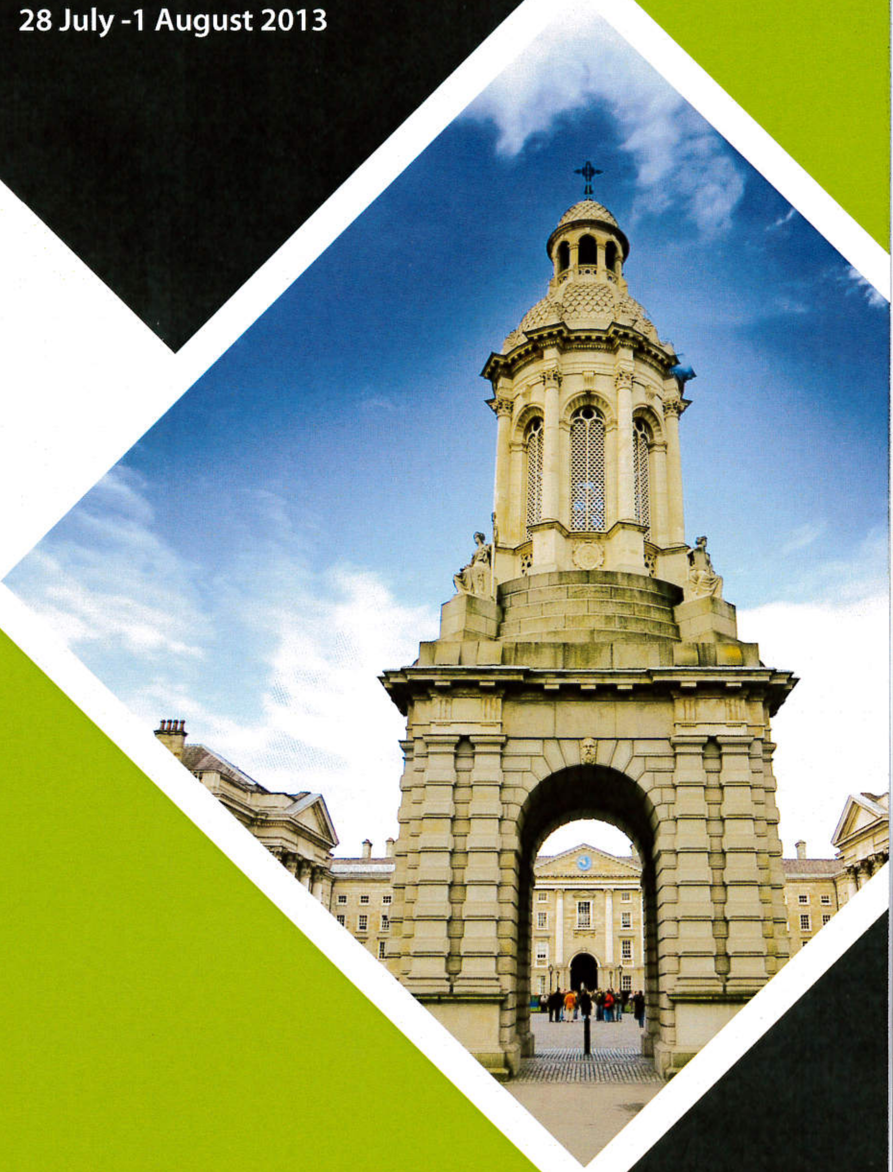
Dublin, Ireland
28 July -1 August 2013

SIGIR
2013
Dublin, Ireland
Baile Átha Cliath



Workshop Proceedings

The 36th Annual ACM SIGIR Conference

Exploration, Navigation and Retrieval of Information in Cultural Heritage (ENRICH)

Association for Computing Machinery

Advancing Computing as a Science & Profession

SIGIR
Special Interest Group
on Information Retrieval

Preface

The first workshop on Exploration, Navigation and Retrieval of Information in Cultural Heritage (ENRICH) will be held as a SIGIR 2013 Workshop on the 1st August 2013 at Trinity College Dublin[1]. Of the 16 original submissions, 3 were selected by the Program Committee to be orally presented as full papers, 3 were selected to be presented as short papers, 5 were selected to be presented as posters to be discussed during the Poster and Demo Session. During this session 2 system demos will also be presented. Thanks are due to the members of the Program Committee for their support in the paper reviewing process. Prof. Jaap Kamps of the University of Amsterdam was invited to give a keynote address entitled "When search becomes research and research becomes search."

A significant portion of the workshop programme has been dedicated to discussions regarding the issues that need to be taken into consideration when developing cultural heritage information access systems. The goal of the discussions is to determine a research roadmap for ENRICH and to outline key challenges to be addressed in the future[2]. This reflects the stated aims of ENRICH 2013, which were to:

1. Discuss the challenges and opportunities in Information Retrieval research in the area of Cultural Heritage.
2. Encourage collaboration between researchers engaged in work in this specialist area of Information Retrieval, and to foster the formation of a research community.
3. Identify a set of actions which the community should undertake to progress the research agenda.

A key challenge currently facing the curators and providers of digital cultural heritage worldwide is to instigate, increase and enhance engagement with their collections. To achieve this, a fundamental change in the way these artefacts can be discovered, explored and contributed to by users and communities is required. ENRICH 2013 is proposed as a forum to discuss how this change can be achieved. Cultural heritage artefacts, in this instance, are digital representations of primary resources; for example, manuscript collections, paintings, books, photographs, etc. Text-based artefacts are often innately "noisy", contain non-standard spelling, poor punctuation and obsolete grammar and word forms. Image-based resources often have limited associated metadata which describes the resources and their content. In addition, the information needs and tasks of cultural heritage users are often complex and diverse. This presents a specific set of challenges to traditional Information Retrieval (IR) techniques and approaches. All of these challenges will be explored as a part of the ENRICH workshop.

New forms of enhanced IR, such as those discussed and proposed in ENRICH, require rigorous evaluation and validation using appropriate metrics, contrasting digital cultural heritage collections and diverse users and communities. These evaluation challenges and opportunities will also be examined and investigated by ENRICH.

ENRICH 2013 will encourage the discussion and design of search techniques, technologies and tools that can help end-users fully exploit the wonderful Cultural Heritage material that is available across the globe.

Séamus Lawless
Maristella Agosti
Paul Clough
Owen Conlan

July 2013

[1] http://www.sigir2013.ie/workshops.html

[2] http://www.cultura-strep.eu/events/enrich-2013/

ENRICH 2013: Exploration, Navigation and Retrieval of Information in Cultural Heritage

Keynote Address

When search becomes research and research becomes search
Jaap Kamps

Paper Session 1 – Full Papers

In Search of Cinderella: A Transaction Log Analysis of Folktale Searchers
Dolf Trieschnigg, Dong Nguyen and Theo Meder

“Gute Arbeit”: Topic Exploration and Analysis Challenges for the Corpora of German Qualitative Studies
Nam Khanh Tran, Sergej Zerr, Kerstin Bischoff, Claudia Niederée and Ralf Krestel

Stereotype or Spectrum: Designing for a User Continuum
Mark Sweetnam, Micheál Ó Siochrú, Maristella Agosti, Marta Manfioletti, Nicola Orio and Chiara Ponchia

Paper Session 2 – Short Papers

Personal Name Extraction from Ancient Japanese Texts
Mamoru Yoshimura, Fuminori Kimura and Akira Maeda

Reconstruction of Apollo Mission Control Center Activity
Douglas Oard, Abhijeet Sangwan and John Hansen

Personalising the Cultural Heritage Experience with CULTURA
Gary Munnelly, Cormac Hampson, Owen Conlan, Seamus Lawless and Eoin Bailey

Poster and Demo Session

The CULTURA Evaluation Model: An Approach Responding to Evaluation Needs in an Innovative Research Environment
Christina M. Steiner, Eva-C. Hillemann, Alexander Nussbaumer, Dietrich Albert, Mark Sweetnam, Cormac Hampson and Owen Conlan

A Social Network Study of Evliyâ Çelebi’s Book of Travels-Seyahatnâme
Ceyhun Karbeyaz, Ethem F. Can, Fazli Can and Mehmet Kalpakli

Publishing Social Sciences Datasets as Linked Data: a Political Violence Case Study
Rob Brennan, Kevin Feeney and Odhran Gavin

GALATEAS D2W: A Multi-lingual Disambiguation to Wikipedia Web Service
Deirdre Lungley, Massimo Poesio, Marco Trevisan, Maha Althobaiti and Vien Nguyen

Generating Automatic Keywords for Conversational Speech ASR Transcripts
Hohyon Ryu and Matthew Lease

The CULTURA Project: Supporting Next Generation Interaction with Digital Cultural Heritage Collections (demo)
Gary Munnelly, Cormac Hampson, Séamus Lawless, Maristella Agosti and Owen Conlan

Personalized Access to Cultural Heritage Spaces (PATHS) (demo)
Paul Clough, Paula Goodale, Mark Stevenson and Mark Hall

Keynote Address

When search becomes research and research becomes search

Jaap Kamps
University of Amsterdam

We have access to unprecedented amounts of information on the Web and in now digitized collections curated by museums, archives, libraries, publishers, or other institutions. The large-scale availability of information about our past and present offers unprecedented opportunities for researchers and other users, but also presents many complexities when combining information from different collections, from different institutions, and from different traditions of documentation.

Big data collections bringing together many sources of information will be the driving force of research in the coming years.

Modern search methods are needed to exploit the availability of massive data of relevance to almost any conceivable research question, hence new search methods become new research methods and search and research evolve in parallel.

In Search of Cinderella: A Transaction Log Analysis of Folktale Searchers

Dolf Trieschnigg
University of Twente
Enschede, The Netherlands
d.trieschnigg@utwente.nl

Dong Nguyen
University of Twente
Enschede, The Netherlands
d.nguyen@utwente.nl

Theo Meder
Meertens Institute
Amsterdam, The Netherlands
theo.meder@meertens.knaw.nl

ABSTRACT

In this work we report on a transaction log analysis of the Dutch Folktale Database, an online repository of extensively annotated folktales ranging from old fairy tales to recent urban legends, written in (old) Dutch, Frisian and a variety of Dutch dialects. We observed that users have a preference for subgenres within folktales such as traditional legends and urban legends and prefer stories in standard Dutch over stories in Frisian. Searches are typically short and aim at large groups of stories (from the same subgenre or collector), or specific stories with the same main character. In contrast, search sessions are relatively long (median of around 2 minutes) and many result pages are viewed (average: 3.4 pages, median: 2 pages). Based on the observations we propose a number of improvements to the current search and browsing interface. Our findings offer insight into the search behavior of folktale searchers, but are also of interest to researchers and developers working on other e-humanities collections.

Categories and Subject Descriptors

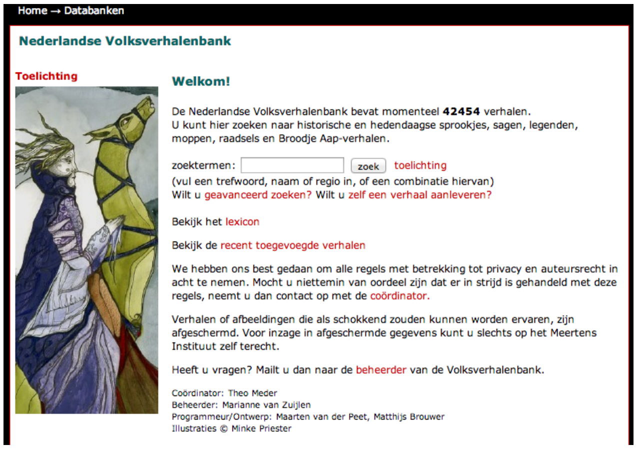
H.2.8 [**Database applications**]: Data mining; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms

Measurement, Human Factors

1. INTRODUCTION

The Dutch Folktale Database contains over 42,000 folktales from different subgenres (such as fairy tales, legends and urban legends), from different time periods (from the Middle Ages to present day), and in different languages (Frisian, Dutch and a large variety of Dutch regional dialects). It has both an archival and a research function: it preserves a part of the Dutch cultural heritage and it allows researchers to investigate the oral tradition of telling stories. The Dutch Folktale Database is maintained by the Meertens Institute in Amsterdam which studies and documents the language and culture in the Netherlands. The initial, offline, folktale database was created in 1994, and in 2004 the database became available online at `http://www.verhalenbank.nl` (currently only in Dutch). A variety of users access the online collection, ranging from folktale researchers, students writing for school projects, journalists investigating urban legends, storytellers expanding their repertoire to a general audience interested in stories related to their local region. Search is the primary mechanism to access the collection and the homepage presents a simple search box to find stories in the collection (see Figure 1).



Nederlandse Volksverhalenbank

Welkom!

De Nederlandse Volksverhalenbank bevat momenteel **42454** verhalen.
U kunt hier zoeken naar historische en hedendaagse sprookjes, sagen, legenden, moppen, raadsels en Broodje Aap-verhalen.

zoektermen: toelichting
(vul een trefwoord, naam of regio in, of een combinatie hiervan)
Wilt u geavanceerd zoeken? Wilt u zelf een verhaal aanleveren?

Bekijk het lexicon

Bekijk de recent toegevoegde verhalen

We hebben ons best gedaan om alle regels met betrekking tot privacy en auteursrecht in acht te nemen. Mocht u niettemin van oordeel zijn dat er in strijd is gehandeld met deze regels, neemt u dan contact op met de coördinator.

Verhalen of afbeeldingen die als schokkend zouden kunnen worden ervaren, zijn afgeschermd. Voor inzage in afgeschermde gegevens kunt u slechts op het Meertens Instituut zelf terecht.

Heeft u vragen? Mailt u dan naar de beheerder van de Volksverhalenbank.

Coördinator: Theo Meder
Beheerder: Marianne van Zuijlen
Programmeur/Ontwerp: Maarten van der Peet, Matthijs Brouwer
Illustraties © Minke Priester

Figure 1: Homepage of the Dutch Folktale Database, including a simple search form.

It is unclear how folktale searchers actually use this search functionality. Given the broad range of users, is the search behaviour comparable to web searchers? Or do they have their own typical search behavior?

In this paper we investigate how users of the Dutch Folktale Database use its search functionality. By means of a transaction log analysis we would like to get more insight on how they search and what they search for. Based on the study we give a profile of folktale searchers and suggest improvements for the current database.

The overview of this paper is as follows. In section 2 we briefly review related work in the area of transaction log analysis. In section 3 we describe the Dutch Folktale Database and its users in more detail based on an online user questionnaire. In section 4 we describe the collection and preprocessing of the transaction log. In section 5 we in-

SIGIR'13, ENRICH Workshop August 1, 2013, Dublin, Ireland.
ACM .

vestigate frequently accessed content of the collection and in section 6 we analyze the users' search behavior. In section 7 we summarize our observations, give suggestions to improve the current database, and present lessons learned for similar projects.

2. RELATED WORK

In this section we briefly review related work in the area of transaction log analysis. For a more comprehensive review see Jansen [4].

Transaction log analysis is an inexpensive way to unobtrusively collect information about a large number of users about their interaction with a web (search) system [4]. Early transaction log analysis primarily focused on the logs of general web search [6, 12]. Jansen et al. [6] analyzed over 50,000 queries on the Excite web search engine. They report that web searchers are uncomfortable using Boolean search operators and other advanced search options. Typically only the first result page is viewed. Silverstein et al. [12] carried out a large log analysis of a web search engine. They report an average query length of 2.3 terms and short search sessions with few queries and result page views.

With the rise in popularity of vertical search, research has focused on analyzing the logs in these specific domains. Mishne and de Rijke [10] analyzed the query log of a large blog search engine. They show that blog searchers primarily search for names and blog themes. Blog search sessions are typically short and only few search results are viewed. Jones et al. [7, 8] investigated a transaction log analysis of a digital library containing technical reports in the area of computer science. Short queries were observed here as well (2.43 terms on average), and most queries (two out of three) did not contain boolean operators. Again, results sets are reported not to be thoroughly inspected by users. Ke et al. [9] investigated the search behavior of Taiwanese users of the ScienceDirect portal which gives access to scientific and technical papers. They report an average query length of 2.3 terms, but do not report on the number of queries during a search session. Huurnink et al. [2] analyzed the transaction log of a audiovisual archive from which material can be ordered. Half of the recorded sessions are shorter than a minute; search sessions resulting in an order are considerably longer (7 minutes). Queries consist mostly of free text keyword search, but also date filters are used. In 9% of the queries the advanced search function is used. Weerkamp et al. [15] report on the analysis of a people search engine log. They propose a classification scheme in this domain at the query, session and user level, for instance by distinguishing queries for more or less popular persons. An average session length of 1.6 queries was observed. Islamaj Dogan et al. [3] investigated one month of query data from PubMed, which provides access to a large repository of biomedical citations. They conclude that PubMed users primarily search for authors, genes/proteins, and diseases and they frequently reformulate their queries. Search sessions consist of 4 queries on average and an average query consists of 3.5 terms. Park and Lee [11] analyzed the transaction log of a web-based IR system in science and technology. They report very short queries (1.4 terms on average) and relatively long users sessions in comparison to web searchers.

To the best of our knowledge no transaction log analyses have been carried out for folktale searchers.

3. THE DUTCH FOLKTALE DATABASE

The Dutch Folktale Database currently[1] contains 42,454 *folktales*. Folktales circulate among people in oral tradition and are part of our folkore and cultural identity. By definition folktales cover a broad variety of subgenres. The most important subgenres in the Dutch Folktale Database are traditional legends (stories with a known place and time and often containing supernatural elements such as witches or ghosts), saint's legends (religious tales about saints, sacred objects and miracles), jokes (short stories for laughter), urban legends (gloomy contemporary stories claimed by the narrator to have actually happened), riddles (question-answer stories) and fairy tales (adventurous stories, playing in an unspecified time and place, often containing magical items). Table 3(a) lists the distribution of the collection over the subgenres. Most of the folktale material has been collected in the nineteenth to twenty-first centuries, but stories from the Middle Ages and the Renaissance are present as well. The stories have been written down in a large number of languages including Frisian, Standard Dutch, 17th century Dutch, Middle Dutch, regional dialects and combinations of languages (also see Trieschnigg et al. [13]). A total of 196 unique language combinations is present in the metadata (based on 92 unique language names). Table 3(b) lists the distribution of the collection over the languages.

Another important metadata field is the type of the folktale. This field refers to international catalog numbers used for indexing folktales such as the Aarne Thompson Uther index [14] and the Brunvand classification of urban legends [1]. Using this field all variations of Cinderella, e.g. told in different languages or in different times, are conveniently grouped.

Other metadata fields include: a summary and keywords in standard Dutch; the geographic region in which the story was told; the name of the storyteller; proper names present in the story; the source where the story came from, for instance a book or received by e-mail; a title of the story (which is frequently not part of an orally transmitted folktale); and the corpus this story is part of. Table 8(a) lists all the metadata fields (as present in the advanced search function).

3.1 User Survey

To get an idea of the user demographics we posted a brief opt-in questionnaire on the homepage of the Dutch Folktale Database, to which 88 people responded between June 2012 and June 2013. A summary of the answers is listed in Table 1 (note that some questions were not answered by all respondents). A majority of the respondents indicated to be male (57%). Table 1(b) lists the indicated age ranges of the users. All age categories are present, but more than a quarter (26%) of the users is between 55 and 64 years old. More than half (57%) of the users is above 45 years old. Table 1(c) lists the highest education the users have received[2]. The users are highly educated: 56% of the respondents indicated to have a university diploma. To the multiple response question what they intend to do with the found information, 64% of the respondents indicated 'personal use' (see Table 1(d)). Another large group of respondents (30%) indicated to use the found information for storytelling. 17% and 14% of the

[1] June 2013

[2] For explanation of the Dutch education names see `http://en.wikipedia.org/wiki/Education_in_the_Netherlands`

respondents indicated scholarly and educative use, respectively. Only 3 respondents indicated to use the information for journalism. In the 'other' category respondents noted inspiration for making art and for developing a guided tour.

We can conclude that the average user of the Folktale Database is highly educated, between 55 and 64 years old, and interested in folktales for personal use or for storytelling (or both).

3.2 The Search and Browsing Interface

The Dutch Folktale Database provides a simple and an advanced search interface. The simple search interface, shown in Figure 1, provides a single search box which searches the keywords, proper names and region metadata fields. The results are shown in a list and are ordered by their id number (e.g. 'ABIJMA22', which consists of a string prefix indicating its collection and a number).

The advanced search interface, shown in Figure 2(a), allows the user to enter separate query terms for each of the metadata fields. The entered values are combined with a Boolean AND. Using check-boxes next to the input field the user can indicate which fields should be shown in the result view. Also the advanced search results are ordered by id number.

When clicking a search result, the user is directed to an overview page of the folktale (shown in Figure 2(b)). This page lists all the metadata fields. The full-text of the story is available through a separate link. The overview page also links to a description of the story type (a catalog page), lists of stories by the same storyteller, lists of stories of the same type, and a map of stories of this type.

4. COLLECTION AND PREPARATION

We carried out a transaction log analysis of the Dutch Folktale Database. We followed the methodology described by Jansen [5] who distinguishes between collection, preparation and analysis of the transaction log. We extracted the transactions from an Apache web server log recorded between April 2010 and Jan 2012, a period of 21 months. From each line we used the following information: the date and time of the request, the IP address or hostname from which the request was issued, the requested page (URL), the user agent (typically the name of the browser) and the response code from the server.

A total of 3,870,947 requests was logged in this period. After removing requests from Web crawlers based on user agent (amounting to 70% of the logged traffic), and removing requests for style sheets and images which are part of the website template, a log of 502,893 requests remained.

