Ingredients for a user interface to support media studies researchers in data collection

Bron, M.; Nack, F.; de Rijke, M.; van Gorp, J.

Publication date
2012

Document Version
Final published version

Published in
EuroHCIR 2012 : European Workshop on Human-Computer Interaction and Information Retrieval

Ingredients for a User Interface to Support Media Studies Researchers in Data Collection

Marc Bron
ISLA, University of Amsterdam
m.m.bron@uva.nl

Frank Nack
ISLA, University of Amsterdam
nack@uva.nl

Maarten de Rijke
ISLA, University of Amsterdam
derijke@uva.nl

Jasmijn van Gorp
TViT, Utrecht University
j.vangorp@uu.nl

ABSTRACT

We describe our efforts to design an interface that supports media studies researchers in collecting data. Based on interviews about their search behavior we arrive at a set of search scenarios and for each we identify IR techniques that provide the required functionality. We end with a discussion about the implementation of such an interface and its re-usability across the humanities.

Categories and Subject Descriptors

H.5.2 [**User interfaces**]: User-centered design

1. INTRODUCTION

Research in the arts and humanities often follows an interpretive, associative method based on historic-cultural materials, including primary sources as well as secondary materials. Humanities researchers increasingly make use of online archives and libraries to collect and compare materials and this is changing the way they work [3]. Although the trend towards e-humanities is being addressed by computer science, current search tools remain ineffective in supporting humanities researchers in data collection [5, 17, 18].

Most support tools for e-humanities research focus either on supporting a single type of search process, are aimed at analysis and organization, or focus on a single collection of primary sources rather than the collection of secondary material across various sources and modalities. Letizia is an example of a user interface that assists a user in browsing the Web by pre-fetching related documents [10]. Flamenco is an interface that supports exploration of image collections through facets [20]. Imagesieve is an exploratory tool for museum archives based on entities [11]. See [15] for an overview of metadata enhanced interfaces for specific digital libraries. Other systems aim to support sensemaking of collected data. CombinFormation, for example, is a creativity support tool for searching, browsing, organizing, and integrating information [9]. Visualization and text analysis tools provide a wider variety of methods to organize and analyze material, e.g., MONK[1] and TaPoR.[2].

In this paper we revisit the search scenarios in which humanities researchers engage. We follow a human centered approach to derive the search scenarios that a tool for data collection in the humanities should support. We focus on a specific group of users, i.e., researchers in the field of media studies. Media Studies concerns the study of production, content and/or reception of various types of media, e.g., social media, film, and television [12]. The search for data in different modalities and across a wide variety of sources make this an interesting group for our analysis. We perform a set of interviews to analyze the information search behavior exhibited by media studies researchers during their research. Our contributions are establishing a set of search scenarios based on these interviews, identifying suitable information retrieval techniques that support these scenarios, and a discussion about the challenges in incorporating these techniques in an interface and its re-usability for other humanities disciplines.

[1] http://monkproject.org/
[2] http://portal.tapor.ca/

2. ESTABLISHING SEARCH SCENARIOS

Most research in the humanities starts out by gathering specific primary sources on a certain topic. When a selection of source materials has been made the search for additional materials starts in order to provide context for the source materials [1, 14]. In terms of information behavior a research project consists of successive information seeking processes each consisting of multiple search processes [19]. Each search process, whether for primary sources or other materials, consists of starting a search, several types of search actions, i.e., browsing, chaining, and monitoring, followed by differentiating, verifying and extracting information [7].

Each of these actions can be observed in the search processes of media studies researchers in various stages of their research cycle [4, 12]. However, the effectiveness of current search tools to support these actions depends on the goal of the search process and the organization of the material. For example, browsing is easily facilitated in an interface by providing facets over the metadata annotations of documents, but metadata is usually unavailable. We will focus on the contextualization stage of the research cycle of media studies, where the primary sources have already been collected and the search for additional material starts. In this stage multiple sources of different modality are searched for and we expect that the analysis of this process will provide search scenarios that facilitate the development of an appropriate search interface.

