the documents. It might indicate that the algorithm, i.e. incorporating usage data, is what makes the difference, rather than the cohort. Looking at finer-grained classes of cohorts than just local users vs. non-local users should uncover which groups (if any) benefit most from the proposed approach or whether it is really the usage data as such that makes the difference.

The main conclusions we can draw from this pilot study is that all variations of the ACO-based algorithm outperform the other methods, and that multi-document summarisation offers the biggest potential, in particular when only choosing the first

\textsuperscript{28}Applying the rather conservative Bonferroni adjustment suggests that there are only two significant differences, namely between Random vs. ACO title-based MDS (First 5), and Centroid SDS vs. ACO title-based MDS (First 5).

\textsuperscript{29}A Bonferroni adjustment results in only one significant difference between Centroid MDS (All) and ACO title-based MDS (First 5).
five outgoing links to generate the summary. In other words, not simply relying on the target document but also incorporating “related” documents can improve the quality of the summary. Future work could look into employing incoming rather than outgoing links.

A different reading would be to conclude that query- or profile-based summaries are better than query-independent summaries. This does not contradict our findings about the potential of applying ACO for profile-based summarisation, although further experiments could explore how well this methodology scores against alternative query- or profile-based approaches.

7. TASK-BASED EVALUATIONS

The findings of the pilot study suggest that profile-based summarisation has the potential to generate document summaries that are better than summaries that do not use a profile, or, more conservatively, that incorporating usage data from a web site helps make better summaries than not using any usage data. We wish to determine whether the apparent improvement in summarisation quality has a measurable effect in the context of a navigation task. More specifically, when using summaries to assist in a navigation task, we wish to know whether profile-based summaries allow users find information more easily and quickly, with improved levels of satisfaction.

Task-based evaluations aim to simulate real tasks as closely as possible while assessing various aspects of users’ performance. Here we will evaluate a navigation task to determine the utility of different summarisation methods. We adopt a standard approach to these task-based evaluations [Diriye et al. 2010; Yuan and Belkin 2010; Kelly 2009]; in these laboratory experiments, we use a within-subjects design to compare three systems. We broke the evaluations into two experiments, involving different baselines. This was because it is difficult to compare more than three systems with the established approach that we adopted.

The experiments were primarily concerned with a navigation task, but subjects were also free to use a local search engine that was part of the web site. In line with Kelly [2009], we chose eighteen subjects, each of whom attempted to complete six search tasks. These tasks required the subjects to find documents that are relevant to predetermined topics. Each subject was expected to completed two different tasks on each
system. In accordance with TREC-9 Interactive Track guidelines [Hersh and Over 2001; Hersh 2002], subjects had ten minutes for each task. They were asked to complete standardised questionnaires.

7.1. Questionnaires
We used questionnaires based on those suggested by the TREC-9 Interactive Track guidelines.30 We made some very minor modifications to the entry questionnaire to reflect the terminology of the British education system, and included Google as an example of a web search engine. A five-point Likert scale was used where appropriate. Note that for the experiments reported here, the term “search” on these standardised questionnaires is intended to be interpreted as “search by navigation”. We used the following questionnaires:

1. Entry questionnaire, which collects demographic and internet usage information.
2. Post-search questionnaire, which is used to assess a user’s perspectives of the systems and the tasks.
3. Post-system questionnaire, which is used to capture the user’s perceptions for each of the systems.
4. Exit questionnaire, which focuses on a comparison between the three systems.

7.2. Protocol and Search Tasks
We followed the procedure adopted in Craswell et al. [2003] to guide the subjects during the task-based evaluation. The experiments were conducted in an office in a one-on-one setting. We have three systems and six tasks and their orders were revolved and counterbalanced. At the start of each session, subjects were asked first to complete an entry questionnaire. After that, subjects were given five minutes’ introduction to the three systems — without being told anything about the technology behind them. Then, each subject was expected to perform two different tasks on each system. The interactions with the system were logged electronically. To finish a task, subjects were instructed to click on a link that was provided. This would record the page they found in the log of the experiment, together with a time stamp. After completing each individual task, subjects were asked to complete a post-search questionnaire. When they had finished both tasks on one system, they were asked to complete a post-system questionnaire. Finally, when subjects finished all the tasks, they were asked to complete an exit questionnaire.

To make the tasks realistic, they were designed using information about commonly submitted queries, as recorded in the search logs of the existing web site. Constructing tasks using query logs is a common approach [Kelly 2009]. The tasks had actually been constructed for a previous experiment within the same domain [Adindla and Kruschwitz 2013; Adindla 2014] to allow cross-study comparisons. They have some similarity with Borlund’s “simulated work tasks” [Borlund 2000], and also take into account the guidelines suggested by Kules and Capra [2008].

We assume that users with common, frequently occurring information needs will benefit most from the framework proposed here: frequent queries submitted on a web site tend to make up a large proportion of all submitted queries [Kruschwitz et al. 2013, for example]. Given the “long tail” of rare queries in a life system, we might wish to ignore such queries in any domain model. Here, this is achieved as a side-effect of the “evaporation factor” in the ACO model.

http://www-nlpir.nist.gov/projects/t9i/qforms.html
Table IV. A basic design with Graeco-Latin square rotation for topic and interface [Kelly 2009].