We used a simple method to identify sessions in the transaction log: requests from the same IP address or hostname with no longer than one hour between two consecutive requests are grouped into a session. A similar method was used by Weerkamp et al. [15] in the context of people search. We realize that in case of a shared or public computer this can result in erroneously grouped requests. However, given the modest number of users, we expect few errors to occur.

In the next two sections, we first analyze frequently accessed content and then we look into sessions and simple and advanced queries.

Table 2: Request types

Request type	Percentage
Story overview	24%
Story text	17%
Homepage	13%
Simple search result	13%
Lexicon entry	11%
List stories of type	5%
Advanced search result	4%
List stories from storyteller	2%
Recent additions	2%
List lexicon entries	2%
Advanced search page	2%
Catalog page	2%
Map	1%
About page	1%
Multimedia	1%
Bibliography	< 1%
Insert story	< 1%

5. POPULAR CONTENT

In this section we first analyze the type of requests made to the website, then we describe which type of folktales was frequently accessed.

5.1 Request Types

We categorized page requests according to request types. In most cases this came down to classifying the URLs after removing all parameters. If, for instance, a user visits the entry page of the website, this is a request of type 'Homepage', which offers the possibility to search. Entering a (simple) search and pressing enter, results in a 'Simple search result' request, clicking the link to the advanced search form in a 'Advanced search page' request. Requests for (the first or later) search result page are labeled 'Advanced search result'. A complete list of the request types is listed in Table 2. 24% of the requests are for story overview pages, these show the metadata fields in a single view. The second most popular request is for the full text of a story, which is accessible from the overview page. Note that these requests include users who use a general web search engine and directly access a page. The simple search interface is accessible only from the homepage and a search result page is equally often requested as the homepage itself. Of the requests for simple search result pages, 73% is a request for the first page, whereas 27% is a request for the second or later page. The requests for advanced search result pages shows a higher percentage of requests for later pages (44%), indicating that during advanced searches more search results are viewed.

We conclude that basic content pages are most frequently requested, including story summaries, story full-text and lexicon entries. More advanced content pages, such as multimedia items, catalog pages and maps are less popular. The most popular way to find a story (on the site) is through a simple search, but also browsing via stories of the same type and stories from the same author is popular.

Table 1: Summary of responses to the user questionnaire (n=88)

(a) Gender

Gender	Frequency	
Male	49	57%
Female	37	43%

(b) Age

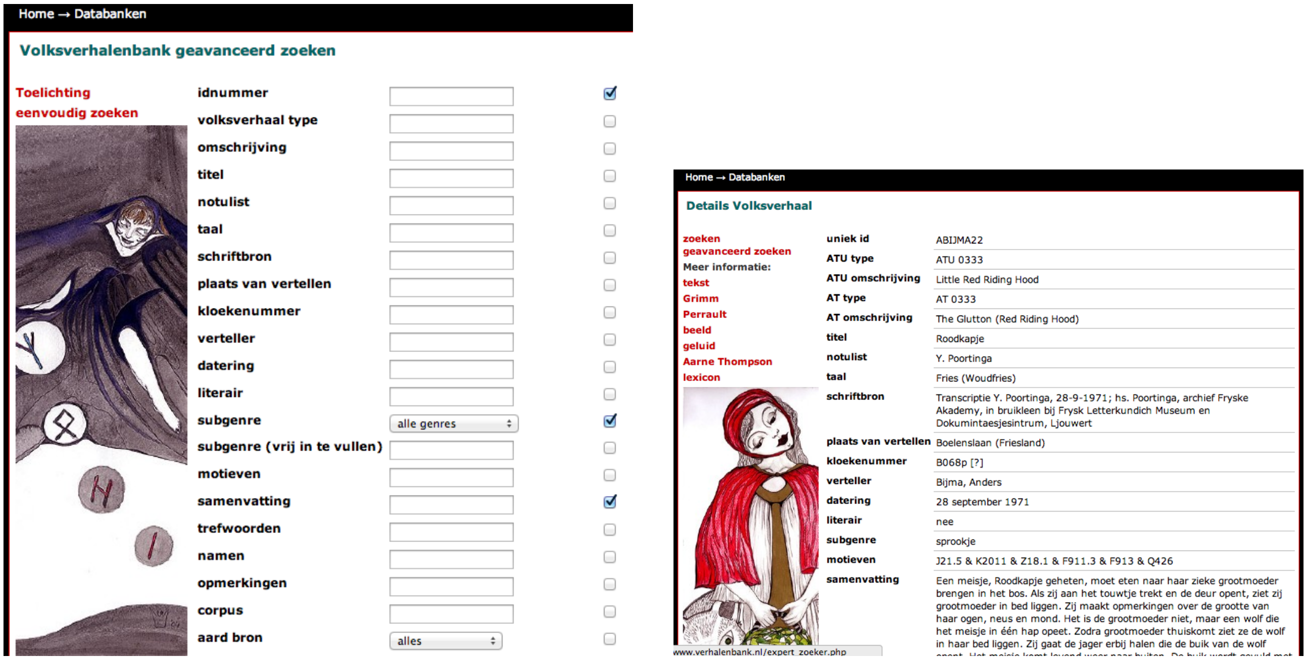
Age range	Frequency	
<15	9	10%
15–24	11	13%
25–34	8	9%
35–44	10	11%
45–54	8	9%
55–64	23	26%
65+	19	22%

(c) Education

Education	Frequency	
No education, elementary school	6	7%
LBO, VBO, VMBO (secondary school)	4	5%
MAVO, first 3 years HAVO/VWO (secondary school)	3	4%
MBO (vocational learning)	7	8%
HAVO/VWO last 2 years (higher secondary school)	17	20%
HBO/WO-bachelor (bachelor)	29	34%
WO-doctoraal, master (master)	19	22%

(d) Usage

Usage	Frequency	
Personal use	56	63.6%
Scholarly use	15	17.0%
Education	12	13.6%
Journalism	3	3.4%
Storytelling	26	29.5%
Other	11	12.5%



Home → Databanken

Volksverhalenbank geavanceerd zoeken

Toelichting
eenvoudig zoeken

idnummer		☑
volksverhaal type		☐
omschrijving		☐
titel		☐
notulist		☐
taal		☐
schriftbron		☐
plaats van vertellen		☐
kloekenummer		☐
verteller		☐
datering		☐
literair		☐
subgenre	alle genres	☑
subgenre (vrij in te vullen)		☐
motieven		☐
samenvatting		☑
trefwoorden		☐
namen		☐
opmerkingen		☐
corpus		☐
aard bron	alles	☐

(a) Advanced search page

Home → Databanken

Details Volksverhaal

zoeken
geavanceerd zoeken
Meer informatie:
tekst
Grimm
Perrault
beeld
geluid
Aarne Thompson
lexicon

uniek id	ABIJMA22
ATU type	ATU 0333
ATU omschrijving	Little Red Riding Hood
AT type	AT 0333
AT omschrijving	The Glutton (Red Riding Hood)
titel	Roodkapje
notulist	Y. Poortinga
taal	Fries (Woudfries)
schriftbron	Transcriptie Y. Poortinga, 28-9-1971; hs. Poortinga, archief Fryske Akademy, in bruikleen bij Frysk Letterkundich Museum en Dokumintaesjesintrum, Ljouwert
plaats van vertellen	Boelenslaan (Friesland)
kloekenummer	B068p [?]
verteller	Bijma, Anders
datering	28 september 1971
literair	nee
subgenre	sprookje
motieven	J21.5 & K2011 & Z18.1 & F911.3 & F913 & Q426
samenvatting	Een meisje, Roodkapje geheten, moet eten naar haar zieke grootmoeder brengen in het bos. Als zij aan het touwtje trekt en de deur opent, ziet zij grootmoeder in bed liggen. Zij maakt opmerkingen over de grootte van haar ogen, neus en mond. Het is de grootmoeder niet, maar een wolf die het meisje in één hap opeet. Zodra grootmoeder thuiskomt ziet ze de wolf in haar bed liggen. Zij gaat de jager erbij halen die de buik van de wolf opent. Het meisje komt levend weer naar buiten. De buik wordt gevuld met

www.verhalenbank.nl/expert_zoeker.php

(b) Story overview page

Figure 2: Screenshots of the search interface

5.2 Accessed Content

A total of 201,129 story requests were in the log. The most frequently requested story amounts to 1.1% of these requests, and is an urban legend about the vanishing hitchhiker. The second and third most frequently requested stories are a traditional legend about being "born with the helmet" and an urban legend about tampered food. Most of the requests for these items originate from popular searches on web search engines. For instance, the urban legend about the vanishing hitchhiker is the first hit when searching the database for a Brunvand catalog number, which is linked from several Google search results.

Tables 3(a) and 3(b) show that the story requests are not evenly distributed over the languages and subgenres present in the database. Traditional legends are frequently requested (45%) and also form the largest subgenre in the collection (55%). However, stories with urban legends and fairy tales, which amount to only a small fraction of the collection, receive many more requests. Most of the stories in the database are in Frisian (42%), but Standard Dutch is the language of most of the requested stories (64%).

In a redesign of the website these statistics can be taken into account to provide easy access to frequently accessed subgenres and languages, or to highlight stories which are currently not frequently requested.

6. SEARCH ANALYSIS

We first look into the characteristics of search sessions. Then we analyze simple and advanced searches in more detail.

A total of 140,136 of sessions is identified, of which 20,162 contain one or more simple or advanced searches. Table 4 lists the session statistics. A session consists of 3.6 requests and takes 3.5 minutes on average; however, half of the sessions consists of only a single request (sessions with only a single request were counted as a duration of 0 seconds). 20,162 sessions (14%) contains at least a single search.

Search sessions have an average duration of more than 10 minutes, but given the large standard deviation this value is deceptive: half of the search sessions is shorter than 2 minutes. Search sessions are quite long: on average 13 requests are made (with a median of 6). Search sessions involving an advanced search (median of 7 minutes) take more than three times longer than simple search sessions.

The sessions are quite long in comparison to search in an audiovisual collection. Huurnink et al. [2] reported a median of half a minute, where folktale searchers use 2 minutes. Also the percentage of advanced searches is higher in this collection (13% vs. 8%). The duration of an advanced search session is comparable to audiovisual searches leading to an order (both 7 minutes), which also could be considered an advanced search. A possible explanation of the differences could be the average age of the users of the folktale database: first-time visitors might need more time to get used to the search system. Users who are more familiar with the website and who use the advanced search function might search faster and more similar to professional audiovisual searchers.

Table 5 lists the querying and viewing statistics of simple and advanced search sessions. On average, 2.4 queries are issued during a simple search session. This is slightly more than the reported number of queries for people search [15] but fewer than an average PubMed search [3]. Twice as many queries are issued during an advanced search session. The number of viewed result pages is surprisingly high: 3.4 result pages are viewed on average (median 2.0). This means that on average 34 search result snippets are viewed. In contrast, the number of viewed story overview pages is relatively low: on average 3.1 (mean 1.0) overview pages are viewed. The full-text of the documents is visited for 71% of the viewed summaries. During advanced search sessions more unique queries are issued, and more result pages, overview pages and full-texts are viewed.

Table 5: Per session statistics

	Simple			Advanced		
Number of	Avg.	Std.	Med.	Avg.	Std.	Med.
Unique queries	2.4	4.4	1.0	4.9	9.8	2.0
Result pages viewed	3.4	7.3	2.0	7.7	15.0	3.0
Summaries viewed	3.1	10.6	1.0	7.9	19.0	2.0
Full-texts viewed	2.2	6.7	0.0	5.2	13.8	1.0

One possible explanation for the large number of viewed result pages might be the fact that the results are not ranked by relevance, but by id number. This might give incentive to look further for relevant documents.

6.1 Simple Queries

Table 6 shows the most frequent simple queries and a description of what it searches for. It is apparent that the most frequent queries are expected to give wrong results or no results at all. Searches for particular collections (e.g. 'cornelius bakker' and its abbreviation 'cb'), subgenre (e.g. 'sage', 'sagen', 'legende', 'stadssage', 'broodje aap') are not supported by the simple search interface, which only searches in the keywords, proper names and region metadata fields. Another frequently issued query is the empty query (yielding no results). This could indicate that visitors do not have a specific information need, but simply want to browse the collection.

Table 7 shows the number of terms per query. On average, a single query consists of 1.4 terms (standard deviation 0.9), but most of the queries (75%) consist of only a single term. In comparison to web queries (Silverstein et al. [12] reports an average of 2.35 terms per query), this is rather short. This might be explained by the fact that the collection is relatively small (around 42,000 documents); also short queries result in a manageable number of results. The same average number of query terms is reported by Park and Lee [11] for searching scientific and technical information.

6.2 Advanced Queries

Advanced queries can be submitted from the advanced search page (see Figure 2(a)), and allows to search in specific fields.

On average, the advanced queries (1.95 terms per query) are slightly longer than simple queries. Table 8(a) lists the use of fields. Advanced searches involving a (part of a) story id are most popular; these are typical known item searches, or searches for known collections[3]. Similar to the most frequent simple searches, searches for particular subgenres and

[3] The story id has a collection-specific prefix

Table 3: Story access per language/subgenre in comparison to occurrence in the database.

(a) Subgenre

Subgenre	Accessed		Database
Traditional legend	91,311	45.4%	54.5%
Urban legend	39,037	19.4%	7.2%
Fairy tale	34,687	17.2%	3.7%
Joke	20,086	10.0%	24.5%
Personal narrative	6,008	3.0%	2.5%
Saint's legend	3,704	1.8%	1.0%
Not assigned	2,370	1.2%	0.6%
Riddle	1,830	0.9%	5.0%
Animal tale	533	0.3%	0.1%
Song	496	0.2%	0.1%
Other	1,067	0.5%	0.8%

(b) Language

Language	Accessed		Database
Standard Dutch	128,759	64.0%	36.8%
Frisian	34,111	17.0%	42.0%
Standard Dutch mixed	12,886	6.4%	3.9%
17th century Dutch	5,058	2.5%	5.7%
Gendts	2,942	1.5%	0.3%
Noord-Brabants	2,874	1.4%	1.7%
Gronings	2,176	1.1%	2.1%
Middle Dutch	2,022	1.0%	1.6%
Flemish	1,577	0.8%	2.4%
Waterlands	1,031	0.5%	0.4%
Other	7,693	3.8%	3.2%

Table 4: Frequency and duration statistics of different types of sessions

		Duration (s)			Number of requests		
	Freq.	Avg.	Std.	Median	Avg.	Std.	Median
All sessions	140,136	213.9	1792.9	0.0	3.6	17.3	1.0
Search sessions	20,162	622.0	1801.3	120.0	13.1	31.3	6.0
Sessions with a simple search	18,923	608.1	1801.4	118.0	12.9	31.0	6.0
Sessions with an advanced search	2,586	1431.0	3104.0	428.5	32.6	61.6	12.0

Table 6: Most frequent simple queries

Query (translation)	Count	Description
cb	1,308	Collection
roodkapje (red riding hood)	816	Main character
sage (traditional legend)	516	Subgenre
sagen (traditional legends)	401	Subgenre
empty	396	
cornelius bakker	331	Collection
legende (saint's legend)	328	Subgenre
bokkerijders (buckriders)	284	Main character
witte wieven (white women)	224	Main character
stadssage (urban legend)	222	Subgenre
broodje aap (urban legend)	221	Subgenre
limburg	216	Region
sprookjes (fairy tales)	198	Subgenre
legenden (saint's legends)	193	Subgenre
moppen (jokes)	173	Subgenre
project (project)	154	
sprookje (fairy tale)	143	Subgenre
sinterklaas	141	Main character
kerst (christmas)	135	Theme
heks (witch)	126	Character type
Other	39,149	
Total	45,675	

Table 7: Number of terms per query

Query length	Frequency	
1 term	34,144	74.8%
2 terms	7,934	17.4%
3 terms	2,014	4.4%
4 terms	646	1.4%
>4 terms	937	2.1%

keywords are popular. The summary field is only infrequently used for searching. Adding this field to the database is quite expensive: an archivist has to read the full-text and write a summary by hand, which is a time-consuming process. It is not used for matching stories, but it is used for displaying search results. Another notable difference in frequency is between region and Kloeke number. The region is a free-text field indicating the place where the story was told, the Kloeke number is a unambiguous identifier indicating a region on the map in the Netherlands and Flanders. Despite the fact that such an unambiguous identifier is present, users prefer to search using a free-text description of the location. A plausible explanation is that searchers do not understand how to use the Kloeke numbers for searching.

Table 8(b) lists which fields are combined in multi-field searches. Most popular are combinations involving a subgenre (e.g. fairy tale, joke or traditional legend). Surprisingly many searches combine a story id with a subgenre, which might seem strange: stories with a similar story id (i.e. from the same collection) can be in multiple subgenres, making an additional filter on subgenre useful. Other useful combinations are a subgenre with one or more keywords,

Table 8: Use of fields in advanced searches

(a) Fields used in advanced searches

Field	Frequency		Examples
story id	4,564	30.0 %	cb, cd, esopet
subgenre	2,885	19.0 %	traditional legend (34%), saint's legend (13%), fairy tale (11%), urban legend (9%)
keywords	1,767	11.6 %	heks (witch), haas (hare), spook (ghost)
type	853	5.6 %	sage, legende, sprookje, at, brun
region	761	5.0 %	Amsterdam, Friesland, Groningen
storyteller	611	4.0 %	cornelius bakker, cb, km
names	575	3.8 %	bartje poep, herk ooievaar, Rotterdam
source type	533	3.5 %	oral (74%), book (8%), article (4%)
description	421	2.8 %	duivel (devil), heks (witch), boom (tree)
collector	403	2.7 %	Koman, Jaarsma, Sophie van Setten
language	291	1.9 %	Dutch (47%), Frisian (14%)
title	290	1.9 %	Roodkapje (Red riding hood), Assepoester (Cinderella)
summary	268	1.8 %	heks (witch), poema (puma)
date	213	1.4 %	1911, 2001, voor 1900 (before 1900)
corpus	162	1.1 %	Overbeke, USA, Jaarsma
source	151	1.0 %	Algemeen Dagblad, Overbeke
motifs	140	0.9 %	k083p, vrouw (woman)
remarks	128	0.8 %	beeld (image/statue), Roosendaal
Kloeke number	104	0.7 %	k150p, k083p, g196p
literary	86	0.6 %	ja (yes), grimm, nee (no)
Total	15,206	100.0 %	

(b) Fields used in multi-field searches

Fields	Frequency	
story id, subgenre	654	21.6 %
keywords, subgenre	333	11.0 %
source type, subgenre	309	10.2 %
region, subgenre	99	3.3 %
storyteller, subgenre	70	2.3 %
keywords, region	68	2.2 %
all fields	60	2.0 %
keywords, names	55	1.8 %
keywords, source type, subgenre	55	1.8 %
subgenre, summary	51	1.7 %
subgenre, type	47	1.5 %
keywords, story id	46	1.5 %
description, subgenre	42	1.4 %
corpus, keywords	41	1.4 %
keywords, language	39	1.3 %
story id, storyteller	34	1.1 %
collector, keywords	34	1.1 %
names, subgenre	33	1.1 %
collector, region	31	1.0 %
subgenre, title	30	1.0 %
Other	903	29.8 %
Total	3,034	100.0 %

and a subgenre with the source type (e.g. book, oral, e-mail). The examples show that searchers have difficulties to understand the advanced search interface. Commonly used queries in the 'type' field are in fact subgenres, which will result in no results.

7. CONCLUSION

In this paper we carried out a transaction log analysis of users of the Dutch Folktale Database. Additionally, we reported on the results of a short survey to determine basic demographics of its users.

Summary of Results

The survey indicated that most folktale searchers are 45 years or older, are highly educated and use the found information for personal use or storytelling.

Basic content items such as story overview pages and story full-text are most frequently accessed. More advanced pages such as maps and catalog pages receive few requests. The most frequently accessed content does not correspond to the distribution of subgenres and languages in the collection. Urban legends and fairy tales are relatively popular under searchers, whereas these subcollections are relatively small. Stories in standard Dutch are more popular than stories in Frisian, despite the large number of Frisian stories in the collection.

Users of the Dutch Folktale Database clearly have their own search behavior. Simple search queries are short (1.4 terms on average), but search sessions are relatively long and many search result pages are viewed. The click-through ratio from result page to story overview page is low, but when a story overview page is viewed frequently also its full-text is requested. Most simple searches aim to find subgenres and collections, but also main characters are frequently searched for.

Advanced searches lead to even longer search sessions with more viewed result pages and story pages. Advanced searches focus on particular stories or collections, subgenres and story types. Multi-field searches typically include a subgenre.

Based on the analyzed searches we hypothesize about two types of users. The first is someone who is familiar with the database and poses advanced queries on the collection to find a known story based on its id, collection or story type. The second is a user who is new to the collection and wishes to explore it. These users are characterized by simple queries on a particular subgenre which results in long result lists. These users are also characterized by empty queries, with the intent to retrieve the complete collection, but which currently returns an empty list.

Recommendations for the Dutch Folktale Database

Based on the results we propose the following improvements for the current Folktale Database.

It turns out that the most popular searches in the simple search interface are semantically incorrect: searching for a subgenre in the simple search interface does not search the subgenre metadata field. Also some of the advanced searches are unnecessarily complex. The search interfaces should more clearly communicate its functionality. Since searching for subgenre is so popular, this should be part of the simple search interface.