Interview method and analysis. We interviewed ten media studies researchers from 3 different institutes with varying levels of experience: 2 PhD students, 5 post-doctoral researchers, 1 assistant-professor, and 2 full professors. Several media are being studied: television (10), radio (2), news papers (2), and documentaries (1). The interview was conducted in a semi-structured style and consisted of three parts: (i) identification of a recent research project; (ii) open questions about search processes and research questions during the project; and (iii) an interactive part in which subjects

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	frequency
newspapers	1	4	-	2	-	1	-	1	1	3	7/10
interview	1	-	1	-	-	-	1	2	-	4	5/10
magazine/tvguide	-	1	-	1	-	-	1	2	1	-	5/10
entity homepage	2	-	1	-	1	1	-	-	-	1	5/10
forum/blog	-	-	-	1	3	1	-	-	-	1	4/10
paper archive	1	-	-	-	-	-	4	2	1	-	4/10
Wikipedia	1	-	-	1	-	1	-	-	-	1	4/10
reports	1	-	1	-	-	-	-	2	-	-	3/10
book	1	-	-	1	-	-	-	-	-	-	2/10
specific site	2	1	-	-	-	-	-	-	-	-	2/10
twitter/facebook	-	-	-	-	1	-	-	-	-	-	1/10

Table 1: Number of times a source is mentioned in an interview. Frequency is the fraction of interviews to which a code applies.

wrote down the search processes on index cards. Interviews lasted about 30 minutes, were tape-recorded and later transcribed.

In our analysis of the interviews we focus on those questions that address the process of secondary material collection. We apply open-coding [16], to identify different types of material, the type of information needs the materials satisfy, and search strategies used by media studies researchers to retrieve the material. Given this data we then identify search scenarios on a more general level. Note that when using quotes, square brackets [..] indicate modifications to the original quote to improve understanding or to protect the anonymity of the subject. For identification purposes interviewees are assigned a number, i.e, I1 to I10.

Interview results. We first consider the different types of materials that media studies researchers search for besides their primary source material, e.g., television programs. Table 1 shows the types of material mentioned during the interviews. The source that is most often used are newspapers. They provide relevant context in terms of: (i) reviews about a television program; (ii) information on events during the period in which a program was broadcasted; (iii) to reconstruct which programs were broadcasted during a time period; and (iv) whether a program itself caused some event or controversy. Interviews are used when the required information is not otherwise accessible. Interviews with directors and producers provide context in terms of productions information, e.g., why a certain format was choses for a program. Interviews with people who watched a program that is no longer available provide information on the attitude of viewers in that time. Magazines and tv-guides are interesting mainly for the reviews of programs they contain or for reconstructing a broadcasting schedule. The context of homepages depends on the entity of interest. The homepage of the production company of a program provides production information, e.g., when it was broadcasted. Alternatively, a person mentioned in a program may be of interest and his/her homepage allows the researcher to find out more about that person. Wikipedia is also a popular source for this type of information. The main use for fora and blogs is to get a sense of people's attitudes towards a certain program. Other sources used mostly to get production information are paper archives and internal reports. Books and other sites, e.g., history site, are used to provide historical context for a program. The lack of use of social media as sources is due to our sample of media researchers, who work mostly with television.

Next, we categorized the various types of information need that the materials satisfy into 3 types: (i) general background information on a topic; (ii) information on specific entities; and (iii) identifying conversational information about an event. Table 2 shows the number of times each type of information need occurred in each of the interviews. Searching for background information is mentioned by most of the interviewees. We found two general topics on which media studies researchers require background information: cultural context and media production context. Cultural context is necessary to understand the reception of a program by society: (I7) "[regarding television shows for women in the 70's] I found that as the topics discussed were more taboo, those topics were still taboo then... That the program was broadcasted at a later time slot and there the difficult topics like divorce and birth control were discussed." Another type of context, i.e., media production context, is necessary to understand why programs turned out in a certain way: (I1) "There are different ways to interpret this rebellion. As producers got more freedom in creating television programs and while the television landscape was still very much divided, more artistic programs could emerge." The goal of gathering this type of context is to learn more about the situation in which a program was broadcasted or created.

