

Interaction-based Information Filtering for Children

Richard Glassey, Desmond Elliott, Tamara Polajnar and Leif Azzopardi
Information Retrieval Group, Department of Computing Science
University of Glasgow, Glasgow, G12 8QQ
{rjg, delliot, tamara, leif}@dcs.gla.ac.uk

ABSTRACT

This paper presents an interaction-based information filtering system designed for the needs of children accessing multiple streams of information. This is an emerging problem due to the increased information access and engagement by children for their education and entertainment, and the explosion of stream-based information sources on most topics.

It has been shown that children have difficulties formulating text-based queries and using interfaces primarily designed for adults. The in-progress system presented in this paper attempts to address these difficulties by employing an interaction-based interface that simplifies the expression of information needs and adapts itself to user interests over time. To overcome issues of content moderation, the system aggregates multiple child-friendly information feeds and performs offline processing to facilitate topic filtering. A set of standing topics are created for initial interaction and subsequent interactions are used to infer and refine which topics the child would most likely want to have presented. A simple and easy-to-use interface is presented which uses relevance information to determine the appropriate size of the document title to display to act as a relevance-cue to the user.

The planned research focuses on validating the interaction-based approach with both child and adult populations to discover the differences and similarities that may exist.

Categories and Subject Descriptors

H.5 [Information Interfaces and Presentation]: User Interfaces—*Interaction styles*; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Information Filtering*

General Terms

Design, Human Factors

Keywords

Information Filtering, Children

Permission to make digital or hard copies of all or part 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 bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

IJIS 2010, August 18–21, 2010, New Brunswick, New Jersey, USA.
Copyright 2010 ACM 978-1-4503-0247-0/10/08 ...\$10.00.

1. INTRODUCTION

Children are increasingly using the Web as a source for their education and entertainment, alongside more traditional sources such as printed books and television [11]. Most concerns regarding children using the Web have understandably focused upon effective moderation of content. More recently, it has been suggested that children require age-appropriate search interfaces and information presentation as much as they require moderated content [8]. Furthermore, studies comparing children's and adult's information seeking experience have also revealed the similarities (preference for keyword searching, but more successful when browsing subject hierarchies) and differences (navigation efficiency, use of browser shortcuts and ability to recover from keyword 'breakdowns' during their task) [3].

The challenges that children face when seeking information are manifold. Firstly, children have difficulty using query-based interfaces. This difficulty has two aspects: there are cognitive barriers in query formulation (not necessarily unique to children); and a lack of typing dexterity results in children having difficulty observing the screen and keyboard at the same time [8]. This leads to typographical errors and failure to notice events such as suggested queries. Secondly, children have a lower attention span and experience difficulty in judging the relevance of a document, which causes them to browse more erratically than adults. This leads to frequent revisiting of previous pages or 'looping', as well as spending less time reading a page [3]. Finally, it has been shown that children prefer a personalised interface that reflects their personality and interests [2, 10, 9], thus creating a more engaging experience.

In response to these challenges, this paper proposes an interaction-based information filtering system designed for children. Information filtering is the task of selecting documents from an incoming stream against a set of persistent information needs [13, 1]. This task has been extensively studied in a simulated environment via the Text REtrieval Conference Filtering Track and less so in interactive environments. In general, a set of topic representations are generated, documents are processed as they are received and a binary decision is made to filter a document for a topic. The filtering decisions can be used to update topic representation or filtering threshold [1]. Simulated studies, using static document collections from single-source information providers, have focused on advances in topic representation [6] and filtering threshold adaptation [15]. In an interactive environment, using dynamic data collections from multiple-source information providers on the Web, studies have fo-