Given the broad (subgenre and collection) queries, there is a need for a browsing mechanism of the collection. This browsing mechanism should at least include subgenres, types and languages. The browsing mechanism could also be used to promote less frequently requested parts of the collection, such as stories in dialects, or riddles and songs.

The free-text region metadata field is more frequently used for searching than the unambiguous Kloeke number, but it can be expected that searching the free-text region field gives undesirable results because of mismatches (e.g. a user searches for the name of the region rather than the village it was annotated with). A more advanced geographic search system which for instance allows querying on a map would make the geographic labeling of more use.

Lessons for Cultural Heritage

If we view our results in the context of cultural heritage in general, we can summarize the following lessons.

The most general advice we can give is to know your user and his needs. Obviously, the search interface should be geared towards these users and frequent search patterns. In the case of the Folktale Database users frequently search for certain subgenres and types which should be clearly accommodated. Search sessions are relatively long: users are willing to put considerable effort in their searches to achieve their goal. The search interface should accommodate these long search sessions, for instance by providing means to reformulate queries, view related stories and visualize search results using different perspectives. Another concrete suggestion would be the use of a basket in which relevant documents can be stored, but which also keeps track of seen stories. Additional tooling could be used to analyze the basket of relevant stories.

But perhaps just as important as the user and his needs is the goal of the access system itself. The goal of the Dutch Folktale Database is to allow the general public to access and get acquainted with a part of the Dutch Folklore. We observed that large subcollections of the Folktale Database are not frequently accessed. We think that this is because the user doesn't know about the existence of the material. In addition, we observed a group of users who does not have a clear information need but wants to explore the collection. Therefore an important lesson is that the access system can be used to put information of interest "on display", analogue to a museum which varies exhibitions to attract different target audiences. For the Folktale Database exhibitions can be stories about a particular region, in a particular dialect or about a specific theme or main character. Using such exhibitions, the system is not only a searchable archive but also a virtual museum which can be browsed and explored.

8. ACKNOWLEDGEMENTS

This research was supported by the Folktales as Classifiable Texts (FACT) project, part of the CATCH program funded by the Netherlands Organization for Scientific Research (NWO).

References

[1] J. H. Brunvand. A type index of urban legends. In *Encyclopedia of Urban Legends.*, pages 741–765. 2012.

[2] B. Huurnink, L. Hollink, W. van den Heuvel, and M. de Rijke. Search behavior of media professionals at an audiovisual archive: A transaction log analysis. *Journal of the American Society for Information Science and Technology*, 61(6):1180–1197, 2010.

[3] R. Islamaj Dogan, G. C. Murray, A. Neveol, and Z. Lu. Understanding PubMed(R) user search behavior through log analysis. *Database*, 2009, Nov. 2009.

[4] B. J. Jansen. Search log analysis: What it is, what's been done, how to do it. *Library and Information Science Research*, 28(3):407–432, 2006.

[5] B. J. Jansen. The methodology of search log analysis. In B. J. Jansen, A. Spink, and I. Taksa, editors, *Handbook of Research on Web Log Analysis*, pages 100–123. IGI Global, Hershey, 2009.

[6] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management*, 36(2):207–227, Mar. 2000.

[7] S. Jones, S. J. Cunningham, and R. McNab. Usage analysis of a digital library. In *Proceedings of the third ACM conference on Digital libraries*, pages 293–294, New York, NY, USA, 1998. ACM Press.

[8] S. Jones, S. J. Cunningham, R. McNab, and S. Boddie. A transaction log analysis of a digital library. *International Journal on Digital Libraries*, 3 (2):152–169, 2000.

[9] H. Ke, R. Kwakkelaar, Y. Tai, and L. Chen. Exploring behavior of E-journal users in science and technology: Transaction log analysis of Elsevier's ScienceDirect OnSite in Taiwan. *Library and Information Science Research*, 24(3):265–291, Jan. 2002.

[10] G. Mishne and M. de Rijke. A study of blog search. In *ECIR 2006*, pages 289–301, Berlin, Heidelberg, 2006. Springer-Verlag.

[11] M. Park and T. Lee. Understanding science and technology information users through transaction log analysis. *Library Hi Tech*, 31(1):123–140, 2013.

[12] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *ACM SIGIR Forum*, 33(1):6–12, Sept. 1999.

[13] D. Trieschnigg, D. Hiemstra, M. Theune, F. de Jong, and T. Meder. An Exploration of Language Identification Techniques for the Dutch Folktale Database. In *Adaptation of Language Resources and Tools for Processing Cultural Heritage workshop (LREC 2012)*, Istanbul, Turkey, May 2012.

[14] H. J. Uther. *The Types of International Folktales: A Classification and Bibliography Based on the System of Antti Aarne and Stith Thompson*, volume 1-3. Suomalainen Tiedeakatemia, Helsinki.

[15] W. Weerkamp, R. Berendsen, B. Kovachev, E. Meij, K. Balog, and M. de Rijke. People searching for people. In *SIGIR 2011*, pages 45–54. ACM Press, 2011.

“Gute Arbeit”: Topic Exploration and Analysis Challenges for Corpora of German Qualitative Studies

Nam Khanh Tran*, Sergej Zerr*, Kerstin Bischoff*, Claudia Niederée*, Ralf Krestel**
*L3S Research Center, Hannover, Germany
{ntran,zerr,bischoff,niederee}@L3S.de
**Bren School of Information and Computer Sciences, University of California, Irvine
krestel@uci.edu

ABSTRACT

Given their long-standing research traditions, a tremendous body of data has been collected in the social sciences by observing or interviewing people regarding their behavior, attitudes, beliefs, etc. The Sociological Research Institute (SOFI) in Göttingen (Germany) carried out a number of studies observing working situation in German automobile and shipyard industry after the rapid economic growth in post-World War II Germany - the so-called German “economic miracle”. Qualitative data in form of worker interviews was collected during the period of over the last 40 years, starting from early 60’s (i.e Volkswagen and German dockyard studies) and findings of these studies made a significant impact on the working situation in German industry. Intelligent access to this heritage of qualitative data would turn such data collection into a valuable source for a secondary research, e.g., for longitudinal (meta)analysis or historical investigations. By using modern information technologies the project “Gute Arbeit” aims at providing intelligent access to qualitative social science data on the subject of “good work”. Topic modeling has gained a lot of popularity as a means for identifying and describing the topical structure of textual documents and whole corpora. However, when applied to the corpora directly, topic modelling leads to poor quality topic models due to the limited number of sociological surveys in our dataset.

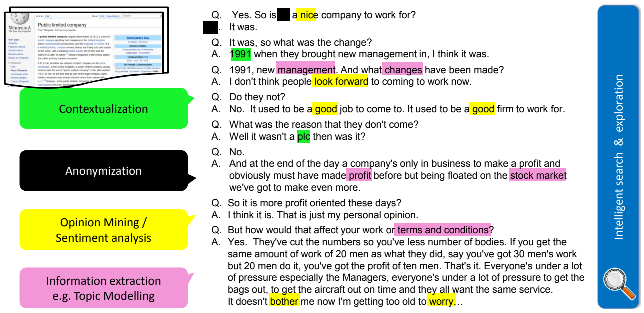
In our previous work we proposed *topic cropping* a fully automated process for selecting and incorporating additional domain-specific documents with similar topical content which can expand a dataset and significantly improve the quality of inferred topic models. We tested our approach on thematically close English and German document corpora and investigated that the produced results for German corpora slightly outperformed those of the English dataset.

Keywords

digital humanities, qualitative data, topic modeling

1. INTRODUCTION

For social sciences, sharing of qualitative primary data like interviews and re-using it for secondary analysis is very promising as data collection is very time consuming. Moreover, some qualitative data sources capture valuable information about attitudes, beliefs, etc. as people had them at other times – “realities” that cannot be captured anymore (see e.g. studies in UK Data Archive). Enabling secondary analysis of data not collected by oneself, analyzing it with new research questions in mind, imposes a lot of challenges, though. In this paper, we focus on the aspect of advanced techniques for facilitating exploration of such data and for improving topic exploration in German digital data archives. Supporting intelligent access to and exploration of data shared for re-use is also a main goal within the digital humanities as it is, for example, expressed in the theme of the Digital Humanities 2013 conference: “Freedom to Explore”. Figure 1 visualizes the main building blocks for intelligent support of secondary qualitative analysis envisaged in this project: Contextualization, Information Extraction, Opinion Mining/Sentiment Analysis, and Anonymization.



Q. Yes. So is [redacted] a nice company to work for?
[redacted]. It was.

Q. It was, so what was the change?
A. 1991 when they brought new management in, I think it was.

Q. 1991, new management. And what changes have been made?
A. I don't think people look forward to coming to work now.

Q. Do they not?
A. No. It used to be a good job to come to. It used to be a good firm to work for.

Q. What was the reason that they don't come?
A. Well it wasn't a plc then was it?

Q. No.
A. And at the end of the day a company's only in business to make a profit and obviously must have made profit before but being floated on the stock market we've got to make even more.

Q. So it is more profit oriented these days?
A. I think it is. That is just my personal opinion.

Q. But how would that affect your work or terms and conditions?
A. Yes. They've cut the numbers so you've less number of bodies. If you get the same amount of work of 20 men as what they did, say you've got 30 men's work but 20 men do it, you've got the profit of ten men. That's it. Everyone's under a lot of pressure especially the Managers, everyone's under a lot of pressure to get the bags out, to get the aircraft out on time and they all want the same service. It doesn't bother me now I'm getting too old to worry…

Figure 1: Modules supporting secondary analysis

By exploiting information retrieval and topic modeling techniques we can mine additional knowledge about themes discussed in primary qualitative data. This way, interview contents can be visualized by means of extracted topics for a quick overview. For example, topics extracted for a collection of studies, cases, or samples show the commonalities of themes while comparing topics of individual studies, cases or samples sheds light on the specifics. Interview topics as well aid an enhanced (automatic) content analysis and retrieval of similar documents. This is especially interesting as qualitative primary documents are often very long, and thus it is hard to easily grasp their thematic coverage – let alone to manually analyze and code them.

Due to the enormous resources required for conducting qualitative research by means of interviews (holding the interview, interview transcription, document coding/analysis), the primary data resulting from such qualitative studies is usually limited to a small number of interviews per study case or sample. Topic models, however, are based on statistics and thus perform better on big data sets (see, e.g. [19]). In our recent work [23] we presented a generalizable framework for using topic modeling given such corpora restrictions as they occur in qualitative social science research. Our fully automated adaptable process tailors a domain-specific Cropping corpus by collecting relevant documents from a general corpus or knowledge base, here Wikipedia. The topic model learned on this substitute corpus is then applied to the original collection. Hence, we exploit state-of-the-art IT-methods adapting and integrating them for usage as research tools for the digital humanities. Our previous experiments were conducted on a dataset of workers interviews in English language.

In this paper we present first topic cropping outcomes for the original German studies. Our results show a slight improvement in quality metrics for the German documents due to, as we believe, some properties of the German language such as wide usage compounds. We plan to evaluate the latter hypothesis in our future work

2. "GUTE ARBEIT": IT TOOLS

Ever since the period of rapid economic growth in post-World War II Germany the working environment has fundamentally changed. The rise of the service sector in industrialized countries, in particular, stimulated discussion about "subjectivization" of work, i.e. changes within post-tayloristic management strategies as well as altered work-ethics and attitudes. By re-analyzing data collected during more than four decades with the help of modern information technologies, Gute Arbeit studies how conceptions of "good work" evolved over time.

However, sharing and re-using data, e.g., for longitudinal (meta)analysis is not common practice so far. Gute Arbeit will enable intelligent access to such qualitative data gathered within diverse sociological studies regarding the workplace. For this, we will adapt and advance computational approaches from the fields of Information Retrieval and Data Mining, thus promoting the area of digital humanities. Gute Arbeit will contribute to the area of digital humanities by, amongst others, providing best practices and guidelines, tools and methodologies on how to facilitate re-use of sensitive primary data as well as on the exploitation of intelligent computational techniques for exploring them.

Mining additional knowledge from such valuable data, re-using it with new research questions in mind or sharing it with other researchers involves many challenges though. Besides warranting participants' anonymity, keeping and visualizing context is crucial to correctly interpret utterances of interviewees not surveyed by oneself. Since "good work" is a subjective concept, a major task will be the automatic extraction of topics and people's opinions regarding these topics. Here, the usefulness of popular text mining strategies like Topic Modeling and Sentiment Analysis has to be proven for the kind of material at hand: qualitative data from structured and unstructured interviews.

In secondary analysis, contextualization is crucial, thus it has taken the center stage in the debate so far. There are different kinds and levels of context of the interview, e.g., conversational, situational, regarding the research project, or institutional/cultural (see [1]). Here, we will focus on using external knowledge bases, e.g. Wikipedia or news corpora, to enrich primary data with background information on the socio-cultural context present at the time of data collection. Information extraction – here of Topics and Named Entities – will be beneficial for advanced exploration and navigation support, to get a quick overview over collection contents, to filter via corresponding facets, or to retrieve similar documents. Exploiting opinion mining for secondary analysis is especially interesting for our project as (1) the sociological research focuses on subjective conceptualizations of work and (2) we assume that opinionated passages with negative or critical statements about certain names or topics may receive special attention with respect to anonymization. In contrast to traditional software for qualitative text analysis, DigDeeper will offer various intelligent tools for searching and exploring (parts of) documents, samples, and collections.

Figure 2 gives an overview of the system architecture.

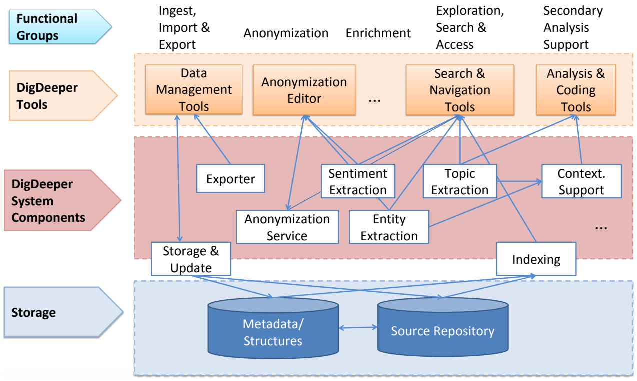


Figure 2: DigDeeper Architecture

Of course, the DigDeeper tool to be developed within the project will also support standard features for qualitative data analysis like coding, annotating, linking, and highlighting.

3. RELATED WORK

Tools for (Secondary) Analysis of Qualitative Data: When it comes to software tools and techniques for supporting the (re-)analysis of qualitative data usually three groups are differentiated. Qualitative data analysis (QDA) tools like ATLAS.ti, MaxQDA, Nvivo are well developed software products enabling the manual coding, annotation and linking of data in a variety of formats. Other features are simple search procedures, the definition of variables, automatic coding of specified text strings, and sometimes also visualization of co-occurrences are features or word frequency counts.

More advanced are tools for (quantitative) content analysis, e.g. General Inquirer, Diction, LIWC, TextPack, WordStat. Software in this category usually builds upon large dictionaries to analyze vocabulary use also semantically. Besides word frequencies, category frequency analysis as well as statistics or filtering for keywords in contexts (KWIC / concordance) are typical features. Programs may offer co-occurrence or correlation analysis of categories or words, ideally accounting for synonyms via the build in dictionaries. Related is cluster analysis and multidimensional scaling for visualizing word or category correlations. Dictionaries can also be used for normative comparison, i.e., to find specifics of vocabulary usage in a document or a collection [12].

Text mining and statistical analysis are advanced techniques exploited to automatically find themes and trends in qualitative data. Tasks are, for example, supervised document classification requiring human input for the label or variable value to be learned, unsupervised clustering of similar documents, or document summarization. Various algorithms as well as standard data preprocessing procedures (stemming, stop word removal, etc.) exist. Via lexicons, patterns and rules information extraction, e.g. of sentiment, can be achieved. To name just a few – mostly commercial – tools that (claim to) provide additional text mining capabilities: Catpac, SAS Text Miner, SPSS TextSmart, WordStat.

In [8], the usage of unsupervised learning methods is discussed, here a self-organizing map (SOM) build upon manually selected terms from interviews, for qualitative data analysis. They argue that such text mining procedures can aid both data-driven, inductive research by finding emergent categories/concepts as well as theory-driven, deductive research by checking the adequacy and applicability of defined schemes. The next section reports in detail on work regarding the related goal of topic modeling for qualitative data – the focus of this paper.

Topic Modeling: Topic modeling is a generative process that introduces latent variables to explain co-occurrence of data points. Latent Dirichlet allocation (LDA) [2] is a further development of probabilistic latent semantic analysis (PLSA) [5] modeling documents using latent topics. LDA was developed in the context of large document collections, such as scientific articles, news collections, etc. with the goal of getting a quick topical overview. The success of LDA led to the application in other domains, such as image processing, as well as other types of documents, e.g. tweets [6] or tags [10].

There is also some work applying topic modeling to transcribed text. In [22], the standard LDA model is extended to identify not only topics but also topic boundaries within longer meeting transcripts. The authors show that topic modeling can be used to detect segments in heterogeneous text. Howes et al. [7] investigate the use of topic models for therapy dialog analysis. More specifically, LDA is applied to 138 transcribed therapy sessions to then predict patient symptoms, satisfaction, and future adherence to treatment using latent topics detected vs. hand coded topics. The authors find only the manually assigned topics to be indicative. Human assessment of the interpretability of the automatically learned topics showed high variance of topic coherence.

Using topic models where there is only limited data, e.g. very short documents or very few documents, has been studied as well. Micro-blogging services, such as Twitter, limit single documents to 140 tokens. Hong and Davison [6] study different ways to overcome this limitation when training topic models by aggregating these short messages based on users or terms. The resulting longer documents yield better topic models compared to training on short, individual messages. Unfortunately, this method only works if the number of short texts is sufficiently large. Using additional long documents to improve topics used for classification was proposed in various approaches: Learning a topic model from long texts and then applying it to short text [21] improves significantly over learning and applying it on short texts only. Learning it on both [24] and applying it on short texts improves further. Jin et al. [9] present their Dual LDA model to model short texts and additional long text explicitly, which outperforms standard LDA on long and short texts for classification. Our focus is not on classification of short documents, we use topic modeling to analyze (long) individual documents and focus more on a careful selection of the corresponding training corpus.

Incorporating domain knowledge for topic transition detection using LDA as is described in [26] addresses this problem using manual selection of training corpora(s). A topic model is trained using auxiliary textbook chapters and used to compare slide content and transcripts of lectures. Because of sparse text on slides and possible speech recognition errors in the transcripts, training a topic model on long, related documents improves alignment of slides and transcript significantly. In contrast, our method does not rely on a manual selection of a training set as cropping is performed in an automated process. Applicability of topic modelling for multilingual IR were identified in [11] the authors attempt to construct accurate and comparable relevance models in the source and target language, and use that models to rank the documents in the target collection. The advantage of this approach is that it does not rely on a word-by-word translation of the query and the relevance of the target collection can be estimated more accurately. In [17] the authors proposed polylingual topic modelling using the Wikipedia interlinked pages. In this work we show that a language could be an important factor for topic modelling and topic cropping quality.

4. EXPERIMENTS

In our experiments we compared German and English document corpora. Both corpora comparably consist of qualitative sociological data, specifically surveys and interviews on topics related to working environment within different industrial areas.

4.1 English Dataset

This corpus consists of qualitative data shared for research purposes via the ESDS Qualidata / the UK Data Archive, which is currently moving to the UK Data Service. We selected four out of the eight cases from the case study on “Changing Organizational Forms and the Re-shaping of Work” [14]. Each case has verbatim transcriptions or summaries of in-depth Face-to-face interviews conducted in England and Scotland between 1999 and 2002. The study surveyed employees from inter-organizational networks as new organizational forms, analyzing how they operate in practice and focusing on the aspect of employment relationship.

- *Airport case*: four airlines, engineering department, airport security, baggage handling, full handling, cleaning company, fire service (30 files online)
- *Ceramics case*: five ceramics manufacturers (32 files)
- *Chemicals case*: a pigment manufacturing plant, two Suppliers, two Transportation specialists, two Business Service Contractors (28 files)
- *PFI case*: Hotel Services Company, Facilities Design Company, Special Purpose Vehicle, NHS Trust Monitoring Team (41 files)

Interviews were held in semi-structured form given guidelines for questions along the main research themes of managing, learning and knowledge development, experience of work, and performance – particularly investigating the links between these topics and changing organizational forms[1]. For example, questions asked for how and why changes in organizational form arose and how much progress has been done on implementation. Regarding learning, interviewees were asked on knowledge and skills required for the jobs, on how and by whom training and learning is organized, or how customer/production pressures are handled. Subjective attitudes and experiences of work were captured via questions on changing patterns in and changing perceptions of team work, working time, pay, contracting, etc. For performance, definition of criteria at different levels, measurement and monitoring as well as source of performance pressure were talked about. In particular, the focus was on links between changing organizational forms and the four broader topics.

4.2 German Dataset

This corpus consist of qualitative data obtained during the time period 2001 - 2009 from the employee interviews of the vehicle manufacturing company “Auto 5000” which was set up inside the Volkswagen complex in Wolfsburg, Germany. This lower cost model company was set up aiming of keeping manufacturing jobs in Germany instead of moving production to other areas of Europe. The stuff was mainly composed of formerly unemployed people and those looking to have more flexible working hours.

The dataset is composed of three parts

- 19 individual interviews with skilled workers (2002)
- 14 individual interviews with production engineers (2003)
- 8 group discussions (2005)

Interviews include the employment history of the former unemployed workers and engineers, shift work and relations between the Volkswagen and “Auto 5000” employees. The average number of the pages per document is about 40.