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	frequency
background	5	-	2	-	4	1	5	4	1	-	7/10
entities	2	-	3	-	1	4	-	-	-	1	5/10
conversational	1	4	1	4	-	-	-	-	-	5	5/10

Table 2: Number of times a type of information need is mentioned in an interview. Frequency is the fraction of interviews to which a code applies.

Regarding context for entities we find that biographic information about people is important: (I6) "For example, if a [person] was mentioned then I would know his name, but not his ethnic background. Part of the analysis was finding out [peoples]' ethnic background" and (I3) "not all journalists are so vain to put themselves online. Especially the ones that are not well known and then I could not find their specialism." For organizations information about the internal culture and policies is important: (I1) "Sometimes you search for policy information. How did the broadcasting company present itself, what does it mean for the broadcasting company... So what role does the program play in the perception of the broadcasting company's own history." These quotes show a particular interest in specific information about entities.

In five cases interviewees engage in a conversational search, i.e., they look for the discussion around certain events. For example three interviewees mention that they look for controversy: (I2) "I searched in newspapers for controversies, you are actually searching in other media for reflection on what happened. And then you find the title of the program. You start with there was a fight, and you need to know what it was about" and (I10) "Of course the programs that are the most controversial, those that fuel public debate and get the most media attention, are the ones that I examined closest." These quotes show how media studies researchers are interested in the reasons that cause a controversy. A similar type of information need is mentioned by two other interviewees interested in multiple views on a topic: (I3) "I was interested in the relation between political issues and whether news programs show multiple views on every topic, for example a government source and an opposition source" and (4) "I chose to organize [political figure]'s story chronologically: her rise, moments of glory, and her fall."

We find that of the information needs described above the search task of finding background information is exploratory, i.e., it is unclear what the actual goal is other than information about a general topic. Regarding the entity information need the search task is very specific, e.g., finding journalists' specialism. The search task may be repeated multiple times for different entities, but the goal remains the same. In the case of conversational search multiple types of search task are required to satisfy the information need. For example, to find multiple views on a political issue, first, an exploratory

Figure 1: Sketch of an exploratory meta search engine, numbers are used to reference components.

search task is required to find the people involved. Second, a more targeted search is necessary to find the attitude of each person towards the issue. We also note that these types of information need occur multiple times during a research project, e.g., (I3) looking for multiple views on political issues and the specialism of journalists reporting on those issues.

Finally, we consider the search strategies used to collect material. The tools used most often are web search engines, for example when searching for controversial documentaries: (I2) "I Google not for documentaries, because I do not know which are controversial. I use keywords of which I know that they are related to controversies such as: conflict." Media studies researchers are trained in searching through archives and so also use various strategies when searching for additional material: (I1) "You have to search in different ways, of course... That was the same in the archive. When you can not find anything on a shelf organized per director, then use decades as a searching criterion. So you are always trying different angles." Another example shows how chaining via web links is used to reconstruct the conversation about an entity: (I4) "right, you end up on a forum with a discussion about her biography. Where one post suggests to look at this and this. Another suggests you should read that. In this way you get a lot of pointers to links in a very organic way, and I collect it in a folder with interesting links." Even for a specific information need such as the nationality of an entity several sources are searched: (I6) "Wikipedia is also a search engine. I needed to know the ethnic background of [people]. You can do this in all sorts of ways, for example fora, but also other sites that provided information about the nationality of [people]."

These quotes show how media studies researchers use multiple strategies and cover multiple sources to get at the information they need. The most popular tool are web search engines that while specialized in navigational search are also used for exploratory and informational type searches.

3. IDENTIFYING IR TECHNIQUES

The types of information need identified in the interviews suggest that an interface that supports multiple data search and collection tasks should support the following search scenarios: (i) general search; (ii) entity information search; (iii) entity relation search; and (iv) information management.