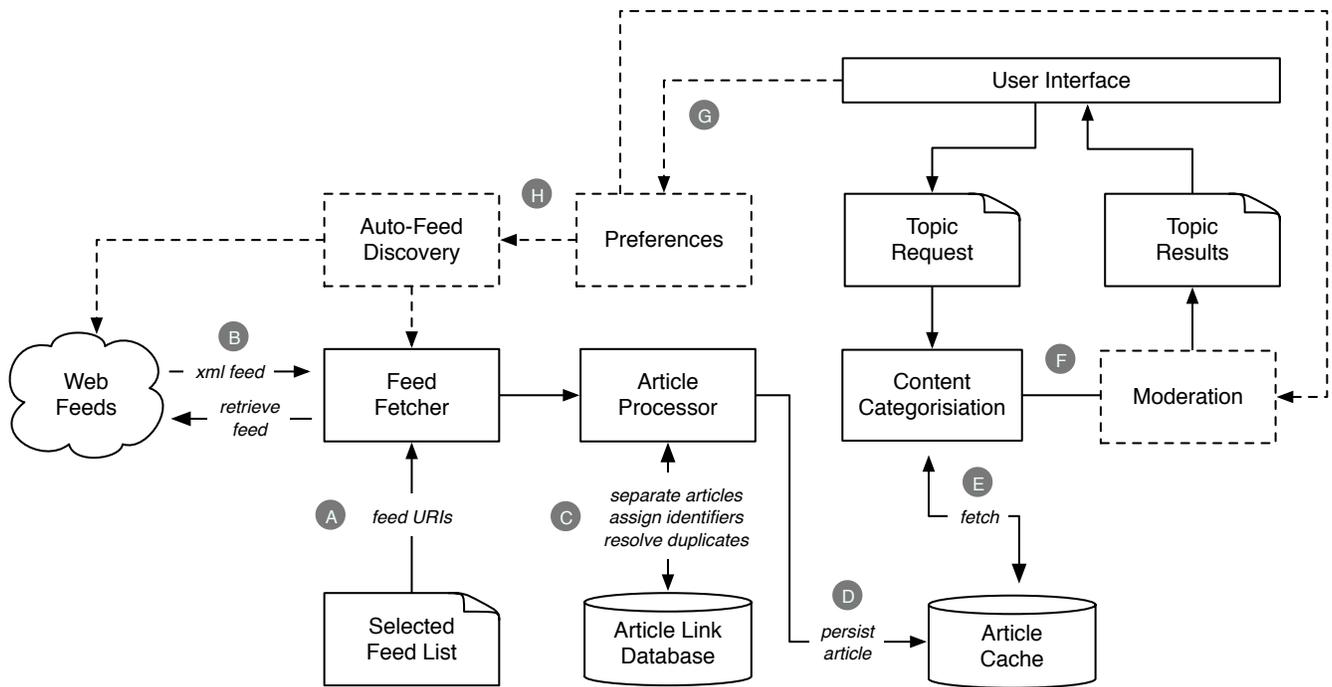


Figure 1: Information filtering prototype architecture. Dashed components are currently part of planned research discussed in Section 3.

cused on validating the acceptance of topic representations and the documents presented to users [12, 14].

Information filtering specifically for children has not been explored nor reported extensively within the literature, but provides an opportunity for supporting their persistent information needs whilst facilitating information encountering [7]. The need for an interaction-based information filtering system can be illustrated by an example involving Maddie, a nine year old girl. She has specific news topics she likes to follow regularly and occasionally needs to supplement the available school resources with newer findings reported on the Internet. Like many girls her age, Maddie likes to follow documents related to her favourite movies and TV shows. She is a particular fan of *High School Musical* and *Twilight*. These serial movies have a continuously updating stream of information relating to rumours of new productions, events, and the cast. Instead of issuing text queries for these movies on a regular basis, including the individual actors and actresses, and trawling through celebrity news sites¹ in order to find this specific information; Maddie would prefer to be able to go to her own personal application and be instantly presented with the latest news on her favourite topics.

This paper aims to address the needs of children like Maddie by designing an interaction-based information filtering system that adapts to the characteristics, demographics, and personal interests of users over time. For example, a rich and engaging interface containing content with more visual information (picture, video and animation) relative to textual information may be preferred by younger children. Textual information itself may be filtered using readability measures to ensure that children do not become frustrated by incom-

prehensible content. Furthermore, the layout and presentation of the user interface should simplify access to the relevant information that children find interesting, without requiring online help or assistance from an adult.

The remainder of this paper describes the progress in developing the information filtering system, in terms of its backend system and user interface, and highlights the proposed research arising from the work.

2. SYSTEM OVERVIEW

The architecture of the system presented in this paper is shown in Figure 1 and is split into four major components: content discovery and pre-processing; content categorisation; user representation; and content presentation. Each of these components is described in more detail throughout this section.