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>I1: 1, 2</td>
<td>I2: 3, 4</td>
<td>I3: 5, 6</td>
</tr>
<tr>
<td>S2</td>
<td>I1: 2, 3</td>
<td>I2: 4, 5</td>
<td>I3: 6, 1</td>
</tr>
<tr>
<td>S3</td>
<td>I1: 3, 4</td>
<td>I2: 5, 6</td>
<td>I3: 1, 2</td>
</tr>
<tr>
<td>S4</td>
<td>I1: 4, 5</td>
<td>I2: 6, 1</td>
<td>I3: 2, 3</td>
</tr>
<tr>
<td>S5</td>
<td>I1: 5, 6</td>
<td>I2: 1, 2</td>
<td>I3: 3, 4</td>
</tr>
<tr>
<td>S6</td>
<td>I1: 6, 1</td>
<td>I2: 2, 3</td>
<td>I3: 4, 5</td>
</tr>
<tr>
<td>S7</td>
<td>I2: 1, 2</td>
<td>I3: 3, 4</td>
<td>I1: 5, 6</td>
</tr>
<tr>
<td>S8</td>
<td>I2: 2, 3</td>
<td>I3: 4, 5</td>
<td>I1: 6, 1</td>
</tr>
<tr>
<td>S9</td>
<td>I2: 3, 4</td>
<td>I3: 5, 6</td>
<td>I1: 1, 2</td>
</tr>
<tr>
<td>S10</td>
<td>I2: 4, 5</td>
<td>I3: 6, 1</td>
<td>I1: 2, 3</td>
</tr>
<tr>
<td>S11</td>
<td>I2: 5, 6</td>
<td>I3: 1, 2</td>
<td>I1: 3, 4</td>
</tr>
<tr>
<td>S12</td>
<td>I2: 6, 1</td>
<td>I3: 2, 3</td>
<td>I1: 4, 5</td>
</tr>
<tr>
<td>S13</td>
<td>I3: 1, 2</td>
<td>I1: 3, 4</td>
<td>I2: 5, 6</td>
</tr>
<tr>
<td>S14</td>
<td>I3: 2, 3</td>
<td>I1: 4, 5</td>
<td>I2: 6, 1</td>
</tr>
<tr>
<td>S15</td>
<td>I3: 3, 4</td>
<td>I1: 5, 6</td>
<td>I2: 1, 2</td>
</tr>
<tr>
<td>S16</td>
<td>I3: 4, 5</td>
<td>I1: 6, 1</td>
<td>I2: 2, 3</td>
</tr>
<tr>
<td>S17</td>
<td>I3: 5, 6</td>
<td>I1: 1, 2</td>
<td>I2: 3, 4</td>
</tr>
<tr>
<td>S18</td>
<td>I3: 6, 1</td>
<td>I1: 2, 3</td>
<td>I2: 4, 5</td>
</tr>
</tbody>
</table>

The tasks assigned were randomised using a Graeco-Latin square design [Kelly 2009] to avoid task bias and potential learning effects (see Table IV). The tasks were constructed based on the following queries: “accommodation”, “undergraduate”, “funding”, “postgraduate”, “staff” and “international”.

Task-1. You have been accepted for a place at the University of Essex at the Colchester campus. Find information on the residences, accommodation information for freshers, contact details and other useful information.

Task-2. You are interested in applying for an undergraduate course in the School of Computer Science and Electronic Engineering (CSEE). Find information for prospective students and the courses being offered within the school.

Task-3. You are going to be a new postgraduate student at the University of Essex. You need to locate a page with useful information about tuition fees and possible funding offered by the university.

Task-4. Find documents that allow you to download the postgraduate prospectus and view maps of the University of Essex campuses.

Task-5. You need to find out who is the head of the Department of Philosophy at the University of Essex, and a list of people or office holders within the school.

Task-6. You have registered as an international student at the University of Essex. You would like to locate information regarding new and current students, freshers information and administration information. Locate documents to help you look for the right information.

7.3. Subjects
In order to obtain a good selection of different types of users, and to avoid bias in the selection process, we sent an e-mail to the same local mailing list (i.e. “small ads” for students) and, for each experiment, we selected the first eighteen volunteers who replied. None of the subjects took part in more than one of our studies. The background of the subjects was mixed, both in the level of their degree (Bachelor, Master and PhD)

---

31 We adopted a previously proposed Graeco-Latin square design but note that the design does not actually completely balance the interface orders.
and discipline (Law, History, Computer Science, Linguistics, Electronic Engineering and Economics). Subjects were informed in advance that they would be reimbursed for participation.

7.4. Significance Tests
Unless otherwise specified, we apply one-way ANOVA (at significance level 0.05) as a parametric measure to test for statistical significance (main effects). Where appropriate, we use Tukey’s HSD tests for post-hoc analysis (at significance level 0.05). For Likert-scale data, we apply Kruskal-Wallis tests with Mann-Whitney U tests for post-hoc analysis also at significance level 0.05.

7.5. Experiment 1: Standard web site versus Single and Multi-Document Profile-Based Summarisation
In this experiment we compare the following three systems:

1. **System A**: a snap-shot copy of the existing site, with no alterations. This is the system that users normally use to find information about the university. It serves as the baseline.
2. **System B**: this system adds a layer of multi-document summarisation (MDS — Section 5.3) on top of **System A**. The summaries are built using the first five outgoing links of the hyperlinked documents.\(^{32}\)
3. **System C**: this is similar to **System B** but uses single-document summarisation (SDS — Section 5.2) of the hyperlinked documents.\(^{33}\)

Both **System B** and **System C** use pop-up tool-tips to display their summaries of hyperlinked documents (as illustrated in Figures 6 and 7). Summaries for the hyperlinked document (either MDS or SDS) are presented as soon as a user hovers the mouse over the link to that document. The summaries disappear when a user moves the mouse away from the link. If a query has been submitted, query-term highlighting is applied to the summaries in **Systems B** and **C**, in line with White et al. [2003], as are the terms extracted from the document title. Note that the university style template is removed before any summarisation is performed.

We compare these three systems in order to test how SDS and MDS compare against the chosen baseline and each other within a single experiment.

In presenting our results, we start with statistics derived from the logs and then look at the questionnaires that our subjects completed.