4.3 Experimental Settings

For tailoring the Cropping corpus, we used top (selected by MI) representative terms identified in the Working corpus analysis phase. The terms were used individually to search for relevant Wikipedia pages using the Bing Search engine. This resulted in a Cropping corpus of about 10.000 documents.

An important parameter in learning the topic model is the number of topics to be learned. With an increasing number of topics, which is a parameter of the topic model learning process, the topics get ever more fine grained. The challenge here is to find a number, which results good topic coverage for the study (all relevant topics are in) and in sufficiently fine grained topics to help in exploring unknown qualitative material, while still being useful for human understanding and for spotting areas with similar topics.

There is no general notion of a “good” number of topics, since this strongly depends on the corpus and the targeted application. We decided to take the diversity of the topics assigned to the study based on the topics learned from the Cropping corpus as a measure for a sufficient number of topics. The intuition behind this is that we need a sufficiently large topic model to cover all aspects of the study. As long as this is not yet reached the diversity still increases with the number of topics. Once the diversity stops increasing substantially the newly added topics are either not relevant for the study or they just provide subtopics by splitting topics, which does not substantially add to the diversity.

5. A GENERAL APPROACH FOR TOPIC CROPPING

The goal of our approach described in [23] is to enable the exploitation of the advantages of topic models, e.g., with respect to capturing latent semantics, even if the considered corpus is too small for their direct application. To obtain this target, in recent work [23] we proposed the topic cropping workflow which is a four step process (see also Figure 3):

1. Analyzing working corpus coverage by selecting characteristic terms
2. Tailoring a Cropping corpus by collecting relevant documents
3. Learning a topic model from the Cropping corpus
4. Applying topic inference to the working corpus

Analyzing Working Corpus Coverage: The goal of this step is to understand the topical coverage of the corpus under consideration. At first glance, this might look like a hen-egg problem: we need to know the main topics of the corpus for building a corpus for learning those topics. For overcoming this, we relied on a method for determining the most relevant terms by using a counter corpus and used the metric of Mutual Information (MI) [13], which measures how much the joint distribution of terms deviates from a hypothetical distribution in which features and categories are independent of each other. The measure ranks higher terms which are frequent in the working corpus but not in general.

[1] For more details see: `http://discover.ukdataservice.ac.uk/catalogue?sn=5041`

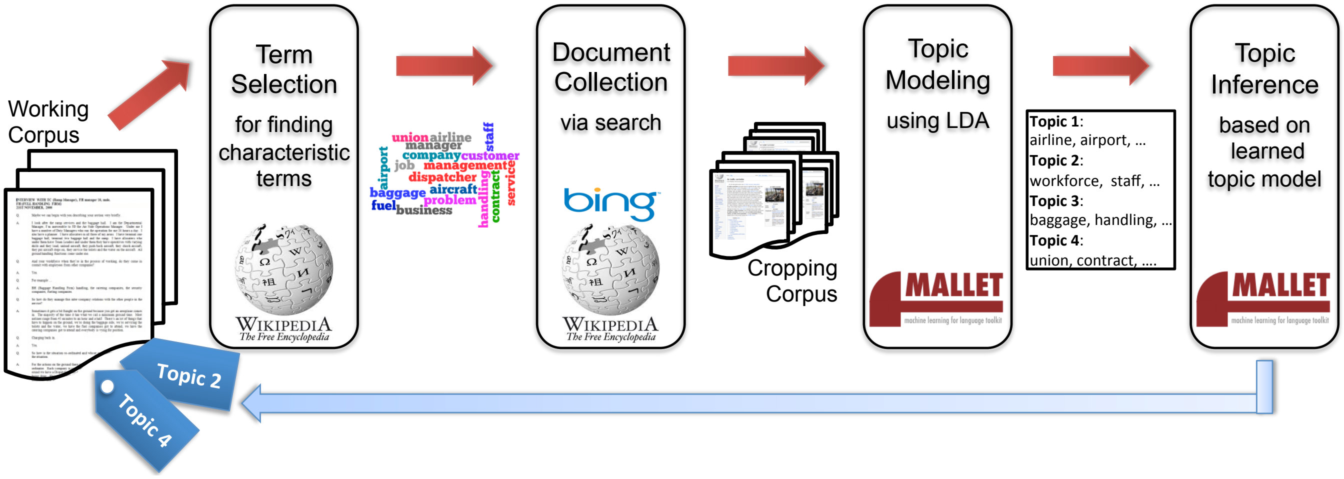


Figure 3: Workflow for Topic Cropping

Tailoring a Cropping Corpus: The top-ranked subset of those terms is used for tailoring the Cropping corpus. We used a general Web search engine to identify the set of highest ranked Wikipedia pages for each of the terms. The Cropping corpus is created from the set union of all those pages. Wikipedia has been selected as the starting point for Cropping corpus creation because of its broad coverage providing information on seemingly every possible topic. Of course it is also possible to use large domain specific corpora or combinations of several corpora.

Learning the Topic Model: We made use of the Mallet topic modeling toolkit [15], namely the class ParallelTopicModel. This class offers a simple parallel threaded implementation of LDA (see [18]) together with SparseLDA sampling scheme and data structure from [25]. LDA is based on a generative probabilistic model that models documents as mixtures over an underlying set of topic distributions.

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$$

where $P(w_i)$ is the probability of the ith word for a given document and z_i is the latent topic. $P(w_i|z_i = j)$ is the probability of w_i within topic j. $P(z_i = j)$ is the probability of picking a word from topic j in the document.

Applying the Topic Model: Using learned models from the previous step, we determine the topics for working corpus using topic inference as offered by the Mallet toolkit (cc.mallet.topics.TopicInferencer). It assigns to each of the topics in the topic model a probability of it being relevant for a study document. As stated in [23], it is not expected that the set of topics learned from the Cropping corpus is exactly the set of topics inherently included in the working corpus. We analyze this issue further in Section 6.

6. EVALUATION

We judge the quality of the automatically detected topics exploiting both, internal (intrinsic) and external (extrinsic) evaluation [13, 20]. In topic analysis an internal evaluation prefers low similarity between topics whilst within a topic high similarity is favored. We adopt this idea by measuring *topic diversity* capturing variance between the different topics in a model and *topic coherence* within the single topics respectively. We additionally measure *topic relevance* externally by comparing with human annotators. In this section, we evaluate both the topics learned directly from the working corpus and those from the Cropping corpus with the same setting and analyze them with respect to these quality dimensions.

6.1 Topic Diversity

Topic diversity is an important criterion for judging the quality of a learned model. The more diverse, i.e. dissimilar, the resulting topics are, the higher will be the coverage regarding the various aspects talked about in our interview data. It has been shown in earlier work that the Jaccard Index is an adequate proxy for diversity [4] and its output value correlates with a number of clusters (topics in our case) within the dataset. Thus, to estimate the average similarity between produced clusters, we employ the popular Jaccard coefficient [13].

Figure 4 shows the change of the average Jaccard similarity, comparing the diversity of topics learned from the working and the Cropping dataset. We observe that topics learned from the Cropping corpus are generally more diverse already in the beginning of the curve, indicating that our approach covers more aspects of the data even for smaller number of topics.

6.2 Topic Coherence

We tackle the task of topic coherence evaluation by rating coherence or interpretability based on an adaptation of the Google similarity distance (NGD), which performs effectively in measuring similarity between words [3]. The more similar, i.e less distant, the representative words within a topic, the higher or easier is its interpretability (see details in [23]).

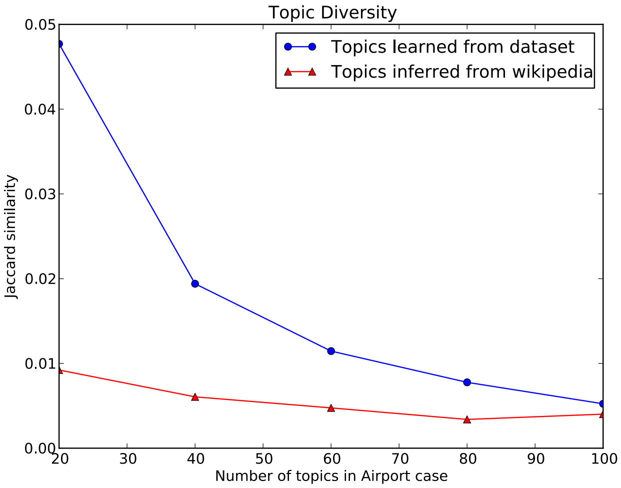


Figure 4: Topic diversity, measured via Jaccard similarity, and its variance for different numbers of topics learned during topic modeling.

Table 1: Example topics with coherence measured via normalized Google distance (NGD), topics inferred from the working corpus (W) or the Cropping corpus (C).

	Topics	**NGD**
English		
W	bag day company baggage	0.44
W	airline service issue baggage	0.38
C	workers labor work employment	0.19
C	employee employees tax employer pay	0.19
German		
W	hörensagen standard schweißerpass block	0.60
W	antwort endeffekt wolfsburg gmbh	0.43
C	arbeitnehmer arbeitgeber gewerkschaften arbeit	0.19
C	unternehmen management ergebnisse mitarbeiter	0.32

For example, for the topic T_i that is presented by a list of words {airline, service, issue, baggage}, its NGD is determined by the average of the scores of all possible word pairs {(airline, service), (airline, issue), (airline, baggage), . . . } (see also Table 1).

To estimate overall topic coherence, we randomly choose a list of 30 learned topics per case ($T = (T_1, ..., T_n)$), compute NGD for each T_j, and then take the average of the list $\mathbf{AvgNGD}(T) = \frac{1}{n}\mathbf{NGD}(T_j)$.

Table 2 reports the average normalized Google distances and their deviations for topics inferred for three English cases as reported in [23] and for the one German case (Auto5000). For both corpora and all cases evaluated, we obtain consistent improvement. This indicates that the topics inferred from the German and English Cropping corpus are also significantly more coherent than those only learned directly from the working corporas (measured significance of a t-test $p < 0.001$).

Table 2: Average (Avg) and standard deviation (SD) of topic coherence of three cases, measured via normalized Google distance (NGD). Topics are inferred from the working corpus (W) or the Cropping corpus (C).

Case	AvgNGD$_W$	SD$_W$	AvgNGD$_C$	SD$_C$
Airport	0.34	0.07	0.21	0.08
Ceramics	0.32	0.08	0.25	0.09
Pfi	0.35	0.1	0.22	0.08
Auto 5000	0.38	0.09	0.29	0.08

6.3 Topic Relevance

While topic diversity and topic coherence can help to estimate the quality of the topics with respect to information-theoretic considerations, validity of our results, i.e., the usefulness of the derived topics for the working corpus, needs to be assessed by human evaluation of topic relevance. Here, we decided to compare our inferred topics with topics assigned by human annotators. For this evaluation, we randomly selected 16 and 8 documents from English and German corpora respectively to be manually annotated by five users. Each document was split into smaller units – typically question and answer pairs – resulting in about 60 units per document. Thus, a total of 1500 units was annotated. We asked users to define topics discussed in each given unit. Each unit could have one or more topics and there were no restrictions on how topics are to be phrased. Typically the topics assigned were single words or short phrases.

Topic relevance is then assessed by automatically matching user defined topics with the learned ones. For this, the terms used by the user for a topic are matched with the top terms learned for a topic by the topic model. We consider it a match if the term used by the user appears in the top terms of the respective topic. By design, this evaluation gives preference to the topic model learned directly from the working corpus since the users tend to use terms that appear in the text. Similarly, the topic models learned directly on the working corpus use exactly those terms for their topics. In order to even out this terminology disadvantage, for English dataset we made use of word synonyms from WordNet [16] to extend sets of topic words before matching. Due to the lack of German WordNet and the language property, we compute these scores for German corpus without any synonym extensions. A learned topic T is considered to be relevant if its representative words and their synonyms $\mathbf{w} = (w_1, ..., w_k)$ share one or more terms with user defined topics $\mathbf{t} = (t_1, ..., t_r)$

$$\mathbf{Rel}(T) = \begin{cases} 1 & \text{if } |\ \mathbf{w} \cap \mathbf{t}\ | > 0 \\ 0 & \text{otherwise} \end{cases}$$

Figure 5 compares topics learned from the documents in the English corpus with respect to the number of relevant topic at rank k, $\mathbf{R}@k = \sum_{i=1}^{k} \mathbf{Rel}(T_i)$, where the rank is determined by the probability of the topic assignment (resulting from topic inference). Similarly, Figure 6 presents the relevance results for the German corpus. It can be seen from the results that the topics learned from Wikipedia reach a comparable level of relevance as those learned directly from the corpus, while being more coherent and diverse.

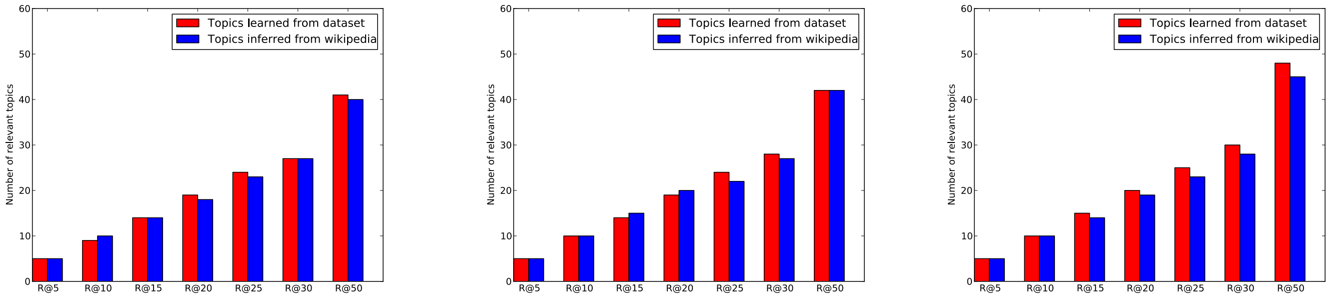


Figure 5: Topic relevance as the number of relevant topics at rank k, for the documents in English dataset

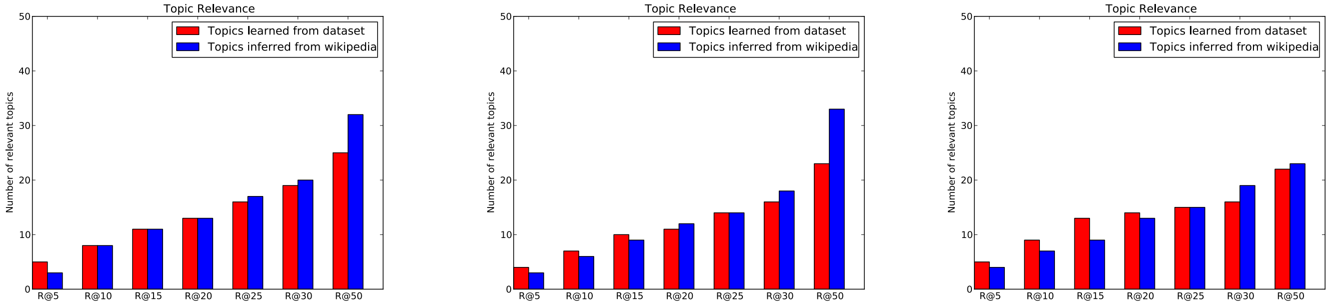


Figure 6: Topic relevance as the number of relevant topics at rank k, for the documents in German dataset

Table 3: Example topics inferred from the German working corpus (W) and the Cropping corpus (C).

Corpus	Topics
W	magdeburg erwartungshaltung april fliessbandarbeit rad blechteile ruck garderlingen bildungsträgern
W	art endeffekt umfeld niveau stendal nummer automobilbauer not bahn
W	antwort fachtalent autos umschulung mal jungs band hammer mechaniker
W	test aufgaben gestaltungswerkzeuge kinderbetreuung leistung maß wissen maschine leiter
W	monat ahnung bereich lack betriebsingenieur brief band fachwissen halle
C	dresden chemnitz dresdner zwickau sachsen radebeul clear style div
C	französischen paris frankreich französische saint jean louis dreyfus les
C	münchen deutscher geboren hans karl friedrich archäologe august verstorben
C	film filme films rolle regisseur filmen schauspieler regie
C	formula verfahren test unternehmen methoden management ergebnisse mitarbeiter methode

6.4 Results for the German Corpus

We conducted the Topic Cropping Procedure for the German corpus and obtained comparable results. Also in this case both, the diversity between topics and the coherence within each topic were increased. Additionally we noticed a slight increase in the relevance, meaning that inferred topic slightly better reflected the users annotations of German compared to the English dataset. In this paper we report the results and let the further investigations about the reason for the increase to the future work. A hypothesis is that the improvement could be due to a particular property of the German language the use of compounds. Compound is a word which consists of more than one word. English examples of compounds are words like: “smalltalk”, “makeup”, “notebook” and so on. In German language it is usual to use compounds and create them “on-the-fly”, if necessary. The English phrase “car body pressing” turns into a single word in German - “Karosseriebau”. Considering topic cropping strategy, the German word is more concise, as a query resulting in documents well focused on the particular topic. Compared to English case, the terms “car”, “body”, “pressing”, each would give a large number of noise documents when searching in Wikipedia. This hypothesis is supported by the fact that German queries in our experiments generally led to less Wikipedia pages, however the relevance of the pages was obviously higher compared to English dataset using our relevance annotation measure described in 6.3.

The Table 3 provides some examples for topics obtained from the working corpus only(“W”) and the Cropping corpus (“C”). In the future work we plan to evaluate the outcomes in more details, however already on the first glance, for German speaking person it will be more difficult to identify and label the topics obtained using straight forward modelling on original dataset compared our cropping approach. For example, the labels for “C” could be (top-down) “East Germany”, “France”, “archaeology”, “movie”, “company management”, whether the labels for “W” are hard to extract.

7. CONCLUSION AND FUTURE WORK

In our recent work [23] we proposed a method for a *fully automated* adaptable process of tailoring a domain-specific sub-corpus from a general corpus (e.g. Wikipedia).

In this paper we present first results of an application of our Topic Cropping approach within the German national BMBF Project “Gute Arbeit” and a large scale qualitative interviews in German language. Our experiments show slight improvements of the results for German dataset and it seems it is due to some specific language properties.

We believe that wide usage of compounds in German language can lead selecting more concise representative query terms for the related document search in Topic Cropping and plan to investigate these details in our future work.

8. ACKNOWLEDGMENTS

The work was supported by the project *“Gute Arbeit” nach dem Boom (Re-SozIT)* funded by the German Federal Ministry of Education and Research (BMBF) (01UG1249C) and by the European project ForgetIT (GA600826). Responsibility for the contents lies with the authors.

9. REFERENCES

[1] L. Bishop. A proposal for archiving context for secondary analysis. *Methodological Innovations Online*, 1(2):10–20, 2006.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] R. L. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383, 2007.

[4] F. Deng, S. Siersdorfer, and S. Zerr. Efficient jaccard-based diversity analysis of large document collections. In *Proceedings CIKM*, pages 1402–1411, 2012.

[5] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings UAI*, pages 289–296, 1999.

[6] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings 1st Workshop on Social Media Analytics*, SOMA, pages 80–88, 2010.

[7] C. Howes, M. Purver, and R. McCabe. Investigating topic modelling for therapy dialogue analysis. In *Proceedings IWCS Workshop on Computational Semantics in Clinical Text (CSCT)*, pages 7–16, 2013.

[8] N. Janasik, T. Honkela, and H. Bruun. Text mining in qualitative research: Application of an unsupervised learning method. *Organizational Research Methods*, 12(3):436–460, 2009.

[9] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings CIKM*, pages 775–784, 2011.

[10] R. Krestel, P. Fankhauser, and W. Nejdl. Latent Dirichlet Allocation for Tag Recommendation. In *Proceedings RecSys*, pages 61–68, 2009.

[11] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In *Proceedings of SIGIR '02*, New York, NY, USA, 2002.

[12] K. H. Leetaru. *Data Mining Methods for the Content Analyst: An Introdution to the Computational Analysis of Content*. Routledge, New York, USA, 2012.

[13] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[14] M. Marchington, J. Rubery, and H. Willmott. Changing organizational forms and the re-shaping of work : Case study interviews, 1999-2002 [computer file], 2004.

[15] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[16] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[17] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. Mccallum. Polylingual topic models. In *In EMNLP*, 2009.

[18] D. Newman, A. U. Asuncion, P. Smyth, and M. Welling. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10:1801–1828, 2009.

[19] D. Newman, E. V. Bonilla, and W. Buntine. Improving topic coherence with regularized topic models. In *Proceedings NIPS*, pages 496–504, 2011.

[20] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Proceedings Human Language Technologies*, HLT, pages 100–108, 2010.

[21] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings WWW*, pages 91–100, 2008.

[22] M. Purver, K. P. Körding, T. L. Griffiths, and J. B. Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings ACL*, pages 17–24, 2006.

[23] N. K. Tran, S. Zerr, K. Bischoff, C. Niederée, and R. Krestel. Topic cropping: Leveraging latent topics for the analysis of small corpora. In *Proceedings of TPDL 2013*, LNCS. Springer: to appear., 2013.

[24] G.-R. Xue, W. Dai, Q. Yang, and Y. Yu. Topic-bridged plsa for cross-domain text classification. In *Proceedings SIGIR*, pages 627–634, 2008.

[25] L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings KDD*, pages 937–946, 2009.