2.1 Discovery and Pre-Processing

The components for content discovery and pre-processing can be seen in Figure 1. Information sources (A) are defined and stored as a list of syndication feed addresses. Each information source is periodically retrieved from the Web using the ROME framework² and split into the individual feed entries (B). Each entry is then checked against a database³ of known entries to determine whether an incoming entry is unique (C). If an incoming entry is unique, a new row is written to the database including the entry URL and the timestamp of the entry using an ISO 8601 compliant format⁴. The ID of this new row is assigned to represent the

²<http://rome.dev.java.net/>

³<http://www.sqlite.org/>

⁴YYYYMMDDTHHMMSS

¹a path potentially fraught with inappropriate content given the media's penchant for reporting every last celebrity detail.

entry position in the order of entries retrieved. The content of the entry is written to disk (**D**) to be processed by the content categorisation component. The format of the content written to disk is shown below:

```
<entry>
  <number>16</number>
  <title>A Title</title>
  <link>http://www.link.com/link.html</link>
  <source>http://www.link.com/link.atom</source>
  <dateTime>20100329T144033</dateTime>
  <content>Some content.</content>
</entry>
```

The entry number is determined as described above, whilst the title, link, source, and content are obtained from the raw entry, and the *dateTime* element is retrieved from the database. Some information sources include full HTML in the content of their entries, which is kept in the data written to disk⁵, however, these elements can be transformed for presentation or other types of data-processing.

2.2 Content Categorisation

The entries written to disk by the content discovery and pre-processing component have no predefined categories. We propose two methods for content categorisation: one based on information retrieval and one based on machine learning. The information retrieval approach uses topic terms as a representation of the persistent information need (or interest) and retrieves a ranked list of documents matching the terms using a standard scoring function, such as cosine similarity or Okapi BM25 (**E**). This ranked list of documents, however, represents the probability of a document being relevant for a given set of topic terms. The documents are re-ranked by document number, to represent the time-sensitive nature of news filtering, whilst maintaining the similarity scores to incorporate as relevance-cues in the presentation component.

An alternative approach would use machine learning to automatically induce categories from the unlabelled documents using clustering or topic modelling techniques, such as latent Dirichlet allocation [5]. These methods, however, generally result in a predefined number of clusters whose labels may be difficult to interpret. In order to avoid this problem, we propose to use the Open Directory Project hierarchy to train a classifier on the documents contained therein and their associated manually-assigned labels. For this purpose we propose to use a state of the art classifier such as the support vector machine or the Gaussian processes classifier [4]. These methods can also be used for semi-supervised learning [4], where we could combine the ODP labelled and the incoming unlabelled documents to retrain the classifier to address topic drift over time.

2.3 User Representation

Users are represented anonymously as an IP address and a list of weighted topics based on interactions with the documents presented in the interface. The goal of user representation is not to profile the user, but to model their interactions with topics of interest, and general characteristics to guide moderation (**F**). In particular, we propose to use the interface and relevance judgements to represent the user's

⁵potential further research may exploit the link structure in entries to improve categorisation.

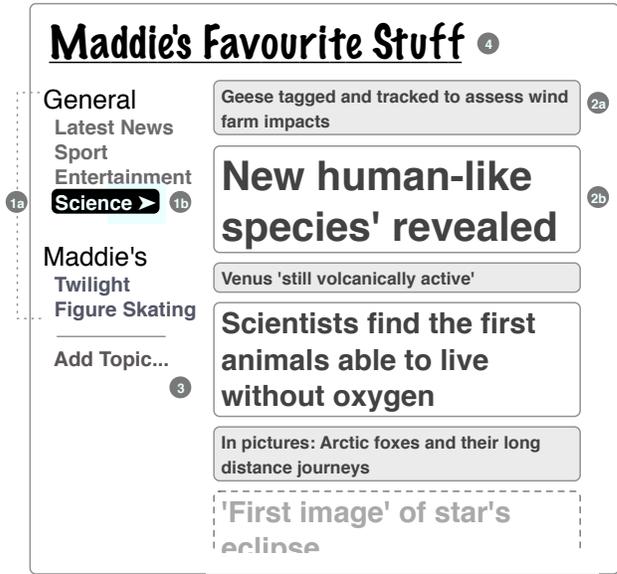


Figure 2: The interface demonstrates the content filtering and relevance cues, based on browsing behaviour.

requirements for specific information. The user model will keep track of the preference for specific topics, but will also customise the topic representations according to the users' interactions with the documents within the topics (**G**). The topics will be adapted using click-through data.

2.4 Content Presentation

The presentation of content seeks to address children's information needs in an age appropriate manner. In this section we describe one of the possible layouts that addresses the main concerns raised in studies observing children interacting with traditional search engine interfaces [8, 9].

Figure 2 shows the interface layout that attempts to ease the cognitive load by minimising query-based interaction.