Background search. When the goal of the search is to find general background information on a certain topic, media studies researchers engage in an exploratory search task over multiple sources, e.g., news archives, libraries, and the Web. To support this task we propose to combine features of an exploratory search engine [13] with those of a meta search engine [6], see Figure 1. A search box (1) is available for the user to type in keywords in response to which a ranked list of result snippets is displayed. To support users in finding material across sources the interface aggregates results from multiple sources. A sidebar (2) shows a list of common information sources, e.g., Wikipedia. Checkboxes are available to select or deselect one or more sources from which results for the keywords are retrieved. Facets (3) are available on the right side and support the user in filtering the result set and learning about the topics covered in the results set. A typical issue for meta search engines is the aggregation of results from different sources. We propose to leave control to the user: for each source we display one result with the option to expand (4) a source and display its results. The user is able to drag and reorder the sources in the sources list in order to select the source that will be shown at the top.

Entity information search. Another scenario is when the goal of a search task is to find more information about an entity, e.g., the ethnic background of a person, or an event, e.g., reality shows causing a controversy. To support this kind of tasks we propose an interface that combines techniques for entity resolution [8], list completion [2] and query by example [21], see Figure 2.

Figure 2: Sketch of an entity information search interface, numbers are used to reference components.

Entity resolution is necessary when only an entity's name is known, e.g., the name Michael Jordan usually refers to the basketball player, but the target could be a researcher or anyone with that name. Typing a name in the entity resolution component (1) and selecting a knowledge base against which to resolve the entity, e.g., Freebase,[3] results in a list of possible targets for the entity. These targets, e.g., a Wikipedia page or homepage, provide an identifier for the entity and context information. This is also useful in the case of events, which only in special cases have specific names, i.e., named events.

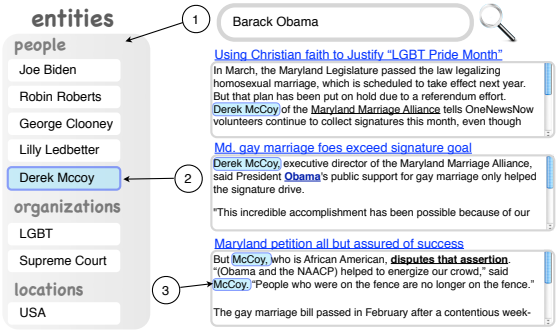
Query by example (2) supports finding information about a topic for which some information (context) is already available. Given a news article about a certain topic, find more documents that describe that same topic. Possible target resources for query by example are video databases and news archives as items from different resources describe an event in different ways.

List completion supports a scenario where a researcher wishes to find a group of entities that all have something in common, e.g., members of the same political party, but he/she has only identified some members of this group. Providing a number of examples the list completion component (3) results in a list of entities that have characteristics in common with the examples, e.g., entering Wikipedia URLs as example entities returns other entities (URLs) from Wikipedia that share characteristics with the examples.

Entity relation search. In some cases multiple types of search processes are required in order to satisfy an information need, i.e., to find multiple views on a topic exploratory and targeted searches alternate. For example, finding who are the opponents and proponents on a political issue and the reasons for their respective views. We propose an interface that facilitates viewing entities in context using dynamic snippets, see Figure 3. Whenever a search query is issued, a sidebar (1) on the left of the interface is populated with entities. To locate candidate entities online named entity recognition

[3] http://www.freebase.com/

Figure 3: Sketch of a search engine with dynamic snippets and entity highlighting, numbers are used to reference components.



is performed (NER). As NER is a costly operation this is done incrementally, i.e., first on the top 10 pages then, if the user paginates, on the next 10 pages, etc. Initially snippets in the result list contain the same text as returned from the source. The snippets, however are dynamic and when hovering over an entity (2) the snippet is updated to show a piece of text from the document in which this entity occurs, highlighting the entity (3). Jumping to an entity's position in result documents and highlighting, allows a user to inspect the context in which an entity occurs without opening each document. To account for the limited amount of space each snippet is made scrollable to enable inspection of other occurrences of the entity.

Information management. In all cases listed above the proposed techniques support finding information, however, this information needs to be stored and organized. Rather than an elaborate information organization environment common to sensemaking tools, we propose to allow the creation and assignment of labels to relevant documents. For example, in the case of the rise and fall of a political figure, documents can be organized according to the start of career, moments of glory and the eventual downfall.