[26] X. Zhu, X. He, C. Munteanu, and G. Penn. Using latent dirichlet allocation to incorporate domain knowledge for topic transition detection. In *Proceedings INTERSPEECH*, pages 2443–2445, 2008.

Stereotype or Spectrum: Designing for a User Continuum

Mark Sweetnam, Micheál Ó Siochrú
Dept. of History
Trinity College Dublin, Arts Building,
Dublin, Ireland
{sweetnam,osiochrm@tcd.ie}

Maristella Agosti, Marta Manfioletti
Dept. of Information Engineering
University of Padua
Via Gradenigo, 6/a
Padua, Italy
{agosti, manfioletti}@dei.unipd.it

Nicola Orio, Chiara Ponchia
Dept. of Cultural Heritage
University of Padua
Piazza Capitaniato
Padua, Italy
nicola.orio@unipd.it,
ponchia.chiara.1@studenti.unipd.it

ABSTRACT

This paper presents work carried out on developing a methodology for identifying and characterising a spectrum of users for a corpus agnostic environment designed to facilitate research into cultural heritage collections. It outlines the approach taken to the design of the CULTURA environment, and addresses the challenges and opportunities presented by this broadly-based user engagement. It also outlines how adaptive strategies have been used within the CULTURA environment to ensure that it effectively addresses the needs of its different user communities.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries - *collection, dissemination, systems issues, user issues.* H.3.5 [**Information Storage and Retrieval**]: Online Information Services - *data sharing, Web-based services.*

General Terms

Management, Design, Experimentation, Human Factors.

Keywords

Cultural heritage, cultural heritage collections, digital cultural heritage collections, digital humanities, different user communities, user studies, adaptivity.

1. INTRODUCTION

The drive to digitize material of historical and cultural significance has largely been underpinned by the twin concerns of conservation and access. The imperative to preserve unique and delicate resources by producing digital surrogates has been, and continues to be, an important driving force for digitisation projects. Important though preservation undoubtedly is, access is more important still. Indeed, the act of conservation implies a need for access – we may access material that we do not bother to preserve, but we do not, in a world of limited and diminishing resources, preserve what we do not wish to access.

To date, the need for access tends to have been handled at institutional or collection level. This access is most often facilitated through online web interfaces, that allow the collection to be searched and records and resources to be viewed.

These interfaces are limited in a number of significant ways. They tend to be specialised, search-based, superficial, and stereotyped. They are specialised because they, most commonly, handle material from one collection, or from one kind of collection. Typically, they require the user to rely on some – more or less complicated – variety of Boolean search to find the results they need. These results are presented in a superficial way – typically as a list of items, without any opportunity for deeper exploration or discovery. And these interfaces are stereotyped in terms of the sort of user they expect and assist. They are designed to address the needs of a monolithic user, whose expectations, experience, and interests are rigidly and unalterably defined.

2. THE CULTURA PROJECT

The CULTURA project[1] has been designed to address each of these limitations (Hampson *et al,* 2012a; Hampson *et al,* 2012b; Bailey *et al*, 2013). The research environment that is being developed by the project is not specialised for a specific collection, but it has been developed using two very different core collections. The 1641 Depositions, which record witness accounts of atrocities committed in Ireland during the 1641 Rebellion, constitute a textual corpus, which has been augmented by manually generated metadata[2]. The Depositions are marked by highly challenging linguistic content, with wide variation in orthography, syntax, and punctuation. The IPSA collection of illuminated medieval manuscripts, a virtual collection of scientific works, including herbals and astrological codices, is a purely visual collection, with extensive metadata[3]. The CULTURA environment has been designed to address the needs of users of both these collections, but also to generalise beyond these specific corpora to a wide range of cultural heritage collections.

In supporting exploration of this range of material, CULTURA supports traditional search-based exploration, but also moves well beyond it. It harnesses a range of innovative normalisation (Lawless *et al,* 2013) and natural language processing technologies that allow entities and relationships to be extracted from the collection and visualised using a range of specially designed visualisations. It also provides for entity-oriented search (Carmel *et al*, 2012), and allows users to crosswalk from one tool to another, ensuring that their exploration of the collection is flexibly supported. The environment also provides a comprehensive set of logging, bookmarking, and

ENRICH'13, August 1, 2013, Dublin, Ireland.

[1] http://www.cultura-strep.eu/

[2] http://1641.tcd.ie/about.php

[3] http://ipsa.dei.unipd.it/en_GB/home

annotating tools that make it a powerful aid to both extensive and intensive work on content collections.

CULTURA, then, makes it possible to work on a range of material. It allows a variety of different investigations and explorations to be undertaken by users. The richness of this functionality is both an opportunity and a challenge. It is an opportunity, because CULTURA has the power and potential to empower a wide range of user types, and a challenge because there is the ever-present danger of the interface buckling under the weight of its own complexity, overwhelming the user with options, and frustrating all but the most devoted and fanatical investigator. The CULTURA project addresses this challenge through its implementation of adaptivity, and its emphasis on addressing a spectrum of users, rather than a single stereotyped user.

3. CHARACTERISING THE USER CONTINUUM

From its inception, CULTURA was conceived as a project that would be driven by the requirements of humanities users, rather than by the pursuit of technological novelty. Thus, sustained interaction with potential users before, during, and after the process of development were a crucial part of the project.

Equally essential to the project was the need to stimulate and support the communities of interest, which form around digital humanities and cultural heritage collections. These communities include a diverse mixture of professional researchers, apprentice investigators (e.g. students of history and art history), informed users (e.g. users belonging to relevant societies or interest groups, cultural or authorities) and members of the general public (both adults and children) with diverse interests and motivations.

Correctly characterising users is an essential element of ensuring that the interface with which they are presented addresses the needs of the user, and so accurate and rigorous user characterisation is a *condicio sine qua non* for the successful implementation of the CULTURA environment. In order to achieve this aim, humanities researchers from the Department of History at Trinity College Dublin – Ireland (TCDH) and the University of Padua – Italy (UNIPD) developed the following taxonomy that is exploited to guide their interactions with users and to inform the design and evaluation of the CULTURA environment:

- **Professional researchers** – established academics, experienced in the general area covered by the resource, *but not necessarily with the specific content of the resource*.
- **Apprentice investigators** – students at advanced undergraduate and post-graduate level. Some knowledge of the historical period and/or cultural context addressed by the resource.
- **Informed users** – researchers who are not professional academics but have knowledge of some aspect addressed by the resource.
- **General public** – adults and children.

Detailed interactions with these users groups, using a range of methodologies including surveys, focus groups, and one-to-one interviews, laid the basis for the design of the CULTURA environment. This approach involved interaction with users at a number of different levels.

For users of all types, focus groups and interviews were of fundamental importance. This was true at the beginning of the design process, when potential users were asked to reflect on their needs, wishes, and preferences for a research environment. Existing users of the 1641 Depositions and of the IPSA material were surveyed to provide their input.

In the case of apprentice researchers, CULTURA researchers have been able to take advantage of the opportunity to work closely with groups of under- and postgraduate students who are involved in the sustained use of the system. In the case of the 1641 Depositions, students in an M.Phil. module based on the Depositions were required to use the CULTURA environment for collaborative and solo research projects. This has been carried out over two years, and a third repetition of the exercise is planned for next year. In the case of IPSA, undergraduate students carried out a number of retrieval tasks on the collection as part of their class assignments (Agosti *et al*, 2012).

In addition to carrying out detailed interviews with professional researchers in a variety of related disciplines, humanities researchers involved in the CULTURA project have undertaken a number of research projects using the research environment. These have resulted in a number of humanities-focused publications (Ó Siochrú and Sweetnam, 2012; Orio and Ponchia, 2013; Sweetnam, 2013a; Sweetnam 2013b), and have been an invaluable means of validating the real life usefulness of the CULTIRA environment.

The experience of project researchers working alongside 'real life' users of the resources has proved tremendously valuable. Where possible, CULTURA has made use of sustained interaction with users. Manifestly, this sort of in-depth interaction is possible only with small cohorts of users. Therefore, where possible, this deep and narrow interaction has been supplemented with broader interactions, typically based around a number of clearly defined tasks. These exercises have particular value in terms of evaluating usability, and are a useful supplement to the sort of detailed engagement described above.

Online surveys play an important role in gathering the input required from users and evaluators. They provide a reliable means of systematically collecting comparable evaluation data, and allow responses to be gathered from users who may be spread over a wide geographical area. The focus of evaluation was on usability and usefulness, and for this reason our starting model has been the triptych model (Tsakonas and Papatheodoru 2006), although a number of interesting measures have been introduced in the literature. such as navigability, searchability, coverage and, in particular, retrievability (Azzopardi and Vinay, 2008).

In addition, because surveys can be partially completed, left, and then completed, this approach allows users to record their impressions at a time and pace that suits their needs. This is a vital feature when unpaid evaluators are used, as it prevents excessive trespass upon their good will.

On the basis of these interactions, use cases were developed and user requirements elicited and refined (Sweetnam *et al*, 2012). One of the key results of this process was a catalogue of user requirements which, in addition to outlining the features required, also recorded the user groups for which each feature was important, and how important each feature was in the case of each group. One of the findings to emerge very clearly from this exercise was the fact that less experienced users ranked more highly tools that allowed them to explore the content collections in a relatively undirected way. By contrast, apprentice and professional researchers were far more likely to require tools that took they directly to specific artefacts, or sets of artefacts that were relevant to their interest.

Engagement with users has been an on-going and iterative process, and user input has guided the development of the environment as a whole, and the individual tools that it comprises. So the development of the environment was and it is user-driven as it is depicted in Figure 1.

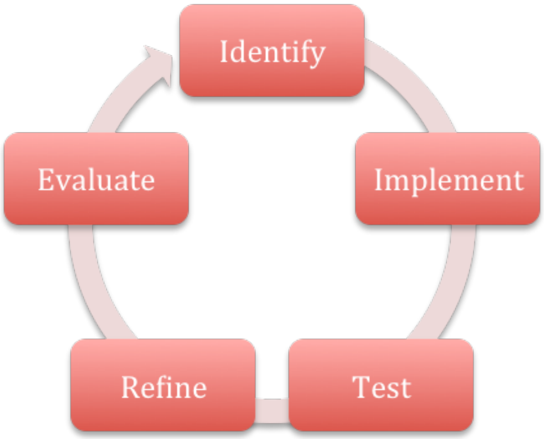


Figure 1. User-driven development.

4. THE CHALLENGES OF THE USER CONTINUUM

The decision to abandon the convenient fiction of a single stereotypical user and embrace the messy, muddled, and demanding heterogeneity of real world users presented real challenges, which had to be overcome or circumvented in the execution of CULTURA's intended approach.

4.1 User Availability

The most basic challenge faced by the humanities researchers was that of locating sufficiently large cohorts of users in each of the categories. These challenges were not uniform across the user spectrum. For example, apprentice investigators – essentially undergraduate and Masters level students – proved relatively easy to recruit, being both plentiful and easily compelled. At TCDH a Masters course on the 1641 Depositions was used throughout the project's lifecycle, as a forum for detailed student interactions with the CULTURA environment, while at UNIPD students in humanities taking classes in computer science used the CULTURA environment as case study for experiencing the interaction with digital collections. These students are a captive audience, but they are also an interested and motivated group of users, who typically valued the opportunity to contribute to the development of a new tool.

In a university context, professional researchers, too, are relatively easily recruited. A number of professional researchers were directly involved in CULTURA, and their insights were crucially important. However, it would be unwise to treat any small group of researchers as though they embodied the Platonic ideal of the humanities researcher, and so it was essential to seek wider input. In an era of spiraling workloads, however, researchers seeking input from expert researchers make heavy demands on the goodwill of their colleagues. In the case of those researchers who work directly on the collections, the promise of exclusive advance access to the interface constitutes something of a *quid pro quo*. This becomes a less compelling argument in the case of non-domain professional researchers. Encouraging involvement from users of this type involves an appeal to their goodwill. Researchers seeking this sort of input require considerable charm, and excellent powers of persuasion. A brass neck, too, does not go amiss.

Informed users are typically not found in a university setting. For both of the CULTURA collections, cultural institutions were an important element of this group. Both TCDH and UNIPD are fortunate to enjoy close relationships with cultural institutions – aided, in Trinity's case, by the close geographical proximity of most of the major Irish cultural institutions to the College. Other interested users differed – for TCDH local historians and genealogists were important, while UNIPD looked to local botanical and astronomical societies, and to members of *Salvalarte,* a voluntary association of individuals who share a great interest in History of Art and Cultural Heritage in general, and which is devoted to the preservation of Paduan culture (Agosti *et al*, 2013). Achieving coherent interaction with these groups required considerable effort. In addition, some unexpected hurdles were encountered. In the context of 1641, for example, TCDH researchers discovered that some professional genealogists were only willing to contribute to the project if they received financial remuneration. While this may be feasible in some projects, in some special circumstances, it is unlikely that many project budgets will allow for user groups to be rewarded in this way.

The reaction of these genealogists does highlight the importance of planning user interactions carefully, in order to ensure that benefit is shared as equally as possible between the project and the participants. While it may not be appropriate to offer financial incentives, it is vital that evaluators feel that they are benefitting from their involvement in the project. That benefit may be defined in terms of advance access to new functionality, or – more vicariously – as an opportunity to contribute to an important advance for scholars in their field, but if high quality feedback is to be obtained, it is essential that users feel a degree of investment in the success of the project.

The need for coherent feedback from informed users led researchers at TCDH to establish a user panel. This group was made up of three representatives each from historical societies, cultural institutions, and genealogists. This small group meant that it was possible to gather consistent, sustained feedback on the developing versions of the CULTURA environment and its tools.

The fourth segment of the user continuum proved, by a considerable margin, the most difficult to address. "General Public" is something of a catch-all category and, by its nature, is heterogeneous and poorly defined. Nonetheless, it is important that the views of these users are captured. In many ways, these are the most demanding users to design for. Their interaction with the system is likely to be casual, and they are easily deterred by obstacles to access. Thus, they provide an acid test for functionality and usability. The CULTURA project has addressed the challenges of interacting with this constituency in a number of ways. Working with schoolchildren was identified as a priority for the project, and researchers were able to develop links with schools in order to provide a number of workshops, which introduced students to CULTURA. These were followed by feedback and evaluation sessions. In addition, the CULTURA environment was offered in a limited public release. User response was gathered using logging, on-page evaluation, and surveys, enabling a holistic picture of the users' experience to be captured.

This sort of broad-based, intensive interaction with users is far more demanding than the sorts of limited user testing of a near complete product that are typical of many digital humanities projects. However, the effort is abundantly justified by the results. The value of user input throughout design and implementation cannot be overstated if the aim is to provide these users with genuinely useful tools that meaningfully assist in investigating and understanding digital cultural heritage collections.

4.2 User Modeling

One of the key challenges addressed by this project is the task of constructing an accurate picture of the characteristics of each user constituency. Beyond this, however, is the challenge of translating those characteristics into a set of technological approaches to address the needs of these users. Both these tasks are complicated by the fact that users do not fall neatly into carefully gradated compartments. Rather, they occupy a continuum, and drawing the line dividing one category of users from another is necessarily and inevitably a somewhat arbitrary exercise.

In addition, there is no reason to suppose that users will only ever occupy one position on that user continuum. Indeed, it is a key aspiration of the CULTURA project to cultivate knowledge and understanding – in other words, to move users along – and up – that continuum. This progression is a feature of the sort of content collections that CULTURA will support. Many, if not most, professional researchers will have encountered material first of all as apprentice researchers, if not as members of the general public. Even if the user who traverses the whole continuum is something of an exception, the CULTURA interface must support their experience at each point.

It is also worth stressing that the humanities user continuum is not hierarchical. No one category of user is more important than any other. The users targeted by CULTURA are not Russian dolls, ranging in significance from the general public up to professional researchers. And if this is true of their significance, it is also true of the complexity of addressing their requirements. Supporting professional researchers is a demanding enterprise, but, then, so too is providing assistance to members of the general public. Indeed, the complexity of supporting the broad heterogeneity of the requirements of the general public is equal to, if different from, supporting the well-defined requirements of professional researchers.

For these reasons, we have chosen not to take what might have seemed an obvious approach to tailoring the CULTURA interface to different types of user. One possible approach to this would have been to define a rich feature set for the professional researchers, and then systematically deplete these features in order to offer a progressively simpler and more streamlined experience to others types of user.

Instead, it has become apparent that CULTURA needs to follow the inverse of this approach. The imperative to support and enable user development means that all users must have access to the full feature set. In addition, some user groups require additional support, and this support is most needed at the general public/interested user end of the user continuum. Users in these categories have consistently highlighted the value of "background briefing" type material to contextualise and illuminate the material in the collection.

4.3 User Support

The results of a great deal of user research – and perhaps of the best sort of user research – are often striking but, at the same time, entirely unsurprising. One of the findings from the user research carried out by CULTURA humanities partners falls firmly into this category. Our interactions with researchers have provided overwhelming confirmation of the – admittedly fairly obvious – fact that researchers are different. Even when working with a small and apparently homogeneous group of researchers who share similar methodological and subject interests, the variety of working and research styles adopted is striking.

To succeed as a research environment, CULTURA needs to accommodate as wide a range of working styles as possible. Too many research interfaces seem to operate as a methodological Procrustean bed, forcing users to conform to the tool, rather than the tool responding to and supporting the individual needs of the user. Thus, the CULTURA interface needs to offer the flexibility to support researcher's work styles. Users from all categories have consistently stressed the necessity to be able not just to locate and annotate the material that is relevant to their research, but to organize it within the interface.

And these users have also provided a comprehensive list of suggestions for achieving this customizability. Most commonly, they have requested the ability to create different workspaces for individual research projects, and a folder functionality that will allow bookmarks and annotations to be categorized and organized. They have requested a high level of control over these folders and projects. They require the ability to copy and move bookmarks and annotations, to allow projects to inherit earlier annotations, but also to start a new project from scratch, or to integrate only some of their existing annotations.

Users have also highlighted the usefulness of being able to export data from the environment. This feature means that CULTURA can be more than a self-contained reservoir – it can become part of a pipeline, a link in a longer chain of analysis, and a cog in a larger research machine. As the tools that contribute to CULTURA are developed, it will be important to ensure that they, make themselves available to wider interoperability and "mashing up". These requirements are important, because researchers are much more likely to make use of an interface that adapts to their requirements, and that allows them to carry out their research as efficiently as possible, without forcing them to change or compromise their preferred approach to research.

5. DEPLOYMENT OF THE COLLECTIONS IN THE ENVIRONMENT

Both of the content collections that are featured in the CULTURA environment have previously been made available by means of more traditional web interfaces and applications.

In the case of the 1641 Depositions, the transcribed text and high resolution page images were made freely available for public access at http://www.1641.tcd.ie/. This site allowed for browsing of the collection, for full-text searching, and for searching across the manually generated metadata. Limited bookmarking was also provided. In the pilot stage of CULTURA, a subset of the depositions data, comprising depositions relating to a single county, was incorporated into the CULTURA environment. Once initial testing had been completed, this coverage was expanded to include the Depositions in their entirety.

Similarly, in the case of the IPSA material, the illuminations were made available through a traditional web interface and application, with search functions and the possibility to link different images. Following the commencement of CULTURA, an analysis of the IPSA content was made between May and October 2012 to choose a significant subset of metadata to be imported and represented in the CULTURA environment for use as a case study to test the new environment and its functions.

The 1641 Depositions and the IPSA collection provide two very different corpora, on which the development of CULTURA can be carried out. As already discussed, this is crucial to the corpus-agnostic principle that underlies the design of CULTURA. In addition, the existence of two pre-existing applications provided a very valuable context for the sort of enhancement being carried out in the CULTURA project. The initial implementations of both 1641 and IPSA provided valuable baseline data on user needs and requirements. Even more crucially, the fact that established user communities had developed around each collection meant that researchers had ready access to experienced users of both collections who were clearly aware of the strengths and weaknesses of the existing modes of interacting with the collections.

6. ADDRESSING USERS WITH ADAPTIVITY

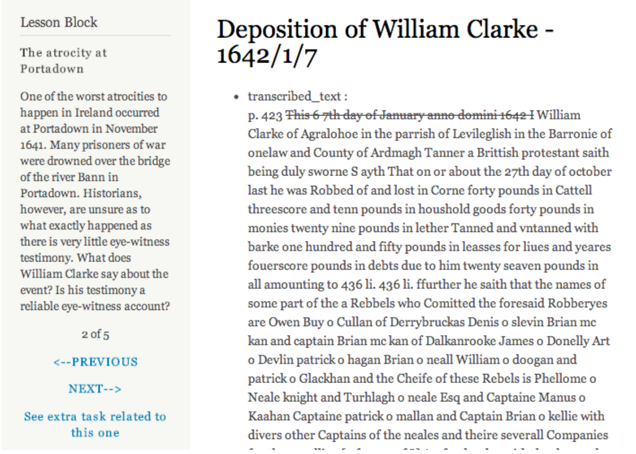
Characterising a diverse range of users, capturing their requirements, and modeling their work process are all vitally important steps in CULTURA's user-driven design. This work would count for little, however, if the resulting environment failed to satisfy these users, to address their requirements, to support their workflows, and to enable them to view culturally significant collection in new and rewarding ways. In order to achieve these ends, the CULTURA environment adopts an adaptive architecture that presents the user with both the content and the tools that are relevant to their investigation at any given time. Adaptivity is an essential part of CULTURA's design, but in the context of this paper there are two aspects especially concerned with supporting the range of users at either end of the spectrum, and we will confine our consideration to these.