The layout facilitates topic browsing by listing topics ranging from general to the more personalised (**1a**). In this example, the *Science* topic has been selected (**1b**) and the list of document titles in the collection is filtered for this topic, ordered by time, are presented to the child. The amount of space allocated to a title indicates the relevance of a document, as shown by comparing a somewhat relevant document title (**2a**) and a strongly relevant one (**2b**). Topics such as *Twilight* can be automatically induced by observing user interaction over time, or by manually adding a *topic request* in the text area (**3**) should the information need be more immediate (e.g. a sudden interest in a new topic, e.g. a recent natural disaster or a breaking celebrity story).

Finally, the banner area can be personalised to encourage a sense of *ownership* (**4**) and the overall look and feel can be modified by choosing from alternative themes (e.g. gender and age-specific colour schemes and graphical cartoon characters or avatars).

The aim of this interface is to de-emphasise the need for query-based search and to allow children to browse based on their topics of interest. By presenting only the document titles, the need to read is reduced, while the variable font size draws attention to more relevant documents. Clicking the

titles will reveal a snippet of the document, using a pleasing animated *drop-and-bounce* effect, and clicking a *more* link will load the the web page. At this stage, the interface is replaced by the web page, however an inner frame could be used, ensuring the user can easily return to their browsing.

3. PLANNED RESEARCH

There are three strands of planned research emerging from the development of the system described in this paper. They are (i) a comparative evaluation of categorisation techniques; (ii) validation of the interaction-based user modelling approach; and (iii) development of an automated content discovery method (based on user modelling). We plan to study the effectiveness of the different categorisation approaches proposed in a simulated environment to help ensure a high-quality information filtering system; adopting the methodology used in the TREC Filtering Track for this purpose.

We also plan to validate the interaction-based approach to user modelling in a user study with children from the 9 – 12 year old demographic. This will also help us evaluate the extent to which this approach addresses the issues raised in the Introduction that children experience. Furthermore, we also plan to study this approach to user modelling with an adult demographic to determine the similarities and differences in experience, and how the approach can be generalised, contributing to the wider issues of providing simple and effective access to personally relevant ‘streams of information’.

Finally, we would like to explore the challenge of developing automated content discovery, instead of using manually selected sources of content. In effect, the system would be capable of continuously searching the Web for appropriate sources of child-friendly information streams tuned to a particular user. As indicated in Figure 1, the Auto-Feed Discovery component (**H**) will be influenced by the user modelling Preference component, with the aim of ensuring that a user’s transient and persistent information needs are satisfied by the system.

4. DISCUSSION

The interaction-based information filtering system presented in this paper attempts to address some of the issues children face when using information access systems. The opportunities and challenges are unique and have received little attention in the information filtering and retrieval literature, despite children having increased access with information systems and the recent explosion in online streams of information available to them.

We have presented a detailed overview of the in-progress system and outlined plans for planned research, in particular a focus on: user interaction modelling, user evaluation and automated discovery and detection of child-friendly information sources. This initial system will then be iteratively improved by involving children in subsequent design phases.

Future work will then focus on exploring what models of interaction emerge from children using a filtering system, e.g. is it truly an unmet need for children and does their interaction align with the similarities and differences discovered when children and adults are compared conducting information seeking tasks [3].

Acknowledgements: This research is funded by the European Community’s Seventh Framework Programme FP7/2007-2013 under grant agreement no. 231507.

5. REFERENCES

- [1] N. J. Belkin and W. B. Croft. Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM*, 35(12):29–38, 1992.
- [2] D. Bilal. Draw and tell: Children as designers of web interfaces. *Proceedings of the 66th Annual Meeting of the American Society for Information Science & Technology*, 40(1):135–141, 2003.
- [3] D. Bilal and J. Kirby. Differences and similarities in information seeking: children and adults as web users. *Information Processing & Management*, 38(5):649–670, 2002.
- [4] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.
- [5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [6] J. Callan. Document filtering with inference networks. In *SIGIR ’96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 262–269, 1996.
- [7] S. Erdelez. *Information Encountering, Theories of Information Behaviour*, pages 179–184. American Society for Information Science and Technology, 2005.
- [8] H. E. Jochmann-Mannak, T. W. C. Huibers, and T. J. M. Sanders. Children’s Information Retrieval. In *Proceedings of Future Directions in Information Access 2008*, pages 64–72, September 2008.
- [9] A. Large and J. Beheshti. Interface design, web portals, and children. *Library Trends*, 54(2):318, 2005.
- [10] A. Large, J. Beheshti, and T. Rahman. Design criteria for children’s web portals: The users speak out. *Journal of the American Society for Information Science and Technology*, 2(53):79–94, 2002.
- [11] A. Large, V. Nasset, and J. Beheshti. Children as information seekers: what researchers tell us. *New Review of Children’s Literature and Librarianship*, 14(2):121–140, 2008.
- [12] J. Liu, P. Dolan, and E. R. Pedersen. Personalized News Recommendation Based on Click Behavior. In *IUI ’10: Proceedings of the 2010 International Conference on Intelligent User Interfaces*, pages 31–40, 2010.
- [13] N. Nanas, A. Roeck, and M. Vavalis. What Happened to Content-Based Information Filtering? In *ICTIR ’09: Proceedings of the 2nd International Conference on Theory of Information Retrieval*, pages 249–256. Springer-Verlag, 2009.
- [14] J. wook Ahn, P. Brusilovsky, J. Grady, D. He, and S. Y. Syn. Open user profiles for adaptive news systems: help or harm? In *WWW ’07: Proceedings of the 16th International Conference on World Wide Web*, pages 11–20, 2007.
- [15] Y. Zhang and J. Callan. Maximum likelihood estimation for filtering thresholds. In *SIGIR ’01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 294–302, New York, NY, USA, 2001.