7.5.1. Subjects. Out of eighteen participants, eight were male and ten were female. Their ages ranged from 19 to 42 (average age: 25.78). Participants were a mix of undergraduate and postgraduate students with different disciplines; six subjects had taken part in previous studies on online searching (but not in any of our studies).

All subjects declared that they use the Internet on a regular basis. The average time subjects have been performing online search is 9.61 years (ten of them between 3 and 20 years). When asked how often they searched for information online, sixteen of the participants selected “daily”. Note that our users (who we would consider typical target users of the system) tend to have a lot of experience using web search systems (mean: 5.00 — on a five-point Likert scale, where 1 means “none” and 5 means “a great deal”) but much less experience using commercial search engines (mean: 2.89).\(^{34}\)

\(^{32}\)ACO title-based MDS (first five documents), the best-performing MDS method in our pilot study

\(^{33}\)ACO title-based SDS, the best-performing SDS method in our pilot study

\(^{34}\)This may look like an anomaly; some explanation is the fact that the relevant question in the TREC-9 Interactive Track entry questionnaire reads like “How much experience have you had searching on commercial online systems (e.g., BRS Afterdark, Dialog, Lexis-Nexis)”
7.5.2. Average Completion Time. Average completion time and number of interactive actions have been commonly used as metrics to compare different interactive information systems [Pitkow et al. 2002; Kelly 2009, for example]. Table V reports the average completion time (and standard deviation) broken down for each task. We measured the time between presenting the task to the users and the submission of the result. There were no cases in which users did not submit an answer.

An analysis of variance showed that the effect of completion time per task was significant (ANOVA $F = 8.01$, $df = 2$, $p < 0.01$). Pairwise post-hoc Tukey tests reveal that two of the comparisons are significant at $p < 0.05$, namely the average time spent on a task on System B and C was significantly shorter than on the baseline.

Table V. Experiment 1: Average completion time (in seconds).

<table>
<thead>
<tr>
<th>System</th>
<th>Task 1 Mean</th>
<th>Task 1 SD</th>
<th>Task 2 Mean</th>
<th>Task 2 SD</th>
<th>Task 3 Mean</th>
<th>Task 3 SD</th>
<th>Task 4 Mean</th>
<th>Task 4 SD</th>
<th>Task 5 Mean</th>
<th>Task 5 SD</th>
<th>Task 6 Mean</th>
<th>Task 6 SD</th>
<th>Overall Mean</th>
<th>Overall SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>166 (73.29)</td>
<td>193 (113.26)</td>
<td>144 (26.31)</td>
<td>255 (137.30)</td>
<td>277 (61.05)</td>
<td>317 (105.85)</td>
<td>225 (112.35)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>161 (62.40)</td>
<td>181 (93.41)</td>
<td>129 (42.09)</td>
<td>209 (65.92)</td>
<td>211 (36.80)</td>
<td>237 (63.38)</td>
<td>188 (72.82)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>164 (35.24)</td>
<td>184 (45.05)</td>
<td>132 (38.85)</td>
<td>203 (90.37)</td>
<td>220 (31.57)</td>
<td>239 (66.34)</td>
<td>190 (65.76)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>164 (59.21)</td>
<td>186 (88.81)</td>
<td>135 (37.03)</td>
<td>222 (104.81)</td>
<td>236 (53.53)</td>
<td>264 (89.03)</td>
<td>201 (87.79)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7.5.3. Average Number of Turns to Finish a Task. We also investigated the number of turns, i.e. the number of steps required to conduct a task (see Table VI). A turn could be viewing a document or inputting a query. On average, users needed 11.51 turns on System A, 9.55 turns on System B and 10.11 turns on System C. There is a main effect in terms of number of turns per task (ANOVA $F = 8.34$, $df = 2$, $p < 0.01$), and pairwise post-hoc Tukey tests reveal the same significant differences as for completion time, i.e. users needed significantly more turns on the baseline system and the fewest number of turns on average on the multi-document summarisation system (marginally fewer than using single-document summarisation).

7.5.4. Task Success. Prior to the evaluation matching documents had been identified by one of the authors. These documents were taken as gold standard to measure task success. In the post-search questionnaire (discussed below) users were asked to state whether they were able to complete their search task successfully. For System A, three answered with ‘No’; for System B and System C, none of them answered with ‘No’. We
also checked task success. The success rate was comparable across all systems. Almost all submitted documents exactly matched the information request as specified by the task (36 on System C, 36 on System B and 33 on System A). Only three of the 108 search tasks did not result in exact matches.

There were two particularly difficult tasks: tasks 5 and 6 (clearly reflected by the user satisfaction values in Table X). Only sixteen of the eighteen users found a correct document for task 6, and seventeen were correctly submitted for task 5. If we look at those two tasks in detail and compare them to the results reported earlier, we find that they have a higher number of turns on average (see Table VI). We also find a higher average completion time for those two tasks (236 seconds for task 5 and 264 seconds for task 6, see Table V).

7.5.5. **Post-Search Questionnaire.** After participants finished each task, they had to fill in a post-search questionnaire. This included the following questions, with answers given on a five-point Likert scale (where 1 indicates “not at all” and 5 indicates “extremely”):

1. “Are you familiar with the search topic?”
2. “Was it easy to get started on this search?”
3. “Was it easy to do the search on this topic?”
4. “Are you satisfied with your search results?”
5. “Did you have enough time to do an effective search?”

Table VII presents the results for question “Are you familiar with the search topic?” Overall, users provided fairly average ratings on the three systems with no significant difference between the systems.

Table VII. Experiment 1: Post-search questionnaire: user familiarity with a search topic, mean scores.