6.1 Narratives

Narratives are a central part of the way in which adaptivity is implemented within the CULTURA environment. The term narrative is used 'to represent the adaptive flow of concepts that are woven together to make a coherent offering to a user. Individual concepts may be grounded with either content or services, or may be further refined with the execution of a sub-strategy' (Conlan *et al*, 2013). These narratives are threads through the collection, linking artifacts and tools related to a particular topic. Expert researchers, guided by the use cases and user requirements outlined during user consultations, designed narratives. They used their specialist knowledge of the collections to create a series of threads through the content. Each narrative has a number of levels. Less expert users are offered a relatively high level narrative, but as users interact with the resources that are presented to them, the system dynamically exposes additional material, resulting in a more complex user experience. These narratives allow for an open-ended and developing engagement with the resource collections, and support the development of users from members of the general public, up to apprentice investigators, or professional researchers.

An example of a high level narrative designed for less expert users are "lessons". These describe the different steps of a short course on a specific topic that is dealt within the collection. Typically, the relevant material will be spread across the diverse parts of the different components of a collection. Adaptive lessons provide structured routes through the collection, exposing the user to artifacts that are relevant to their topic of interest.

These narrative lessons have been implemented for both collections in CULTURA, and the user experience of these narratives has validated their usefulness for both types of content collection. In the case of 1641, a series of lessons were developed for use with schoolchildren who were encountering the 1641 Depositions for the first time. These users were presented with a sidebar containing a brief explanation of the context of the individual deposition being viewed, along with prompts for further research (see Figure 2). Users are able to move backwards or forwards within the lesson, or to branch off into further narratives, covering particular areas that especially pique their interest. At any point, the user can leave the narrative pathway to carry out their own detailed investigations. They are able to resume the lesson when desired.



Lesson Block

The atrocity at Portadown

One of the worst atrocities to happen in Ireland occurred at Portadown in November 1641. Many prisoners of war were drowned over the bridge of the river Bann in Portadown. Historians, however, are unsure as to what exactly happened as there is very little eye-witness testimony. What does William Clarke say about the event? Is his testimony a reliable eye-witness account?

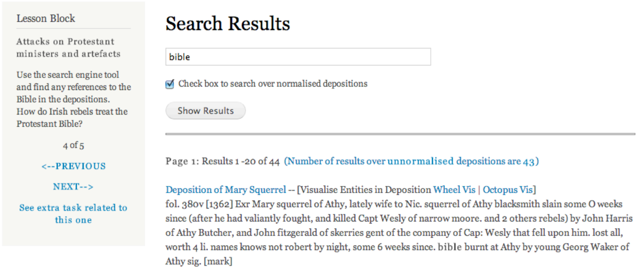
2 of 5

Deposition of William Clarke - 1642/1/7

- transcribed_text :
p. 423 ~~This 6 7th day of January anno domini 1642 I~~ William Clarke of Agralohoe in the parrish of Levileglish in the Barronie of onelaw and County of Ardmagh Tanner a Brittish protestant saith being duly sworne S ayth That on or about the 27th day of october last he was Robbed of and lost in Corne forty pounds in Cattell threescore and tenn pounds in houshold goods forty pounds in monies twenty nine pounds in lether Tanned and vntanned with barke one hundred and fifty pounds in leasses for liues and yeares fouerscore pounds in debts due to him twenty seaven pounds in all amounting to 436 li. 436 li. ffurther he saith that the names of some part of the a Rebbels who Comitted the foresaid Robberyes are Owen Buy o Cullan of Derrybruckas Denis o slevin Brian mc kan and captain Brian mc kan of Dalkanrooke James o Donelly Art o Devlin patrick o hagan Brian o neall William o doogan and patrick o Glackhan and the Cheife of these Rebels is Phellome o Neale knight and Turhlagh o neale Esq and Captaine Manus o Kaahan Captaine patrick o mallan and Captain Brian o kellie with divers other Captains of the neales and theire severall Companies

Figure 2. 1641 Lesson plan.

In addition to tying together chains of documents, these narrative lessons can include the other tools that contribute to the CULTURA environment. So, for instance, a user of the 1641 Depositions who is following a lesson based on a particular individual or location can be presented with the results of a free text search carried out over normalised or unnormalised data (see Figure 3) or of an entity-oriented search displaying the relationship of the entity in question to other entities within the collection (See Figure 4).



Lesson Block

Attacks on Protestant ministers and artefacts

Use the search engine tool and find any references to the Bible in the depositions. How do Irish rebels treat the Protestant Bible?

4 of 5

Search Results

bible

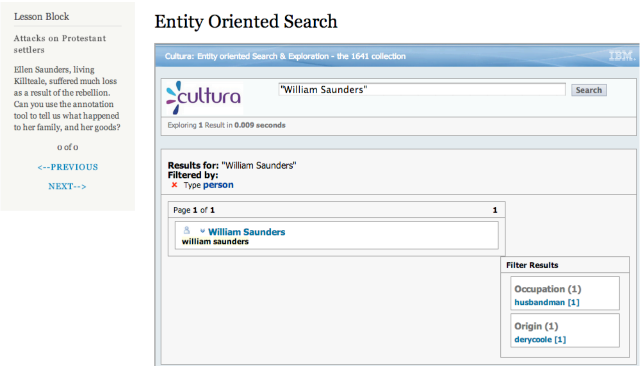
☑ Check box to search over normalised depositions

Page 1: Results 1 -20 of 44 (Number of results over unnormalised depositions are 43)

Deposition of Mary Squerrel -- [Visualise Entities in Deposition Wheel Vis | Octopus Vis]
fol. 380v [1362] Exr Mary squerrel of Athy, lately wife to Nic. squerrel of Athy blacksmith slain some O weeks since (after he had valiantly fought, and killed Capt Wesly of narrow moore. and 2 others rebels) by John Harris of Athy Butcher, and John fitzgerald of skerries gent of the company of Cap: Wesly that fell upon him. lost all, worth 4 li. names knows not robert by night, some 6 weeks since. bible burnt at Athy by young Georg Waker of Athy sig. [mark]

Figure 3. Exploration of search results.

The visualization features outlined below are also available to users of the 1641 Depositions, and, in a similar way, allow the collection to be explored on the basis of the links and relationships between people, places, and events. Individually, these tools offer a range of exciting new ways of looking at the Depositions. The lesson narratives add a further layer of richness, integrating these views within a coherent narrative structure.



Lesson Block

Attacks on Protestant settlers

Ellen Saunders, living Killteale, suffered much loss as a result of the rebellion. Can you use the annotation tool to tell us what happened to her family, and her goods?

0 of 0

<--PREVIOUS

NEXT-->

Entity Oriented Search

Cultura: Entity oriented Search & Exploration - the 1641 collection

"William Saunders" Search

Exploring 1 Result in 0.009 seconds

Results for: "William Saunders"
Filtered by:
× Type person

Page 1 of 1 1

William Saunders
william saunders

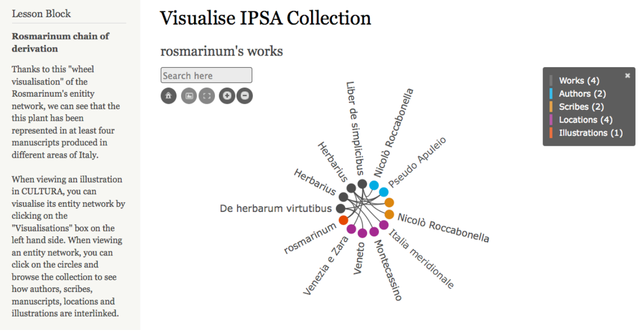
Filter Results

Occupation (1)
husbandman [1]

Origin (1)
derycoole [1]

Figure 4. Entity-orientated search.

For users of the IPSA collection, one of the lessons provided examines the development of botanical illustrations in Italy and allows the user to learn about the specific features of the software environment. To show the development of botanical illustrations over the centuries, the plant rosemary was selected. Rosemary has been illustrated in different illuminated manuscripts and allows the user to appreciate the chain of derivation in the illustration over time. Since it is not easy to show the user the different illustrations of the same plant that are present in different manuscripts on the same screen without the user having a disorientation problem (Conklin, 1987), a new type of visualisation has been designed. This "wheel visualisation" of the rosemary's entity network, is shown in Figure 5.



Lesson Block

Rosmarinum chain of derivation

Thanks to this "wheel visualisation" of the Rosmarinum's enitity network, we can see that the this plant has been represented in at least four manuscripts produced in different areas of Italy.

When viewing an illustration in CULTURA, you can visualise its entity network by clicking on the "Visualisations" box on the left hand side. When viewing an entity network, you can click on the circles and browse the collection to see how authors, scribes, manuscripts, locations and illustrations are interlinked.

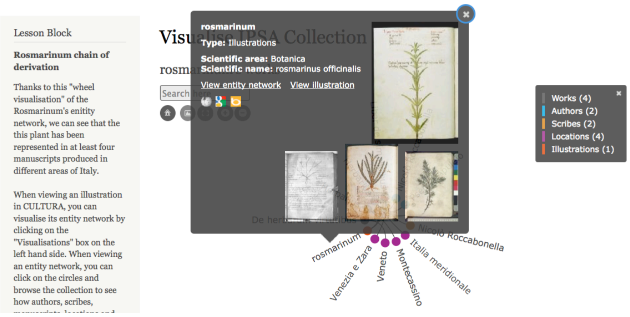
Visualise IPSA Collection

rosmarinum's works

Search here

Figure 5. "Wheel visualisation" of rosemary's entity network.

With this type of visualisation the user can appreciate that this plant has been represented in at least four manuscripts produced in different areas of Italy.



Lesson Block

Rosmarinum chain of derivation

Thanks to this "wheel visualisation" of the Rosmarinum's enitity network, we can see that the this plant has been represented in at least four manuscripts produced in different areas of Italy.

When viewing an illustration in CULTURA, you can visualise its entity network by clicking on the "Visualisations" box on the left hand side. When viewing an entity network, you can click on the circles and browse the collection to see how authors, scribes,

Visualise IPSA Collection

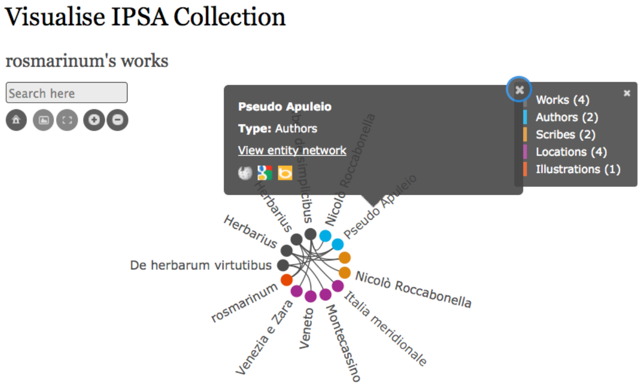
rosmarinum's works

rosmarinum
Type: Illustrations
Scientific area: Botanica
Scientific name: rosmarinus officinalis
View entity network View illustration

Figure 6. The four illustrations of the rosemary in the different manuscripts.

If the user continues to follow the sequence of the lesson, he can see a specific illustration of the plant, otherwise he can use the "dots" of the wheel to follow new routes. For example the user can ask to see the illustrations of the rosemary that are present in the four different manuscripts and the result is that of Figure 6.

If the user clicks on one of the authors, as for example, Pseudo Apuleio, the system displays a callout specifying the type of entity (see Figure 7) and allows the user to see the entity network of the author.



Visualise IPSA Collection

rosmarinum's works

Search here

Pseudo Apuleio
Type: Authors
View entity network

Figure 7. The screen showing that "Pseudo Apuleio" is an author and that is possible to see his network.

The entity network of the author is prepared by the system on the fly to answer this specific request of the user. In this way the user is adapting the interaction of the system to his interests, and he leaves the structure of the lesson and he uses the environment to adapt it to his interests.

The entity network of Pseudo Apuleio is very rich, since his herbal was widely copied in the Middle Ages (Figure 8).

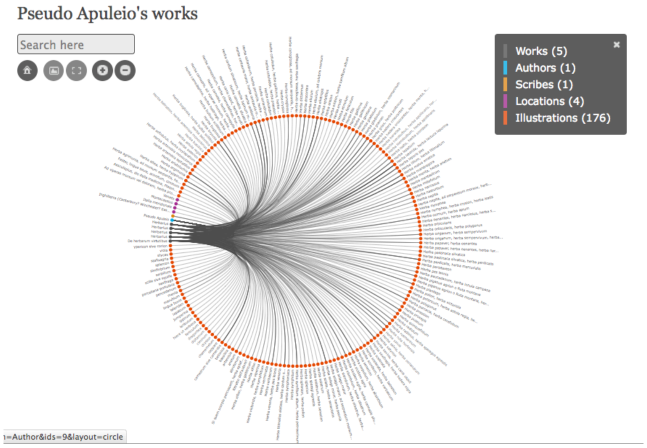


Figure 8. Pseudo Apuleio's works.

By clicking on each of the Pseudo Apuleio's works (a red dot for each work), the user is presented with a contextual square containing the essential information on the work together with an illustration if the dot refers to an illustration of a plant. From that square the user can navigate in the "wheel" of that specific plant and he can continue to explore into the digital cultural heritage collection.

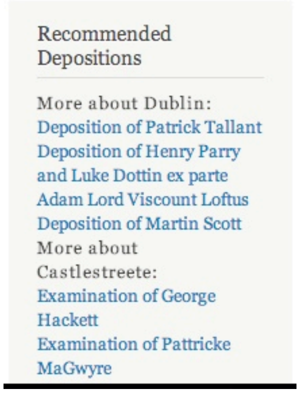
As these examples demonstrate, these narratives are a powerful tool. They allow the diverse affordances of the CULTURA system to be brought together in a way that greatly enhances the user's understanding of the collection. They guide users without limiting them, by providing relevant guidance when necessary, and allowing the user to follow additional sub-paths in order to delve deeper into a particular topic, or to leave the narrative path entirely in order to undertake independent exploration, secure in the knowledge that a single click will return them to the narrative. The content in both of these collections is notable for its complexity, and the tools that the CULTURA environment provides add further layers of richness. Such a wealth of material is challenging for users, and adaptive narratives provide important support, guidance, and direction, without the need to conceal or throttle any of the complexity and the opportunities for discovery what this complexity brings.

In addition, it is clear from the examples considered above that the ability to create dynamic pathways through a collection, that can respond to the needs and interests of the user is entirely independent from the specific type of digital cultural heritage collection that the system is managing. The system can handle non-pictorial material, or illustrations relevant to any possible subject of interest or medium – ranging from illustrations of an ancient manuscript to photographs of a modern book.

These adaptive narratives are a powerful and generalised tool of interaction and navigation with any sort of digital collection related to cultural heritage.

6.2 Recommendations

While the adaptive narratives provided by the CULTURA environment allow users to be flexibly supported as they interact with a wide range of material in a variety of ways, they are essentially pre-packaged ways of looking at the material in the collections. A more dynamic form of support is provided by the recommender element of the user model. This system tracks the artifacts that a user interacts with, and the level of that interaction. So, for instance, viewing an item gives it a low weight, but bookmarking a record or annotating material within it result in a higher rating. These weightings are used to present the user with items that are related to their previous use of the collection.



Recommended Depositions

More about Dublin:
Deposition of Patrick Tallant
Deposition of Henry Parry and Luke Dottin ex parte
Adam Lord Viscount Loftus
Deposition of Martin Scott
More about Castlestreete:
Examination of George Hackett
Examination of Pattricke MaGwyre

Figure 9. The CULTURA Recommender.

6.3 User Model

Not all users are equally accepting of adaptivity. Professional researchers, especially, do not seem to like the idea of an automated system controlling – however benignly, and with however good an intention – their access to the material on which they are carrying out research. For these users, transparency and scrutability are irreducibly essential. The CULTURA environment supports this requirement in a number of ways. Firstly, when the adaptive recommender system offers content recommendations to the user, it makes it clear why the material is being suggested. Secondly, the environment offers the user extensive control over their user model, allowing them to see what terms are influencing their recommendations, and to alter the relative weights given to each of these terms; so the user can control the "user model" as it is depicted in Figure 9 where it is shown that the user can interact with the weight has been associated to a word or a concept, modifying it in relation to specific requirements. In line with the results presented in (Ruthven *et al*, 2003) on the involvment of users manually weighting their interest, it is likely that only motivated users (e.g. researchers) will exploit the user model control. Yet it is part of our approach to provide the continuum of users with similar tools.

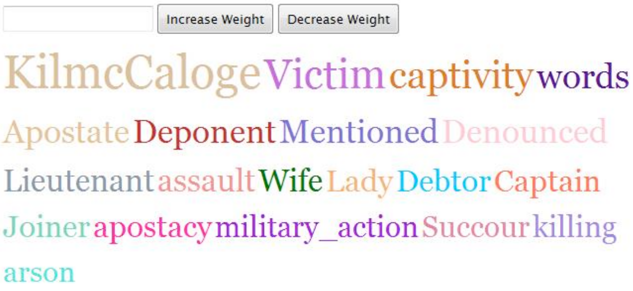


Figure 10. User model control.

7. CONCLUSIONS AND FUTURE WORK

From its inception, the CULTURA project set itself some ambitious goals. The adaptive research environment which it aspired to create went well beyond the existing state of the art. In moving beyond the specialised, search-based, and stereotyped norm, CULTURA offers a new model for access to and interaction with cultural heritage collections. It demonstrates the value of an adaptive interface, that responds dynamically to support the user, whatever his or her level of experience with the environment, or familiarity with the content. This flexibility has been both required and facilitated by CULTURA's attempt to address the needs not of a single stereotyped of user, but of a broad spectrum of user constituencies.

The design of this adaptive environment has been based firmly on input from users. CULTURA researchers have invested considerable effort in developing contacts with users from a broad range of user communities, and in working closely with those users to provide authentic and detailed user input plays its proper role in shaping CULTURA. This engagement has been both sustained and broad-based. This level of engagement is demanding, but essential in helping to underwrite the real-world usefulness and thus the long-term sustainability of the resultant environment.

User engagement as modeled by CULTURA is ongoing and iterative. Thus far, it has informed each phase of the project's development. The user input feedback loop will continue to play a vital role in the ongoing implementation and evaluation of the CULTURA system.

8. ACKNOWLEDGMENTS

The work reported has been partially supported by the CULTURA project, as part of the Seventh Framework Programme of the European Commission, Area "Digital Libraries and Digital Preservation" (ICT-2009.4.1), grant agreement no. 269973.

The authors would like to thank all the CULTURA partners for the useful discussions on many aspects related to the user engagement with cultural heritage collections.

The authors are most grateful to the referees for their very helpful comments.

9. REFERENCES

Agosti, M., Benfante, L., Manfioletti, M., Orio, N., and Ponchia, C. (2012). Issues to Be Addressed for Transforming a Digital Library Application for Experts into one for Final Users. In: Ioannides, M., Fritsch, D. , Leissner, J., David, R., Remondino, F., and Caffo, R. (Eds), *Euromed 2012, 4th Int. Conf., Progress in Cultural Heritage Preservation*, Short Papers. Multi-Science Publishing Co Ltd, Brentwood, UK, 2012, pp. 89-94.

Agosti, M., Manfioletti, M., Orio, N. and Ponchia, C. (2013). Evaluating the Deployment of a Collection of Images in the CULTURA Environment. In: *Proc. of the Int. Conf. on Theory and Practice of Digital Libraries (TPDL 2013)*, Valletta, Malta. LNCS Vol. 8092, Springer, Berlin Heidelberg (in press).

Azzopardi L. and Vinay V. (2008). Retrievability: an evaluation measure for higher order information access tasks. In: *Proc. of the 17th ACM Conf. on Information and Knowledge Management*, pp. 561-570.

Bailey, E., Sweetnam, M., O'Siochru, M., and Conlan, O. (2013). CULTURA: Supporting Professional Humanities Researchers. Accepted for publication at *Digital Humanities 2013*, University of Nebraska–Lincoln, 16-19 July 2013.

Carmel, D., Zwerdling, N., and Yogev, S. (2012). Entity Oriented Search and Exploration for Cultural Heritage Collections: the EU CULTURA project. *World Wide Web 2012 European project* track (Companion Volume), pp. 227-230.

Conklin, J. (1987). Hypertext: An Introduction and Survey. *IEEE Computer*, Vol. 20, n. 9, pp. 17-41.

Conlan, O., Staikopolous, A., Hampson, C., O'Keeffe, I., and Lawless, S. (2013). The Narrative Approach to Personalisation. In: *New Review of Hypermedia and Multimedia.*

Hampson, C., Agosti, M., Orio, N., Bailey, E., Lawless, S., Conlan, O., and Wade, V. (2012a). The CULTURA project: supporting next generation interaction with digital cultural heritage collections. In: *Proc. of the 4th Int. Euromed Conference, Limassol, Cyprus.* Springer, Heidelberg, pp. 668-675.

Hampson, C., Lawless, S., Bailey, E., Yogev, S., Zwerdling, N. and Carmel, D. (2012b). CULTURA: A metadata-rich environment to support the enhanced interrogation of cultural collections. In: *Proc. of the 6th Metadata and Semantics Research Conference, Cádiz, Spain* (pp. 227-238). Springer, Heidelberg.

Lawless, S., Hampson, C., Mitankin, P., and Gerdjikov, S. (2013). Normalisation in historical text collections. Accepted for publication at *Digital Humanities 2013*, University of Nebraska-Lincoln, 16-19 July 2013.