4. DISCUSSION

In this paper we have established several search scenarios in which media researchers engage and recommend IR techniques that provide support in these scenarios. There are however two unresolved issues: (i) will these techniques work for each source or do they have to be tuned towards the characteristics of each data collection; and (ii) is an interface that incorporates these techniques re-usable by other humanities researchers? We believe the first point can be addressed by carefully documenting the characteristics of collections and the dependence of the retrieval performance of IR techniques on these characteristics. Whether linking a video archive with a news archive is a different task from linking a photo archive with a news archive will depend on the agreement of the characteristics of the datasets. To address the second point, we believe it is necessary to separate the functionality and the sources into modules and allow the user to compose the interface required for the search task at hand. These modules also have to be configurable, for example, the facets entities and phrases in a facets search component may be useful in some cases, while others require events and years.

Our main next step is to take these requirements and realize an interface that supports the various information search tasks of media researchers and is re-usable across the humanities.

Acknowledgements

This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.066.512, 612.-061.814, 612.061.815, 640.004.802, and partially by the Center for Creation, Content and Technology (CCCT).

References

[1] D. Altheide. *Qualitative media analysis*. Sage Publications, Inc, 1996.

[2] K. Balog, M. Bron, and M. de Rijke. Query modeling for entity search based on terms, categories, and examples. *ACM Transactions on Information Systems (TOIS)*, 29(4):22, 2011.

[3] C. Borgman. The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly*, 3(4), 2009.

[4] M. Bron, J. van Gorp, F. Nack, M. de Rijke, A. Vishneuski, and S. de Leeuw. A subjunctive exploratory search interface to support media studies researchers. In *SIGIR'12*, 2012.

[5] E. Collins and J. Michael. How do researchers in the humanities use information resources? *Liber Quarterly*, 21(2), 2012.

[6] D. Dreilinger and A. Howe. Experiences with selecting search engines using metasearch. *ACM TOIS*, 15(3):195–222, 1997.

[7] D. Ellis and M. Haugan. Modelling the information seeking patterns of engineers and research scientists in an industrial environment. *J. Doc.*, 53(4):384–403, 1997.

[8] V. Jijkoun, M. Khalid, M. Marx, and M. de Rijke. Named entity normalization in user generated content. In *Proceedings of the second workshop on Analytics for noisy unstructured text data*, pages 23–30. ACM, 2008.

[9] A. Kerne, E. Koh, S. M. Smith, A. Webb, and B. Dworaczyk. combinformation: Mixed-initiative composition of image and text surrogates promotes information discovery. *ACM Trans. Inf. Syst.*, 27(1):5:1–5:45, 2008.

[10] H. Lieberman. Interfaces that give and take advice. *In HCI in the New Millennium*, pages 475–484, 2001.

[11] Y. Lin, J. Ahn, P. Brusilovsky, D. He, and W. Real. Imagesieve: Exploratory search of museum archives with named entity-based faceted browsing. *ASIST'10*, 47(1):1–10, 2010.

[12] B. Lunn. User needs in television archive access: Acquiring knowledge necessary for system design. *JoDI*, 10(6), 2009.

[13] G. Marchionini. Exploratory search: from finding to understanding. *Commun. ACM*, 49(4):41–46, 2006.

[14] C. Palmer. Scholarly work and the shaping of digital access. *JASIST*, 56(11):1140–1153, 2005.

[15] A. Shiri. Metadata-enhanced visual interfaces to digital libraries. *JIS*, 34(6):763–775, 2008.

[16] A. Strauss and J. Corbin. Basics of qualitative research: Grounded theory procedures and techniques. *Basics of Qualitative Research Techniques and Procedures for Developing Grounded Theory*, 270, 1990.

[17] E. Toms and H. O'Brien. Understanding the information and communication technology needs of the e-humanist. *J. Doc.*, 64(1):102–130, 2008.

[18] J. Unsworth. Tool-time, or 'haven't we been here already?' ten years in humanities computing. *Transforming Disciplines: The Humanities and Computer Science*, 2003.

[19] T. Wilson. Models in information behaviour research. *J. Doc.*, 55(3):249–270, 1999.

[20] K. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *SIGCHI'03*, pages 401–408, 2003.

[21] M. Zloof. Query-by-example: A data base language. *IBM systems Journal*, 16(4):324–343, 1977.