<table>
<thead>
<tr>
<th>System</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
<th>Task 6</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4.16 (0.68)</td>
<td>3.33 (1.10)</td>
<td>3.83 (0.89)</td>
<td>2.50 (1.11)</td>
<td>2.50 (1.38)</td>
<td>4.00 (0.81)</td>
<td>3.38 (1.23)</td>
</tr>
<tr>
<td>B</td>
<td>4.16 (1.06)</td>
<td>4.83 (0.37)</td>
<td>3.50 (1.38)</td>
<td>3.50 (0.76)</td>
<td>2.33 (0.94)</td>
<td>2.50 (1.25)</td>
<td>3.47 (1.34)</td>
</tr>
<tr>
<td>C</td>
<td>2.66 (1.24)</td>
<td>3.16 (1.34)</td>
<td>4.33 (0.74)</td>
<td>3.83 (1.46)</td>
<td>3.00 (1.73)</td>
<td>2.66 (1.24)</td>
<td>3.27 (1.46)</td>
</tr>
<tr>
<td>Overall</td>
<td>3.66 (1.24)</td>
<td>3.77 (1.27)</td>
<td>3.88 (1.09)</td>
<td>3.27 (1.33)</td>
<td>2.61 (1.41)</td>
<td>3.05 (1.31)</td>
<td>3.37 (1.35)</td>
</tr>
</tbody>
</table>

Table VIII presents the results for question “Was it easy to get started on this search?” Overall, users indicated that it was easy to get started when doing the search with no significant difference (though it was marginally easier to use the baseline, as one would perhaps expect given the absence of what some might consider to be “distracting” tool-tips).

Table IX presents the results for question “Was it easy to do the search on this topic?” Overall, users indicated that they found it easy doing the search on the topics on the three systems, with no significant difference between them.
Table VIII. Experiment 1: Post-search questionnaire: ease of getting started, mean scores.

<table>
<thead>
<tr>
<th>System</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
<th>Task 6</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>A</td>
<td>4.33</td>
<td>(0.74)</td>
<td>3.83</td>
<td>(0.89)</td>
<td>3.83</td>
<td>(1.10)</td>
<td>3.66</td>
</tr>
<tr>
<td>B</td>
<td>4.00</td>
<td>(0.81)</td>
<td>4.00</td>
<td>(1.15)</td>
<td>4.00</td>
<td>(0.81)</td>
<td>4.16</td>
</tr>
<tr>
<td>C</td>
<td>3.16</td>
<td>(0.89)</td>
<td>3.50</td>
<td>(0.76)</td>
<td>4.33</td>
<td>(0.74)</td>
<td>4.33</td>
</tr>
<tr>
<td>Overall</td>
<td>3.83</td>
<td>(0.95)</td>
<td>3.77</td>
<td>(0.97)</td>
<td>4.05</td>
<td>(0.84)</td>
<td>3.94</td>
</tr>
</tbody>
</table>

Table IX. Experiment 1: Post-search questionnaire: ease of performing task, mean scores.

<table>
<thead>
<tr>
<th>System</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
<th>Task 6</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>A</td>
<td>4.66</td>
<td>(0.47)</td>
<td>3.33</td>
<td>(1.37)</td>
<td>3.66</td>
<td>(0.47)</td>
<td>2.83</td>
</tr>
<tr>
<td>B</td>
<td>3.83</td>
<td>(0.89)</td>
<td>4.33</td>
<td>(0.74)</td>
<td>3.83</td>
<td>(0.89)</td>
<td>2.50</td>
</tr>
<tr>
<td>C</td>
<td>3.66</td>
<td>(0.94)</td>
<td>3.66</td>
<td>(0.74)</td>
<td>4.50</td>
<td>(0.50)</td>
<td>4.66</td>
</tr>
<tr>
<td>Overall</td>
<td>4.05</td>
<td>(0.91)</td>
<td>3.77</td>
<td>(1.08)</td>
<td>4.16</td>
<td>(0.68)</td>
<td>3.88</td>
</tr>
</tbody>
</table>

Table X gives the results for question “Are you satisfied with your search results?” Overall subjects appear to be satisfied with their results (with no significant differences between the systems).

<table>
<thead>
<tr>
<th>System</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
<th>Task 6</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>A</td>
<td>4.50</td>
<td>(0.50)</td>
<td>3.83</td>
<td>(1.34)</td>
<td>4.00</td>
<td>(0.00)</td>
<td>3.66</td>
</tr>
<tr>
<td>B</td>
<td>5.00</td>
<td>(0.00)</td>
<td>5.00</td>
<td>(0.00)</td>
<td>4.66</td>
<td>(0.68)</td>
<td>4.66</td>
</tr>
<tr>
<td>C</td>
<td>5.00</td>
<td>(0.57)</td>
<td>4.00</td>
<td>(1.15)</td>
<td>4.33</td>
<td>(0.74)</td>
<td>4.66</td>
</tr>
<tr>
<td>Overall</td>
<td>4.83</td>
<td>(0.50)</td>
<td>4.27</td>
<td>(1.14)</td>
<td>4.16</td>
<td>(0.60)</td>
<td>4.33</td>
</tr>
</tbody>
</table>

We had given subjects at most ten minutes to conduct one task. We wanted to see if the time allocated was sufficient. Table XI presents the results for the question “Did you have enough time to do an effective search?” Overall, users indicated that they had enough time when using the three systems, with no significant difference between them.