Ó Siochrú, M and Sweetnam, M, (2012) 'The 1641 Depositions and Portadown Bridge', *Seanchas Ard Macha*, 24:1, 72-103.

Orio, N. and Ponchia, C. (2013). IPSA – Imaginum Patavinae Scientiae Archivum. In: Rivista di Storia della Miniatura, 17, Firenze, Centro Di, (in press).

Ruthven I.,Lalmas M., and van Rijsbergen K. (2003). Incorporating user search behavior into relevance feedback. *Journal of the American Society for Information Science and Technology*, 54(6), pp. 529-549.

Sweetnam, M., Agosti, M., Orio, N., Ponchia, C., Steiner, C., Hillemann, E., O'Siochru, M., and Lawless, S. (2012). User Needs for Enhanced Engagement with Cultural Heritage Collections. In: Zaphiris, P., Buchanan, G., Rasmussen, E., and Loizides, F. (Eds), *2nd Int. Conf. on Theory and Practice of Digital Libraries (TPDL 2012),* Paphos, Cyprus. LNCS Vol. 7489, Springer, Berlin Heidelberg, pp. 64-75.

Sweetnam, M. (2013)a 'The Ministry in 1641', *Journal of Irish Scottish Studies*, forthcoming.

Sweetnam, M. (2013)b 'Caroline Preaching: Texts, Contexts, and Challenges', *Yearbook of English Studies,* forthcoming.

Tsakonas, G. and Papatheodorou, G. (2006). Analysing and evaluating usefulness and usability in electronic information services. *Journal of Information Science*, 32, pp. 400-419.

Personal Name Extraction from Ancient Japanese Texts

Mamoru Yoshimura
Graduate School of Information Science and Engineering, Ritsumeikan University, Japan
1-1-1 Noji-higashi, Kusatsu,
Shiga 525-8577, Japan
is046080@ed.ritsumei.ac.jp

Fuminori Kimura
Kinugasa Research Organization, Ritsumeikan University, Japan
56-1 Toji-in Kita-machi, Kita-ku,
Kyoto, Kyoto 603-8577, Japan
fkimura@is.ritsumei.ac.jp

Akira Maeda
College of Information Science and Engineering, Ritsumeikan University, Japan
1-1-1 Noji-higashi, Kusatsu,
Shiga 525-8577, Japan
amaeda@is.ritsumei.ac.jp

ABSTRACT

Text analysis of ancient Japanese language is difficult due to the lack of language tools to segment a sentence into words. There exists some morphological analysis tools for ancient Japanese in a specific period, but there are no such tools that can be used for general purpose. Even if morphological analysis tools were not available, it would be beneficial for a certain kind of text analysis to be able to extract named entities, such as personal names, from ancient Japanese texts. In this paper, we propose a method of personal name extraction from ancient Japanese texts based on Support Vector Machine (SVM) using features of character appearance and probabilistic word segmentation information. Experimental results showed that our proposed method were able to extract personal names from ancient Japanese texts with approximately 4% better F-measure when utilizing our proposed word segmentation information.

Categories and Subject Descriptors

H.2.8 [**Data mining**]

General Terms

Algorithms, Experimentation.

Keywords

support vector machine, named entity extraction, chunking, ancient Japanese texts, personal name

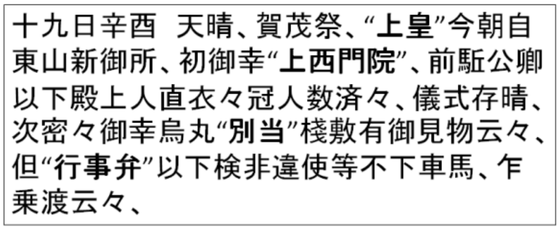
1. INTRODUCTION

Ancient writings are increasingly being digitized in text form. This leads to a possibility of applying natural language processing techniques to digitized ancient writings. Natural language processing techniques for modern Japanese relies on morphological analysis tools in order to separate words from sentences, to identify the part of speech of a word, and so on. The situation is the same for ancient Japanese. However, it is usually impractical to use a morphological analyzer designed for modern language to analyze text written in an ancient language. It is also difficult to segment sentences into words, because there are no dictionaries that can be used for morphological analysis except for Japanese in some specific periods.

However, the following things become possible, if it is able to extract personal names from ancient Japanese texts. One is to utilize it for the construction of ancient Japanese dictionaries. Another is to utilize it for text analysis and text mining of ancient Japanese writings.

In this paper, we propose a method of personal name extraction from ancient Japanese texts. This method uses Support Vector Machine (SVM) in order to learn the rules for named entity extraction automatically. We used the term “personal name” to describe a person name, an alias of a person, and the official position name for a person.

We use three Japanese ancient writings, namely “Hyohanki”, “Azumakagami” and “Gyokuyo” as the corpora. These ancient writings were written between late Heian era and early Kamakura era (12th-13th century). These writings are written in the style of Kanbun (a style of ancient Japanese which is based on classical Chinese). We conduct experiments of personal name extraction from these Japanese ancient writings in order to verify the effectiveness of the proposed method.



十九日辛酉　天晴、賀茂祭、“**上皇**”今朝自
東山新御所、初御幸“**上西門院**”、前駈公卿
以下殿上人直衣々冠人数済々、儀式存晴、
次密々御幸烏丸“**別当**”桟敷有御見物云々、
但“**行事弁**”以下検非違使等不下車馬、乍
乗渡云々、

Figure 1. Excerpt of “Hyohanki”.

2. RELATED WORK

2.1 Support Vector Machine

Support Vector Machines (SVM) is one of the supervised learning models with associated learning algorithms that analyze data and recognize patterns. SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. We used “LIBLINEAR” [1] for implementing the proposed method. “LIBLINEAR” is a machine learning library that is specialized for linear prediction.

2.2 Named Entity Extraction

In the methods of named entity extraction using SVM, it is popular to divide an input sentence into analysis units of proper sizes (tokens) and to group one or more analysis units into one token. The statuses of these grouped tokens are represented by a chunk tag set. “IOB2” chunk tag set is one of the best performed chunk tag set in previous studies. In “IOB2” chunk tag set, “B” tag is attached to the first token of a named entity, “I” tag is

Conference’10, Month 1–2, 2010, City, State, Country.

attached to the following tokens in the named entity after “B”, and “O” tag is attached to tokens that are not a part of named entities.

In this paper, we use these tags as personal name tags. By doing so, personal name extraction is regarded as the learning of rules to classify each token in the input sentence into one of personal name tags. Table 1 shows an example of personal name tag attachment for a part of a sentence in “Hyohanki”. The character string “上皇” is a personal name (in this case, a official position that means a retired emperor). The first character “上” is attached the “B” tag, the following character “皇” is attached the “I” tag.

Table 1. An example of personal name tagging.

character	IOB2 tag
上	B
皇	I
今	O
朝	O
自	O

2.3 Named Entity Extraction using Support Vector Machine

Yamada et al. [2] proposed the grouping of tokens using SVM. This research reported that SVM is effective for the grouping of tokens. However, personal names shorter than a morpheme will not be able to be extracted if the result of morphological analysis is used as the analysis unit.

Asahara et al. [3] proposed a method that adopts a character as the analysis unit in order to solve this problem. This method can perform named entity extraction even if there is a difference in word boundary between the result of morphological analysis and named entities. However, it is impossible to attach the part of speech information to each character directly. For this purpose, this method also uses “Start/End” (SE) chunk tag set, which represents the position in a word. In this method, “B” tag is attached to the first character of a word, “E” tag to the last character of it, and “I” tag to the middle of it.

Table 2 shows an example tag attachment by SE chunk tag set for a sample sentence. This sample sentence is a part of the text in “Hyohanki”. This sample sentence is divided into words as shown in figure 2 using our proposed word segmentation method, which is explained in the Section 3.

- 「上皇今朝自東山新御所、」

↓

- 上皇　今朝　自　東山　新　御所　、

Figure 2. Excerpt of the segmentation result.

Table 2. An example of tagging for positions in a word.

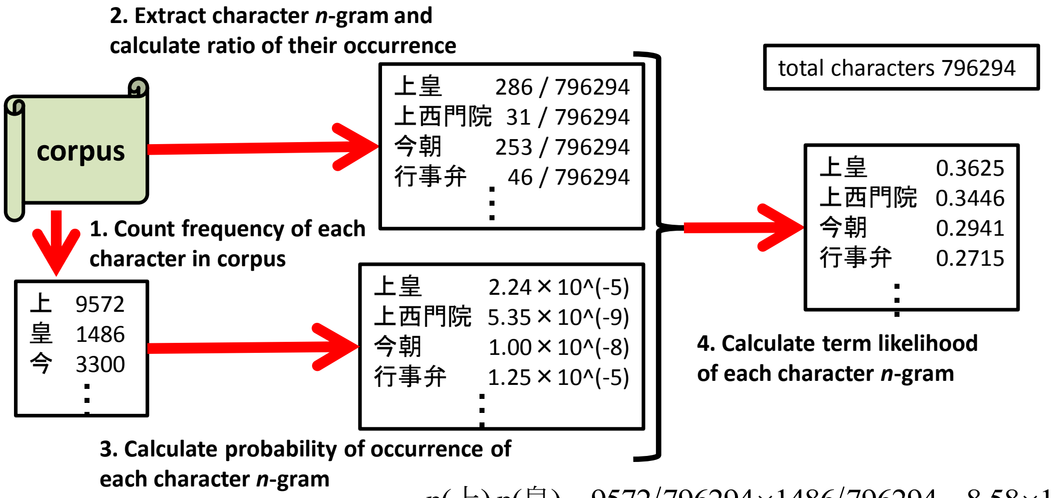
character	SE tag with segmentation result
上	B-上皇
皇	E-上皇
今	B-今朝
朝	E-今朝
自	S-自

3. APPLYING THE METHOD TO ANCIENT JAPANESE TEXTS

Named entity extraction from modern Japanese texts usually utilizes part of speech and morpheme information obtained from a morphological analyzer and inputs them into SVM to conduct leaning and estimation. It also uses information of scripts, i.e. hiragana, katakana, and kanji.

In our proposed method, we target ancient Japanese writings written in Kanbun (a style of ancient Japanese which is based on classical Chinese). We cannot utilize the results of morphological analysis for the proposed method, because there is no morphological analyzer for ancient Japanese in Kanbun style. Therefore, we cannot obtain sufficient information from these texts.

In order to solve this problem, we use the word segmentation method that from our previous research [4]. We calculate the likelihood of character n-grams to be a word, and extract character n-grams with higher likelihood as ancient Japanese words.



$$p(\text{上})p(\text{皇}) = 9572/796294 \times 1486/796294 = 8.58 \times 10\hat{}(-15)$$

Figure 3. Processing flow of term likelihood calculation.

In this method, we first calculate the likelihood of each character n-gram to be a word, and then select the n-grams with high likelihood. These extracted character n-grams can be considered as possible words. We call this likelihood “term likelihood”. It is assumed that a string that is a correct word will appear more frequently than the strings generated by randomly combining characters that constitute the word. Therefore, the n-grams with higher term likelihood can be considered to have higher possibilities to be correct words.

Figure 3 shows the processing flow of the term likelihood calculation. First, we count the frequency of each character in the corpus. Second, we extract character n-grams from the corpus and calculate the probabilities of their appearances in the corpus. Third, we calculate the joint probability of characters in each n-gram. Finally, we calculate the term likelihood of each n-gram.

This word segmentation method has not achieved sufficient precision of segmentation. Besides, this method has the problem for longer analysis unit. This word segmentation method can find correct word boundary in a high accuracy but tends to divide words into smaller character strings than a word. Therefor, we adopt the same approach as [3], which uses character as the analysis unit. The difference is that we use our proposed segmentation method instead of morphological analyzer. We consider the segmentation results of this method as words. These results are utilized for attaching “SE” tag to each character in the text.

4. PROPOSED METHOD

Our proposed method first learns the extraction rules of personal names from an annotated training corpus, and then extracts personal names from ancient Japanese texts by using SVM. We adopt a character as the analysis unit, and group them in order to extract personal names. The proposed method uses the IOB2 tag set as the personal name tag, which is reported to be effective in [2]. The proposed method classifies each character into one of three personal name tags in order to extract personal names. Besides, we also adopt the word segmentation method mentioned in section 3. The proposed method uses the results of this word segmentation as words, and attaches the SE tags according to these results.

The proposed method consists of the following three steps:

1) Word segmentation of input text (described in section 3)

2) Feature extraction for chunking by the method using the SE tag set (described in section 4.1)

3) Classification and grouping of tokens by the method using the IOB2 tag set (described in section 4.2)

4.1 Feature extraction for chunking

In this subsection, we explain the features used to learn the rules of personal name extraction.

Each character has three pieces of information; its own character, SE tag with segmentation result, and IOB2 tag. IOB2 tag is attached in the last step. In this step, the features of each character consists of information in two characters before and two characters after the character, and the character itself.

Table 3 shows an example of features. This example is the case for the focusing character “今” when the input text is “上皇今朝自”. The gray cells in table 3 are the information used for learning “今” in the next step.

Table 3. An example of features. This example is the case for a character “今”.

position	character	SE tag with segmentation result	IOB2 tag (attached in next step)
i-2	上	B-上皇	B
i-1	皇	E-上皇	I
i	今	B-今朝	O
i+1	朝	E-今朝	O
i+2	自	S-自	O

4.2 Classification and grouping of tokens by SVM

In this step, the proposed method conducts the classification and grouping of tokens by SVM. The features extracted in the previous step are entered to SVM and the method estimates the personal name tag for each character. The gray cells in table 3 are used for the features of the i-th character. However, the personal name tags at the i-2 and i-1 positions are known when learning but unknown when testing. Therefore, the personal name tag estimated at each position is used as the feature for the next character. Since the estimation is done one character by one character from the beginning of a sentence, we have estimated personal name tags for i-2 and i-1 positions when estimating the tag for i-th character. For example, when estimating the tag for the character “今” in the i-th position, the estimated tags “B” for “上” at the i-2 position, and “I” for “皇” at the i-1 position is used.

5. EXPERIMENTS

We conducted experiments of our proposed method of personal name extraction from ancient Japanese texts. For the experiments, we used three ancient Japanese writings, namely “Hyohanki”, “Azumakagami”, and “Gyokuyou”. For these documents, there are personal name indices that are manually compiled by scholars and are available in digital form. These indices were used as the correct answers for both training and test data. Table 4 shows the number of characters and the number of personal names contained in each document used in the experiments. For evaluation, we calculated precision, recall, and F-measure for each document by the 5-fold cross-validation.

In order to verify how the information of term segmentation influences the extraction accuracy, we conducted comparable experiments of using/not using the feature of term segmentation results shown in Table 3. The results of these experiments are shown in Tables 5 and 6.

Table 4. The number of characters and the number of personal names in each document used in the experiments.

	Number of characters	Number of personal names
“Hyohanki”	796,294	22,488
“Azumakagami”	787,250	39,909
“Gyokuyou”	1,934,754	22,823

Table 5. The experimental results of not using the feature of term segmentation.

	Precision	Recall	F-measure
"Hyohanki"	0.6149	0.5272	0.5676
"Azumakagami"	0.6829	0.6117	0.6454
"Gyokuyou"	0.6697	0.5973	0.6314

Table 6. The experimental results of using the feature of term segmentation.

	Precision	Recall	F-measure
"Hyohanki"	0.6699	0.5956	0.6086
"Azumakagami"	0.7151	0.6697	0.6917
"Gyokuyou"	0.6857	0.6507	0.6678

We also conducted an experiment for examining the variation of accuracy when varying the amount of text used for training and testing. We used "Hyohanki" for this experiment. We calculated F-measure for several different portions of "Hyohanki" text by the 5-fold cross-validation. The experimental conditions are the same as the previous experiments except for the amount of text used. The result of this experiment is shown in Figure 4.

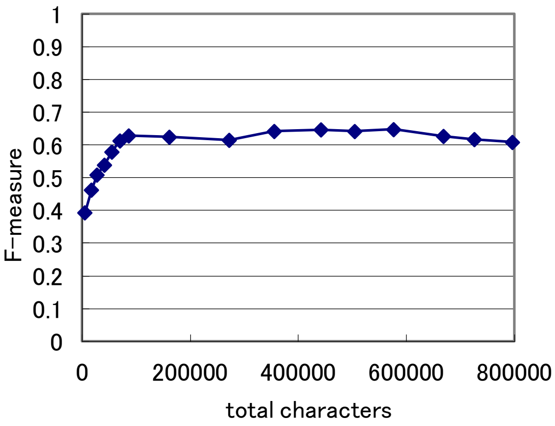


Figure 4. The relations of accuracy and the number of text used.

6. DISCUSSION

Table 5 and 6 show the experiment results for "Hyohanki", "Azumakagami" and "Gyokuyo". Table 5 is the case of not using the feature of term segmentation, and table 6 is the case of using it. These results indicate that the case of using the feature of term segmentation increased extraction accuracy than the case of not using the feature of term segmentation in each three ancient writings.

Figure 4 shows the relationship between total number of characters in corpus and accuracy of person name notation. The result of figure 4 shows that the F-measure increase in proportion to the number of character in range of 0 to about 90,000 characters, and it stabilizes in the range of over 90,000 characters. This result indicates that the proposed method with more 100,000 characters in corpus performs sufficiently. There is a possibility that sufficient amount of data in ancient writings is not prepared because they are limited for amount of data. Therefore, it is important to perform with fewer data amount in corpus.

7. CONCLUSION

In this paper, we proposed a method of named entity extraction from ancient Japanese digitized text based on Support Vector Machine using features of character appearance and probabilistic word segmentation information. We improved accuracy of person name extraction by using the result of our proposed word segmentation method as feature for SVM. We verified that the proposed method is effective for extracting personal names in ancient Japanese texts on which morphological analyzers are not available.

In our future work, we should improve the accuracy of the proposed method. We will focus on notations appearing before and after the personal name. Besides, we are planning to apply our proposed method to extract place names as well as personal names, which will make our method more useful for various kinds of text analysis of ancient Japanese writings.

8. ACKNOWLEDGMENTS

This work was supported in part by the MEXT-Supported Program for the Strategic Research Foundation at Private Universities "Sharing of Research Resources by Digitization and Utilization of Art and Cultural Materials" from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, MEXT Grant-in-Aid for Young Scientists (B) 23700302 "Information Extraction and Visualizing from Archaic Documents", and JSPS Grant-in-Aid for Scientific Research (C) 24500300 "Research on Integrated Information Retrieval from Multilingual Digital Archives" from Japan Society for the Promotion of Science (JSPS).

9. REFERENCES

[1] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* 9 (2008), pp. 1871-1874.

[2] Taku Kudo and Yuji Matsumoto. Chunking with Support Vector Machines. In *Proceedings of The Second Meeting of the North American Chapter of the Association for Computational Linguistics for Computational Linguistics on Language technologies* (NAACL2001) (2001), pp. 1-8.

[3] Masayuki Asahara and Yuji Matsumoto. Japanese Named Entity extraction with redundant morphological analysis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language* (NAACL2003) (2003), pp. 8-15.

[4] Mamoru Yoshimura, Fuminori Kimura, and Akira Maeda. Word Segmentation for Text in Japanese Ancient Writings Based on Probability of Character N-grams. In *Proceedings of the 14th International Conference on Asia-Pacific Digital Libraries* (ICADL2012) (Nov.2012), pp. 313-316.

Reconstruction of Apollo Mission Control Center Activity

Douglas W. Oard
College of Information Studies and UMIACS
University of Maryland, College Park, MD USA
oard@umd.edu

Abhijeet Sangwan, John H. L. Hansen
Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, Richardson, TX, USA.
{abhijeet.sangwan,john.hansen}@utdallas.edu

ABSTRACT

Some cultural heritage institutions have large and growing spoken word collections, the items of which often live in isolation from each other and from other parts of those collections. This paper describes the design process for construction and evaluation of a system to automatically construct links between spoken conversations that address different aspects of the same event. We see this as one step, among many, towards building richer interconnections between part of the same collection, and between collections. Our development environment is a collection of several thousand hours of recordings made in the Mission Control Center during the Apollo space missions of the 1960's and 1970's.

Categories and Subject Descriptors

H.3.7 [**Information Systems**]: Information Storage and Retrieval – digital libraries.

General Terms

Design, Experimentation.

Keywords

Cultural heritage, content linking.

1. INTRODUCTION

As physical access to archival collections becomes increasingly available, both as a result of digitization initiatives and as a result of substantial investments in making those digitized assets available on the Web, providing equally facile and capable "intellectual access" – the ability of users to find and use what they want -- has risen in importance. Intellectual access to archival collections raises a number of issues; in this paper we address one specific format issue: building links between different items in massive spoken word collections. At present, few collections are dominated by spoken word materials, but as prices drop, recording devices proliferate, and digital sustainability investments level the playing field between media, we can expect that to change [6]. We therefore believe the time is right to begin to explore these questions.

Many complex multi-party activities are coordinated using speech. Examples include air traffic control, military command centers, and human spaceflight. Common to all of these settings is that no single person can listen to everything that is happening, and thus no single person can actually come to know precisely how all that that happened was actually interconnected. Indeed, the quantities of recorded speech can become so vast that in some cases no single person could ever even listen to all of it, much less make sense of it all. As the cost of recording and storing audio continues to decline, scholars who seek to make sense of our past will therefore need new tools to help focus their attention on the small parts of this cornucopia that that they most need to hear, and on which parts of that they need to actually hear together. In this paper, we describe the design of a system for this exploration of the speech recorded in the National Aeronautics and Space Administration (NASA) Mission Control Center (MCC).