PuppyIR

Interaction-based Information Filtering for Children

**Richard Glassey
Desmond Elliott
Tamara Polajnar
Leif Azzopardi**

Introduction

This poster presents an interaction-based information filtering system designed to meet the needs of children. It has been shown that children have difficulties using interfaces primarily designed for adults.

- Children have **difficulty using query-based interfaces**.
- Children have a **lower attention span and experience difficulty in judging the relevance of a document**, which causes them to browse more erratically than adults.
- Children tend to prefer a personalised interface** that reflects their personality and interests.

The system presented here attempts to address these difficulties by employing an interaction-based interface that simplifies the expression of information needs and adapts itself to user interests over time based on interactions.

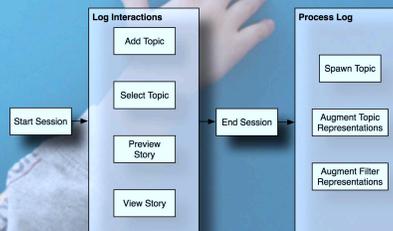


Frontend

- The layout facilitates topic browsing by listing a few general categories followed by an augmentable list of personalised topics (a); and the current topic is highlighted (b).
- The content is located in the middle portion of the screen. The titles of relevant stories are ordered by time (a), while the amount of relevance is indicated by the size of the title font (b).
- Topics such as *Twilight* are induced through interaction or explicitly requested.
- The banner area can be personalised to encourage a sense of *ownership* and of the overall look and feel can be modified by choosing from alternative themes.

Interaction-based feed browsing

The increase in the availability of up-to-date information provided through syndicated feeds, induces the need for safe browsing environments designed specifically for children. Our system aims to provide topic-relevant information that can evolve through time with user interests and news content. In order to reduce the reliance on keyword queries, the system keeps track of user interactions and uses the content of the viewed stories to recalculate relevance of incoming information. The representation of the topics is augmented each time a user closes a session. In addition new topics are spawned from the general ones if the user shows interest in a particular emergent thread.



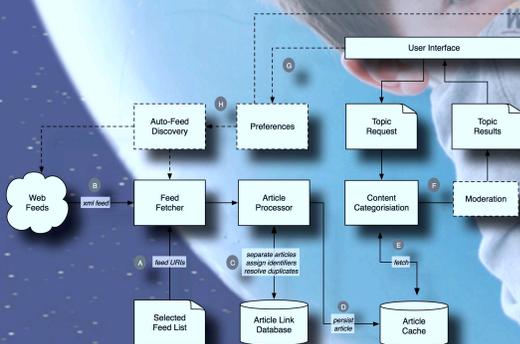
System Layout

- Information sources are defined and stored as a list of syndication feed addresses,
- which are then split into the individual feed entries.
- Each entry is then checked against a database to determine whether an incoming entry is unique.
- The content of the entry is written to disk,
- to be processed by the content categorisation component.
- User data is collected in order to guide content moderation and personalisation.
- This information can be pooled across all the users and used to discover new relevant feeds.

Planned Research

- Based upon the system presented, three strands of research are planned:
- a comparative evaluation of IR and Machine Learning categorisation techniques
 - validation of the interaction-based user modelling approach to refine topic representation
 - development of a method of automated content discovery in line with user interests

Future work will focus on exploring the interaction models that emerge when children use a topic-based filtering service and discover the similarities and differences between children and adult when they use such a system.



www.puppyir.eu

Acknowledgements: This research is funded by the European Community's Seventh Framework Programme FP7/ 2007-2013 under grant agreement no. 231507.