<table>
<thead>
<tr>
<th>System</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
<th>Task 6</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>A</td>
<td>4.83</td>
<td>(0.37)</td>
<td>4.33</td>
<td>(0.47)</td>
<td>4.33</td>
<td>(0.47)</td>
<td>3.50</td>
</tr>
<tr>
<td>B</td>
<td>4.33</td>
<td>(1.10)</td>
<td>5.00</td>
<td>(0.00)</td>
<td>4.66</td>
<td>(0.47)</td>
<td>4.83</td>
</tr>
<tr>
<td>C</td>
<td>4.50</td>
<td>(0.50)</td>
<td>4.33</td>
<td>(0.74)</td>
<td>4.66</td>
<td>(0.47)</td>
<td>4.83</td>
</tr>
<tr>
<td>Overall</td>
<td>4.55</td>
<td>(0.76)</td>
<td>4.55</td>
<td>(0.59)</td>
<td>4.55</td>
<td>(0.49)</td>
<td>4.38</td>
</tr>
</tbody>
</table>

We also get a picture of the users’ perceptions of the difficulty of the tasks in general without differentiating between the three systems; the results are shown in Table XII.
7.5.6. Post-System Questionnaire. After two search tasks were performed on one system, participants filled in a post-system questionnaire. Table XIII gives a breakdown of the results when asked about the system they had just used. A Kruskal-Wallis test suggests one significant finding only, namely for the question “How easy was it to use this information system?” \( \chi^2 = 10.8941, df = 2, p < 0.001 \). System A is significantly worse than the average at the 95% confidence level. Systems B and System C are significantly better than average. Post-hoc Mann-Whitney U tests with Bonferroni adjustment show that both System B and System C are significantly easier to use than System A (at \( p < 0.05 \)) but there is no difference between System B and System C. This finding is perhaps a bit surprising and certainly worth noting as System A is the standard web site.

The box plots in Figures 8, 9 and 10 provide a different representation of these results.

Table XIII. Experiment 1: Post-system questionnaire, mean scores. Bold values are significantly better and underlined values are significantly worse than the average according to the \( z \)-score with 95% confidence level.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>System A</th>
<th>System B</th>
<th>System C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean ( z )</td>
<td>Mean ( z )</td>
<td>Mean ( z )</td>
<td>Mean ( z )</td>
</tr>
<tr>
<td>How easy was it to learn to use this system?</td>
<td>4.05 (-0.52)</td>
<td>4.11 (0.22)</td>
<td>4.11 (0.29)</td>
</tr>
<tr>
<td>How easy was it to use this system?</td>
<td>3.44 (-4.65)</td>
<td>4.27 (2.59)</td>
<td>4.22 (2.06)</td>
</tr>
<tr>
<td>How well did you understand how to use the sys.?</td>
<td>4.16 (-1.22)</td>
<td>4.33 (0.88)</td>
<td>4.22 (0.33)</td>
</tr>
</tbody>
</table>

---

Fig. 8. Learn to use.  
Fig. 9. Easy to use.  
Fig. 10. Understand to use.
7.5.7. Exit Questionnaire. In the exit questionnaire, users were asked to answer questions concerning the search experience they had in the experiment. Overall, users did understand the nature of the tasks (mean: 4.00, on a scale where 1 represents “not at all”, and 5 represents “completely”); to some extent they found this task similar to other tasks that they typically performed (mean: 3.22); and they found an observed difference between the systems (mean: 4.00).

For the question ‘Which of the three systems did you like the best overall?’, users expressed a strong preference for System B and System C: seven users preferred System C, ten users preferred System B, one user preferred System A, and none found no difference. Furthermore, when asked ‘Which of the two systems did you find easier to use?’, a large majority of users judged that System B and System C were easier to use than the baseline system (System A: one, System B: nine, System C: seven users, no difference: one). Users were also asked ‘Which of the three systems did you find easier to learn to use?’ and the results show no clear preference (System A: four, System B: six, System C: five users, no difference: three).

7.5.8. Concluding Remarks. To summarise the results of this experiment: applying profile-based summarisation to assist users in navigation tasks can significantly outperform a standard web site without such assistance in terms of time and turns needed to conduct a task. Multi-document summaries are marginally better than single-document summaries. We need to note however that the results might be distorted by the fact that the baseline system looks different to the other two systems (which are indistinguishable). Nevertheless, we deliberately used the existing web site because that is the system in actual use, and it represents the most natural/common way of navigating a web site.

In order to address the issue of the variable presentation format, and to assess the impact of the personalisation of the summaries, we conducted an additional user study.

7.6. Experiment 2: Generic versus Single and Multi-Document Profile-Based Summarisation

We adopted the same experimental setup as in Experiment 1, but with a different baseline System A. In this experiment, all the three systems looked almost identical to the user, more specifically:

1. System A: adds a layer of centroid-based single-document summarisation of hyperlinked documents, presented as pop-up tool-tips over the existing site. This algorithm is designed for traditional summarisation. It is a widely used baseline [Radev et al. 2004, for example].
2. System B: as before (ACO title-based MDS (first five documents)).
3. System C: as before (ACO title-based SDS).

Figure 1 is a screen shot of System A. While there is still query-term highlighting if a query has been submitted, unlike System B and System C (Figures 6 and 7) there is no highlighting of terms extracted from the document title.35

7.6.1. Subjects. Out of eighteen participants, thirteen were male and five were female. All subjects declared that they use the Internet on a regular basis. The average time subjects have been doing online searching is 12.33 years (nine of them between 3 and 17 years). When asked how often they searched for information online, all of the participants selected “daily”.

7.6.2. Average Completion Time and Number of Turns. Table XIV gives a picture of the average completion time. As in the previous study, it takes significantly longer on the

---

35This difference stemmed from the nature of the summarisation algorithms.
baseline system (System A) to conduct a task than on any of the other two systems despite all three systems providing pop-up summaries (ANOVA $F = 10.31, df = 2, p < 0.01$ followed by pairwise post-hoc Tukey tests at $p < 0.05$).

Table XIV. Experiment 2: Average completion time (in seconds).

<table>
<thead>
<tr>
<th>System</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
<th>Task 6</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>A</td>
<td>157 (35.84)</td>
<td>223 (71.89)</td>
<td>164 (26.94)</td>
<td>240 (72.04)</td>
<td>337 (73.26)</td>
<td>395 (85.04)</td>
<td>253 (108.33)</td>
</tr>
<tr>
<td>B</td>
<td>131 (34.85)</td>
<td>191 (82.68)</td>
<td>123 (37.89)</td>
<td>199 (49.92)</td>
<td>236 (49.54)</td>
<td>257 (50.57)</td>
<td>189 (72.54)</td>
</tr>
<tr>
<td>C</td>
<td>132 (36.17)</td>
<td>194 (32.60)</td>
<td>120 (26.62)</td>
<td>223 (70.37)</td>
<td>237 (66.70)</td>
<td>269 (79.38)</td>
<td>196 (77.95)</td>
</tr>
</tbody>
</table>

Table XV. Experiment 2: Average number of turns to complete a task.

<table>
<thead>
<tr>
<th>System</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
<th>Task 6</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>A</td>
<td>6.78 (1.49)</td>
<td>10.83 (3.38)</td>
<td>8.33 (1.97)</td>
<td>11.83 (0.68)</td>
<td>13.00 (1.91)</td>
<td>16.00 (2.88)</td>
<td>11.13 (3.59)</td>
</tr>
<tr>
<td>B</td>
<td>5.64 (1.49)</td>
<td>9.00 (3.30)</td>
<td>5.50 (1.11)</td>
<td>10.16 (1.77)</td>
<td>11.50 (1.60)</td>
<td>13.50 (4.03)</td>
<td>9.21 (3.59)</td>
</tr>
<tr>
<td>C</td>
<td>6.01 (2.13)</td>
<td>10.16 (2.11)</td>
<td>5.66 (2.28)</td>
<td>10.00 (2.08)</td>
<td>11.50 (2.21)</td>
<td>13.50 (3.50)</td>
<td>9.47 (3.63)</td>
</tr>
</tbody>
</table>

Similar to the previous experiment, a significantly shorter completion time goes hand in hand with significantly fewer turns (ANOVA $F = 28.88, df = 2, p < 0.01$ followed by pairwise post-hoc Tukey tests at $p < 0.05$). Task success is identical in all three systems (with correct answers in all cases).

It looks like System A in this study – Centroid SDS – had consistently worse completion times than the system without pop-ups in the first study – the standard Web site without summarisation support – (cf. Tables V and XIV). These summaries may have required more effort to parse due to the summarisation method used. It also could be that users could not scan them easily for the keywords since there was no highlighting. Since this difference is not present in the number of turns (cf. Tables VI and XV), that adds evidence to think that participants had to spend time trying to read and understand the centroid-based summaries.

7.6.3. Post-Search Questionnaire. Table XVI presents the results for question ‘Was it easy to get started on this search?’. The lower average ratings for each task on the baseline system suggest that it was slightly easier to get started on System B and System C, but there was no main effect.

Table XVI. Experiment 2: Post-search questionnaire: ease of getting started, mean scores.

<table>
<thead>
<tr>
<th>System</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
<th>Task 6</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>A</td>
<td>4.50 (0.50)</td>
<td>3.66 (0.74)</td>
<td>3.83 (0.89)</td>
<td>3.66 (0.94)</td>
<td>3.50 (0.76)</td>
<td>2.83 (0.68)</td>
<td>3.66 (0.91)</td>
</tr>
<tr>
<td>B</td>
<td>4.66 (0.47)</td>
<td>4.00 (1.15)</td>
<td>4.33 (0.47)</td>
<td>4.33 (0.47)</td>
<td>3.83 (0.37)</td>
<td>3.16 (0.68)</td>
<td>4.05 (0.81)</td>
</tr>
<tr>
<td>C</td>
<td>4.66 (0.47)</td>
<td>3.83 (0.68)</td>
<td>4.50 (0.50)</td>
<td>4.33 (1.10)</td>
<td>3.66 (0.94)</td>
<td>3.16 (0.68)</td>
<td>4.02 (0.92)</td>
</tr>
</tbody>
</table>

Table XVII includes the results for question ‘Was it easy to do the search on this topic?’ The tendency is the same (slightly lower ratings for System A) but again there is no statistical significance.
7.6.4. Post-System Questionnaire. In the post-system questionnaire a Kruskal-Wallis test again indicated a significant result for the question “How easy was it to use this information system? ” ($\chi^2 = 18.6279$, df = 2, $p \ll 0.001$). Post-hoc Mann-Whitney U tests with Bonferroni adjustment show that users found System B and System C significantly easier to use than System A (at $p < 0.05$), no other differences are significant (see Table XXI and box plots in Figures 11, 12 and 13). This confirms the finding of the first experiment in that profile-based summarisation outperforms a sensible baseline.
Table XXI. Experiment 2: Post-system questionnaire, mean scores. Bold values are significantly better and underlined values are significantly worse than the average according to the z-score with 95% confidence level.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>System A</th>
<th>System B</th>
<th>System C</th>
</tr>
</thead>
<tbody>
<tr>
<td>How easy was it to learn to use this information system?</td>
<td>4.00 (-1.36)</td>
<td>4.16 (0.68)</td>
<td>4.16 (0.68)</td>
</tr>
<tr>
<td>How easy was to use this information system?</td>
<td>3.50 (-5.96)</td>
<td>4.66 (4.11)</td>
<td>4.44 (1.84)</td>
</tr>
<tr>
<td>How well did you understand how to use the inf. sys.?</td>
<td>4.50 (-1.22)</td>
<td>4.56 (1.21)</td>
<td>4.56 (0.01)</td>
</tr>
</tbody>
</table>

7.6.5. Exit Questionnaire. In the exit questionnaire, users overall strongly preferred System B and System C: no user preferred System A, ten users preferred System B, six users preferred System C, and two found no difference. Furthermore, a large majority of users judged that System B and System C were easier to use than the baseline system (System A: none, System B: eight, System C: seven users, no difference: three). There was no difference between the three systems in the ease of learning to use (System A: none, System B: two, System C: one user, no difference: fifteen). Overall, users understood the nature of the tasks (mean: 3.88); to some extent they found this task similar to other tasks that they typically performed (mean: 2.94); and they found marginal difference between the systems (mean: 3.72).

7.6.6. Concluding Remarks. The second task-based evaluation study on navigation support largely confirms the results of the first study: we have a measurable benefit of using cohort-personalised summaries when using a baseline that is visually almost identical except for the content of the summaries. Furthermore, the results indicate that using the cohort-profile to generate “personalised” summaries has a measurable benefit over using a sensible summarisation baseline when assessed in terms of the time taken, and the users’ overall preference.

7.7. Discussion

7.7.1. General Observations. The results of the two experiments indicate that applying profile-based summarisation to assist users in navigation tasks can significantly outperform generic summarisation as well as a standard web site without such assistance. The results of the task-based evaluations indicate that multi-document summaries are marginally better than single-document summaries although the difference was more marked in the pilot study.

In the first experiment we used three systems that are perhaps not directly comparable; the baseline system looks slightly different to the other two systems (which are
indistinguishable). That was the reason to apply an alternative baseline in the second experiment, namely centroid-based summarisation.

To check for consistency across the two studies we conducted paired t-tests comparing average completion time and average number of turns for System B and System C (the systems used in both experiments) and found no significant differences (at $p < 0.05$) when comparing the overall results for the corresponding systems.

7.7.2 Comparison with Related Work. It is difficult to make direct comparisons between our findings and those of previous studies. To the best of our knowledge, there has been no comparable study that addresses profile-based navigation on a web site. In a web search context there have been many studies (see Section 2) but the setup differs in at least two important dimensions, namely the document collection and the mode of search. However, one interesting finding in a web search study aimed at showing results in context concludes that inline summaries were more effective than summaries presented as hover text [Dumais et al. 2001]. One of our basic assumptions has always been to not touch the structure of the site at all but simply add a layer on top of it. However, perhaps inline summaries could be investigated as an alternative in future, also because even with adding an overlay it is not necessarily guaranteed that one is not intercepting with the content owner's javascript and hence altering the presentation to a certain degree anyway.

There have been other studies on the same web site. In a navigation context, the suggestion of commonly visited pages (derived from query and click logs) as an overlay window was shown to cut down task completion time and is preferred by users over the unaltered web site [Saad and Kruschwitz 2011], a follow-up study applied an ACO approach (to click graphs rather than query graphs as we do) and this was shown to outperform a baseline that suggests links simply based on popularity [Saad and Kruschwitz 2013]. A different study on the same web site adopts a search context [Adindla and Kruschwitz 2013] whereby query support is offered through a conceptual graph derived from the syntactic relationships extracted from the document collection. That study applied exactly the same tasks we employed and apart from the finding that search assistance did help overall it is worth noting that the average completion time of the baseline system is virtually identical to ours and that Tasks 5 and 6 were also found to be the most difficult ones.

7.7.3 User Feedback. We primarily focussed on a statistical analysis but would like to report some feedback received from users in the exit questionnaires. We observe that there was a common theme in that many users found the summarisation systems to be very similar, e.g. in the first experiment “there should be no big difference between the two” and “I found System C very similar to System B”, and when compared against a centroid-based baseline “I think that all the systems were the same” and “I didn’t notice a very big difference between these systems”.

Regarding the content and presentation format of the summarisation-based systems users liked the highlighting of keywords, liked the overall idea (“[...] provides better information when moving the mouse over any link. This helps me to decide on clicking the link or not.”), the ease of use and potential to save time. They also noted problems, the potential for confusion was mentioned several times, e.g. “additional information [...] could be off-putting and confusing.” and “The page contains a lot of links which creates a lot of ambiguity and misconfusion [sic]”. One user noticed “Sometimes in Systems B & C the help box did not provide useful information.” One user suggested: “It’s more fun if we have colorful windows popping up without so many words crammed

---

36Unlike in the first study where only three subjects found no difference in learning to use the different systems in the second study a large majority of fifteen subjects found no difference between the systems.
together. These words are too small and boring sometimes." Perhaps future work in this
direction should explore moving away from extractive summaries and, for example,
use term clouds instead [Jones and Li 2008]. Technical issues included “The clouds
with the comments could load faster”.

Finally, we also received some feedback that supports the observation that web sites
are difficult to navigate, e.g. “I found parts of the university web site difficult to
search for specific information [...] I found it easiest just to type the information [...] rather than through the menus and sub-menus”. In line with this, someone suggested
to include more direct links for new students on the university home page which fits
very much the overall idea of our work. Regarding the overall experience one comment
reads “nice systems making search on the uni website easier”.

We also received some evidence that our tasks were realistic and targeted at the
right user group, e.g. “The tasks are well chosen, because they are close to our life as
students. It’s an interesting experience” and somebody else commenting “I learned some
new information about university funding”.

7.7.4. Limitations. There are a number of limitations. First of all, task-based evalua-
tions aim to model realistic tasks as closely as possible but they still remain somewhat
artificial. More extensive tests with real users in realistic contexts will have to be con-
ducted to validate any of the findings of this study.

An obvious limitation of such a study is that the results are based on data from a
single web site and the findings may or may not be transferable to other document col-
clections, the same caveat was raised when assessing the quality of query modification
suggestions on the same web site [Kruschwitz et al. 2013].

This is also true for any comparison of the results with studies conducted on web
logs. Site logs may have very different characteristics. For example, web queries tend
to be around 2.35 words long on average [Silverstein et al. 1998; Jansen et al. 1998;
Beitzel et al. 2004; Beitzel et al. 2007, for example], our queries are shorter, on average 1.81 query terms [Kruschwitz et al. 2013] which is consistent with other results
reported on web sites and intranets [Stenmark and Jadaan 2006]. As a result the query
logs collected on a web site will likely be very different from web query logs. This also
affects the average session length (and hence the profile acquisition process), in our
case 1.53 queries per session, 1.73 on a different web site (the Utah government web
site) [Chau et al. 2005] but more than two queries on average for web logs [Silverstein
et al. 1998; Jansen et al. 1998; Jansen et al. 2007].

The major bottleneck in conducting research into using any form of query logs is the
difficulty in obtaining realistic and large-scale log data which is also the reason why
it is nearly impossible to conduct and report studies on a selection of large-scale logs
collected on different sites. For this study we used the actual log files of a reasonably
sized university web site collected over three years. We therefore hope that our results
could serve as a benchmark for future studies on different web sites and using different
profiling approaches.

7.7.5. Summary. The main conclusion that we derive from the statistical evidence is
that profile-based document summarisation can lead to significant measurable ben-

ACM Transactions on Information Systems, Vol. 0, No. 0, Article 0, Publication date: August 2014.
outperform sensible baseline approaches but the findings merit further investigations to drill down further. This is also true when it comes to the issue of different cohorts of users: in this study we provided a framework for cohort-based summarisation to assist in navigation scenarios, but have not evaluated the impact of using models derived from different cohorts.

7.7.6. Future Directions. There is much scope for future work. One major route is to exploit the adaptive nature of the ACO-based profile. In the presented work we took a snapshot of the profile and did not try to exploit the benefits of the ACO model as a continuously updated representation of the cohort’s collected search knowledge. This will need to be done in longitudinal studies that capture an element of time. It has already been demonstrated that the ACO-based model has the potential to improve over time [Albakour 2012, for example]. There is certainly a need to adapt as illustrated by rebranding exercises and the introduction of new terms.

7 Note also that the outlined approach should need no customisation to new web sites and can be applied out-of-the-box. This avoids the problem of having to manually or semi-automatically adapt general-purpose knowledge structures to a new collection or domain. It would be interesting to see the approach applied to sites of different types and sizes.

Two other major directions would include clickthrough information in the model acquisition process and the application of the profile to finer-grained cohorts of users. Both of these aspects are straightforward and simply assume that clickthrough information is available and an organisational structure is in place, respectively.

Alternatives to applying an ACO approach for the construction of the profile are also possible, e.g. the use of query flow graphs which have shown to be effective in deriving query suggestions in a web site context [Kruschwitz et al. 2013]. Random walk models can be explored in place of the simplistic selection of related concepts as used in our summarisation approaches.

8. CONCLUSIONS

We presented the use of profile-based summarisation for navigating a local web site. The profile was acquired by exploiting past users’ searching patterns in an attempt to capture the ‘knowledge’ of previous users in order to help a new one who might have the same information need, a form of group-based context we try to exploit. We can now return to the initial research questions and answer them using the evidence from the studies we conducted.

(1) Can web site navigation benefit from the automated summarisation of results? We found evidence that summarisation does have potential benefits when applied to a web site. We also found that profile-based summarisation outperformed a range of alternative approaches. We find that the profile-based summarisation approach outperformed the baseline systems which included the web site without any support (the existing default web site) as well as a site that employed generic summarisation, i.e. one that did not incorporate information about past user search patterns.

(2) Will a domain model/profile capturing the search behaviour of a group of users be beneficial for the summarisation process?

In all our studies we found that a model that employs past users’ search paths and patterns significantly improves the quality of the summarisation process for the

[37] Two examples of commonly used concepts on the web site used in this study are the term Freshers’ Week which has been replaced by Welcome Week, and the term FASer (Feedback Assessment Submission electronic repository) which has replaced OCS (Online Coursework Submission system).
given scenarios. We have employed fairly simple methods both for acquiring the profile as well as profile-biased summarisation. There is much scope for further exploration, including the use of clickthrough data in the model construction process, the utilisation of the adaptive nature of the model, as well as the characterisation and modelling of more fine-grained user groups.

(3) *Will such methods result in measurable (quantifiable) benefits such as shorter sessions, fewer interactions etc?* Based on two task-based evaluations we can conclude that we can indeed assist users in finding information more quickly and in fewer interaction steps. However, to validate the findings in a realistic context one needs to test the methods on real users in real time, e.g. via A/B testing [Kohavi et al. 2007], which will then also test less frequent queries.

We conclude that overall there is much potential in mining past user query logs to build a model that reflects not individual user needs but collective needs, by way of some form of community profile. Using an ant colony optimisation analogy allows the model to learn patterns over time but also forget them (the adaptive nature of the model has not been investigated in this paper but that opens interesting avenues for future work). There are many avenues for future work. We hope this study will serve as a benchmark for future investigations into web-site navigation.

**Acknowledgements**

We would like to thank the five anonymous reviewers as well as the editors of this special issue for substantial, insightful and very constructive feedback (we received seventeen pages of reviews in the first round and another eleven page in the second round!). This feedback has helped us to make major improvements to the paper and we hope this final version has addressed most of the concerns and suggestions.

**REFERENCES**


A. Alhindi et al.

national ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’12).
ACM, New York, NY, USA, 185–194. doi: http://dx.doi.org/10.1145/2349283.2348312


Profile-Based Summarisation for Web Site Navigation


ACM Transactions on Information Systems, Vol. 0, No. 0, Article 0, Publication date: August 2014.