2. THE MCC AUDIO

A total of 38 people flew in Apollo spacecraft on 15 missions between 1968 and 1975. Of those, 24 flew to the moon on 9 missions; 12 of those people walked on the moon. Together, the flights spanned more days than a typical person works in a year, with 30-track audio recorders running continuously during that time in the MCC. The result is about 100,000 hours of recorded audio, with perhaps about 10% being speech and 90% silence.

The MCC was organized hierarchically, with one flight director, a dozen or so flight controllers, and a corresponding set of "back rooms" (more properly, "staff support rooms") that supported each flight controller. One "loop" (i.e., intercom circuit) connected the flight director with the flight controllers, and each back room had a separate loop to connect them with the flight controller who they supported. There were also several additional loops that two or more flight controllers could select when necessary to facilitate coordination activities that did not need to be heard by the full flight control team. Two special loops were also recorded, one between the spacecraft and the MCC, and a second for the news media that included those communications along with public affairs commentary. These circuits were recorded using a 30-track tape recorder that ran continuously; specific loops could be assigned to specific channels on the tape recorder, and it was common to record at least all of the circuits mentioned above.

During low workload periods, flight controllers would typically monitor three circuits simultaneously: (1) the flight director loop, (2) the loop to their own back room, and the (3) space-to-ground communications. They would typically alternate between talking on the first two of those; only the CAPCOM (a title derived from the older name "capsule communicator") would normally talk to the astronauts. We can, therefore, trace the flow of information from a specific flight controller's back room to that controller, then from that controller on the flight director loop to the CAPCOM, then (with the flight director's concurrence) from the CAPCOM up to the spacecraft. Indeed, during certain mission phases there were voice recorders running on board the spacecraft so we can hear how the astronauts discussed and acted on information that they received from the ground. The entire system included several dozen people, and during high workload periods there were many more simultaneous conversations going on that any one person could listen to. Flight controllers were able to monitor the back room loops of other controllers, and often did when specific activities on those loops might affect their own decisions.

All communication with the spacecraft was transcribed twice, once in near real time for use by the press and the second time (more carefully) for use in post-mission analysis. None of the other loops were routinely transcribed, however. Indeed, with the exception of the flight director loop, most of the other recorded loops have never even been replayed. Today, the tapes are stored in the National Archives and Records Administration (NARA), which has no system capable of playing them. There is such as system at the NASA Johnson Space Center, however, although this machine can currently play only 2 of the 30 tracks at a time. We are currently working with NASA to create or to otherwise gain access to a 30-track replay (and digitization) capability.

3. THE SPEECH LINKING TASK

Because the Apollo missions were carefully choreographed using a meticulously planned timeline, Mission Elapsed Time (MET) since launch provides a natural means for organizing access to the resulting flood of information. Time-based reconstruction has is widely used, include in aircraft accident investigations and for mission replay on military training ranges, and it is well matched to the linear nature of audio. We have, therefore, constructed a mission replay system for the Apollo missions that we call the Apollo Archive Explorer in which audio, video, transcripts, maps, planned and actual event timelines, and post-flight commentary are presented as a unified time-synchronized reconstruction of a mission [7]. Our focus in this paper is on the design of an additional capability that will add multi-channel audio scene reconstruction to the Apollo Archive Explorer.

The basic capability is simple; we can mix audio from multiple sources, some in one ear, some in the other, some in both; some at higher volume, some at lower. Such a capability replicates to a degree the (monophonic) capability that flight controllers had at the time to listen to multiple loops at once. Fight controllers were, however, highly trained to manage that complexity, and experienced in recognizing the voices on those loops. For modern users, we will need to provide some tools to help them manage and make sense of the complexity.

We envision three kinds of tools. First, we can use speech activity detection and speaker identification to identify who is speaking when and then to indicate that graphically in some way. Our initial design for this uses a sketch of the MCC console layout to simply indicate which flight controller is speaking (by lighting up the depiction of that console); in future work, we expect to build a similar visualization showing the back rooms. We do not yet have the back room loops digitized, so we are focusing initially on integrating the flight director loop, the space-to-ground communications, and the audio recorded aboard the spacecraft.

The speaker identification problem is simplified somewhat in this setting because flight controllers are typically the only people who speak on both the flight director's loop and (usually one) back room loop. There are some times when more than one flight controller is present at the same console, but there are also long periods in which only a single flight controller is present. We can therefore cluster speakers across one back room loop and the flight director loop, thus easily identifying which speaker is the flight controller who owns that specific back room loop. Once we know that, we can listen for first-name references to specific members of the back room staff, thus labeling most of the remaining clusters. Because we don't yet have the back room loops digitized, we are initially training a speaker identification system for the Apollo 11 flight director loop by hand-annotating a portion of the recording in a more conventional way. One early result from this work is that the shortness of some utterances (e.g., polling flight controllers for their agreement to proceed to the next mission phase) is challenging for conventional speaker identification techniques. We therefore plan to build in some interaction models that leverage specific forms of stylized interactions that often result in short utterances in this setting, and we plan also to leverage limited-vocabulary isolated-word speech recognition because it is common for an interaction to begin with the statement of a name or a position title that indicates who is being addressed.

Although the initial use we will make of speaker identification will be to indicate to the user of the Apollo Archive Explorer who is speaking when, our most important use of speaker identification will be as a basis for speaker-dependent Large-Vocabulary Continuous Speech Recognition (LVCSR). Our initial experiments with speaker-independent LVCSR (trained on other sources) yielded results that are not sufficiently accurate for content linking, so improving LVCSR accuracy is on our critical path. In addition to creating speaker-dependent models for each flight director and each flight controller (about 60 people total, because flight control teams worked in shifts), we are also now building domain-specific language models. The Apollo program is one of the most extensively documented undertakings in all of human history, so there is no shortage of text that can be used for language modeling. Much of this text was originally in printed form and is now available from the NASA Technical Reports Server (NTRS) as scanned PDF, for which Optical Character Recognition is easily performed. Of course, OCR introduces errors, so there will surely be a quality-quantity tradeoff to explore.

For our initial LVCSR experiments have focused on 11 hours of spacecraft communications with the MCC [8]. We trained a language model using text from a number of sources including transcripts, books, and technical reports. Although there was existing OCR for all of these materials, we obtained better results from rerunning OCR with a more modern system We used the resulting text to train a word trigram language model with a 38,000 word vocabulary. We ran forced alignment on the 11 hours of audio, finding that only 3 hours aligned well enough to be used; we rejected the remaining 8 hours due to (i) inaccurate transcripts, (ii) inaccurate timestamps in the transcripts, and (iii) poor quality audio. The remaining 3 hours of audio was split equally into adaptation and evaluation sets. Using a conversational telephone acoustic model (trained on a mixture of the switchboard and Fisher datasets) as baseline, the adaptation set was used to perform MLLR followed by MAP adaptation (resulting in adapted acoustic models). Upon decoding the evaluation dataset, we obtained a word error rate (WER) of 77% for the adapted system, which was a substantial improvement over the 92% word error rate of the baseline system without adaptation (i.e., trained only on Fisher and switchboard, but with our Apollo language model). We are now working on using the in-band transmitter keying tones ("quindar tones") to improve the time alignment and thus gain access to additional training data, and of course we can ultimately train on one entire mission and then test on subsequent missions. Moreover, much more material exists from which richer language models could be built. And, of course, the MCC loops that are our principal focus are far more tractable acoustically than the spacecraft communication circuits. We therefore foresee little difficulty in eventually sufficiently accurate transcripts to support content alignment.

Other issues that may affect the accuracy of speaker identification and speech recognition include unmodeled variations such as (i) background noise (such as side conversations in MCC or pumps running aboard the spacecraft), (ii) band limited recording equipment (particularly for recordings that were later transmitted from the spacecraft to the ground and recorded there), and (iii) time-varying channel characteristics (which can result both from the characteristics of the analog take recorders used at the time and from the need to replay these tapes on much older equipment today) [1,9]. Additionally, despite the image of "right stuff" astronauts and flight controllers when they are speaking on the radio, the "off the record" recordings of the mission control loops and the interaction among the astronauts exhibit clear variations in speech production due to the whole range of human emotions such as stress, anxiety, and joy. Physical, emotional and cognitive state are well known to influence speech production, and the resulting variations can adversely affect both speech recognition and speaker identification [2]. Moreover, the Apollo astronauts spoke in an exceptionally diverse range of physical environments, including under extreme g-forces during launch and reentry, in low pressure pure oxygen during moonwalks, and (later in life) in oral history interviews. This exceptional range of diversity in working environments in itself offers some remarkable research opportunities for speech processing systems. Indeed, those unique opportunities were one of our principal initial motivations for undertaking this project.

There is a large body of research that has focused on problem of "speech under stress." The usual approaches are to attempt to remove variability in speech (introduced due to environment, channel and speech production) in either the feature domain or the model domain [3], and we should be able to apply some of those techniques as well. Unlike the speech corpora on which much of this earlier research has been performed, we have very large amounts of speech from a relative small number of people. That offers us an unprecedented opportunity to investigate long-term adaptation techniques that could ultimately have broad applications beyond this specific task (for example, personal speech systems such as Siri face similar challenges).

Once we have adequately accurate LVCSR (for which prior work in a query-based ranked retrieval setting suggests requires will require word error rates below about 50%), we can begin to build content linking systems. We already have some experience with content linking in this setting from an experiment we reported in the main conference in which we linked mission events from the transcript of the communication between the spacecraft and the MCC to question-answer pairs from the oral history interviews with the same astronauts that were recorded many years later [5]. In that work, simple sliding window bag-of-words techniques yielded a mean reciprocal rank at 3 of about 0.5 for mission events for which a substantial mention existed in the oral history. Importantly, however, we have not yet tackled the important problem of automatically determining when no link should be made. Such a capability will be essential for content linking between the MCC loops. For this we will need to switch from a ranked retrieval design to one based on supervised machine learning for text classification, and for that we will need training data. Thus system design naturally leads us to the question of test collection deign.

4. TEST COLLECTION DESIGN

Our goal is to discover when two loops should be linked, so our test collection must contain some ground truth for those kinds of links. The synchronized nature of our task greatly simplifies the search space – we seek only to link two loops at the same time, not to build links that span different mission phases or that span different missions. This constraint results in a simple form for a link, it is specified by a start time, an end time, and a pair of loops to be linked during that interval.

In our initial thinking, we can see two employment scenarios that lead to two link types. In one scenario, the listener hears something being discussed on the flight director loop and wishes to hear the conversations between the flight controller involved and his (flight controllers were all men in Apollo) back room. In the other, more challenging, scenario, the listener hears something on one back room loop and wishes to know if there are related conversations ongoing on other back room loops. We plan to identify some ground truth links of each type.

There are at least three ways of identifying such events in a mission. First, NASA prepared a post-flight mission report in which every engineering anomaly that occurred during the mission was identified. For example, there were water leaks in one of the two Apollo spacecraft during both the Apollo 11 and the Apollo 15 missions. The resulting database (present in multiple scanned documents, not actually yet as a database) offers one possible source for events that would have prompted discussion on one or more loops. Second, over the past several decades, authors and documentary film makers have mined the records of the Apollo program for compelling human interest events. For example, the commander of the first lunar mission (Apollo 8) became ill between the Earth and the Moon. These events, less well codified but now nonetheless well known, offer an alternative source of events that could have prompted discussion on multiple loops. A third obvious source for events is the sequence of planned mission events from the pre-flight flight plan. For example, a planned television broadcast from lunar orbit would require coordination among flight controllers responsible for spacecraft attitude and communication systems. From these three sources, it seems reasonable to conclude that identifying a broad range of events for which links might be built should be straightforward.

The more interesting question will be whether people can agree on the proper span for a link. Here we are helped a bit by the fact that the onset time of a link might reasonably be of greater importance to the listener than its termination time, for the simple reason that once the listener chooses to listen to something they can make the decision of when to stop listening on their own. We will, therefore, ask annotators to mark both onset and termination times, but we will initially evaluate (and assess inter-annotator agreement) based solely on the onset time errors. As we have done previously for retrieval of unsegmented speech, we plan to initially use a one-sided linear penalty function as our evaluation measure [4].

Ultimately, of course, we will need to conduct user studies to learn which kinds of links users are most interested in seeing, and what kinds of errors actually prove to be most troublesome for them. But this is a chicken-and-egg sort of problem, in that we cannot study how users would use a system until such a system exists. So we necessarily anticipate a spiral development model in which we first build a plausible system, and then we iteratively refine that system as we learn more about how it will be used.

5. CONCLUSION

We often think of cultural heritage as involving things that are centuries old, and often at best incompletely documented. As time progresses, however, we will surely encounter more collections like that created by the Apollo missions, where our problems will be not how best to make the most of that which is scarce but rather how best to make the best use of that which is abundant. The Apollo missions, flown as they we now nearly half a century ago, offer an outstanding laboratory with which to begin that quest.

6. ACKNOWLEDGEMENTS

The author is grateful to the rest of "Houston We Have a Solution" project team for the rich discussions and initial implementation and experimental effort that have brought us to this point, including Lakshmish Kaushik, Chengzhu Yu, Joseph Malionek and Bryan Toth. This material is based upon work supported by NSF under Grant 1218159. Opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. REFERENCES

[1] Murat Akbacak and John H.L. Hansen. Environmental sniffing: noise knowledge estimation for robust speech systems. *IEEE Transactions on Audio, Speech, and Language Processing,* 15(2) 465-477, 2007.

[2] John H. L. Hansen, Abhijeet Sangwan and Wooil Kim. Speech under stress and Lombard effect: Impact and solutions for forensic speaker recognition. *Forensic Speaker Recognition*, Springer, New York, 2012, pp.103-123.

[3] Wooil Kim and John H.L. Hansen. Time–Frequency Correlation-Based Missing-Feature Reconstruction for Robust Speech Recognition in Band-Restricted Conditions. IEEE Transactions on *Audio, Speech, and Language Processing,* 17(7)1292-1304, 2009.

[4] Baolong Liu and Douglas W. Oard, One-Sided Measured for Evaluating Ranked Retrieval Effectiveness with Conversational Speech, in SIGIR 2006, pp.673-674.

[5] Joseph Malionek, Douglas W. Oard, Abhijeet Sangwan and John H.L. Hansen, Linking Transcribed Conversational Speech, in SIGIR 2013, 4 pages.

[6] Douglas W. Oard, Unlocking the Potential of the Spoken Word, *Science*, 321(5897)1787-1788, 2008.

[7] Douglas W. Oard and Joseph Malionek, the Apollo Archive Explorer, in JCDL 2013, 2 pages.

[8] Abhijeet Sangwan, Lakshmish Kaushik, Chengzhu Yu, John H.L. Hansen and Douglas W. Oard, Houston we Have a Solution: Using NASA Apollo Program to Advance Speech and Language Processing Technology, in Interspeech 2013, 5 pages.

[9] Umit Yapanel and John H.L. Hansen. A New Perceptually Motivated MVDR-Based Acoustic Front-End (PMVDR) for Robust Automatic Speech Recognition. *Speech Communication,* 50(2)142-152, 2008.

Personalising the Cultural Heritage Experience with CULTURA

Gary Munnelly
Knowledge and Data Engineering Group
Trinity College
Dublin, Ireland
munnelg@scss.tcd.ie

Cormac Hampson
Knowledge and Data Engineering Group
Trinity College
Dublin, Ireland
hampsonc@cs.tcd.ie

Owen Conlan
Knowledge and Data Engineering Group
Trinity College
Dublin, Ireland
Owen.Conlan@scss.tcd.ie

Eoin Bailey
Knowledge and Data Engineering Group
Trinity College
Dublin, Ireland
baileyeo@scss.tcd.ie

Seamus Lawless
Knowledge and Data Engineering Group
Trinity College
Dublin, Ireland
Seamus.Lawless@scss.tcd.ie

ABSTRACT

This paper discusses the CULTURA project and its attempts to personalise a user's experience of exploring cultural heritage collections regardless of their level of expertise. The services provided within the environment are introduced with respect to the four phase personalisation model used by CULTURA. In addition to these services CULTURA's comprehensive user model is detailed, in order to explain how a user's actions can affect the environment in which they work.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – Information filtering, query formulation, relevance feedback, retrieval models, search process, selection process. K.3.1 [**Computers and Education**]: Computer Uses in Education – Collaborative learning, computer-assisted instruction (CAI), computer-managed instruction (CMI).

General Terms

Algorithms, Management, Measurement, Design, Reliability, Experimentation, Human Factors.

Keywords

Personalisation, adaptive environments, cultural heritage, digital collections, user modelling, assisted learning.

1. INTRODUCTION

Cultural heritage provides a valuable resource through which future generations can be informed and learn about the origins of the modern world. In spite of being of importance, accessing the information which is inherent in artefacts is often a difficult task, particularly for the uninformed or inexperienced individual.

For a novice, searching through large quantities of information and interpreting data present a challenge. The tasks require that the researcher have some degree of knowledge regarding what they are looking for and how it will be represented in the corpus. If there are multiple differing references to an entity (e.g. 17th century Irish depositions alternatively refer to Sir Phelim O'Neill as "The O'Neill" and "The Rebel Leader"), the net effect can be that the user does not gain the full benefit of the content. The overwhelming size of the archives may even discourage the novice from researching the topic in any form.

For the expert, the challenge lies in discovering how the information is stored and what tools are available to traverse it. In general, this person knows what they are looking for and have a strong sense of how it will be documented. However they are hindered by their unfamiliarity with the services which allow them to explore the data and how these will respond to their queries.

Fundamentally, a solution to these problems should provide for the naive nature of the novice without diminishing the experience of the expert. This paper discusses the CULTURA environment, a subset of the tools it provides and how they are helping in this endeavour. It should be noted that, for the purposes of illustration, personalisation is discussed in relation to two extreme classes of user – the novice and the expert – but CULTURA is intended to cater to all levels of experience and adapt to reflect the research interests of each individual user.

Section 2 introduces the four phase model which CULTURA employs for personalisation and describes some of the tools which aid in its implementation. Section 3 describes the user model which represents the user within the system and upon which much of the personalisation is based. Section 4 discusses how these tools are being evaluated to ensure that they are providing the expected benefits for the users.

2. Personalisation in CULTURA

CULTURA is a three year, FP7 funded project [1]. Its main objective is to pioneer the development of personalised information retrieval and presentation, contextual adaptivity and social analysis in a digital humanities context. For the purposes of personalisation, CULTURA utilises a four phase model as illustrated in Figure 1.

During the guide phase of the model, the individual travels along a sequenced path through the content of the corpus. This path can be explicitly extended with new resources by the user, or implicitly by the system which can analyse a user's level of interest. Thus the system leads the visitor through the lesson while encouraging and advising them on relevant places to explore. Ultimately it is envisioned that users will naturally transition from the guide phase to the explore phase as they become more engaged with the content.

The explore phase is where CULTRA encourages users to spend the majority of their time. Here, the user is performing personal, free-form investigations into the contents of the corpus. The environment furnishes them with the tools they require to engage and access information, but ultimately they are driven by their own desire to learn.

The reflect phase affords a user the opportunity to examine their own user model and influence how it is being constructed. Changes made to the user model directly influence the behaviour of CULTURA and how it responds to a user. Here, the system is actively engaging with the researcher, giving them the ability to correct its interpretation of their actions and adjust itself as required.

In the suggest phase, the environment considers what it thinks the researcher is interested in and offers proposals for additional content which may be relevant. The information contained in the user model is collated and used to make conjectures regarding what the individual is trying to achieve and what may be of relevance to their current interests. These recommendations are presented to the user through various means within the environment

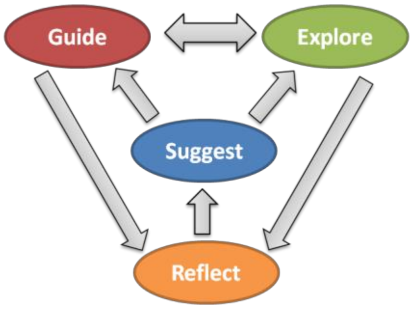


Figure 1: CULTURA's four phase personalisation model

The application of this model is facilitated by a number of services that have been incorporated into the CULTURA environment.

2.1 User Narrative

In Section 1 it was noted that both novice and expert researchers suffer from a lack of guidance when using a new system. The narrative module of the CULTURA environment is designed to provide this guidance by furnishing users with a selection of lessons which they may follow to increase their knowledge about the site and its contents [2]. As such, this module is associated with the guide phase of the personalisation model.

An individual narrative is comprised of a number of parameterised links between the various resources and services in the CULTURA environment. These links form a sequenced path which the user can traverse at their own leisure. At each stage of the narrative, the user is presented with an artefact (or a form of visualisation of an artefact) from the corpus and a body of text which describes that stage of the lesson. This text may contain a description of the artefact, instructions for the user to follow or some other form of information.

Due to the open nature of a lesson's textual content, a narrative may additionally be used to construct a tutorial on the functionality of the site, making users aware of the various features which are available to them. This type of narrative is more likely to be of interest to the expert user than narratives which describe the artefacts themselves.

While a user is following a narrative they retain the use of other site functionality such as annotation and bookmarking of content. This information is gathered by the user model in an attempt to identify topics within the artefacts which interest them. While following the narrative, the recommender block (see Section 2.2) will continue to be updated with suggestions based on this data. Thus, even as a researcher follows a lesson plan, CULTURA encourages them to explore under their own volition.

Should a user choose to deviate from the path of a narrative (perhaps to follow an interesting subject returned by the recommender) the module allows for this. The individual's progress through a particular narrative is recorded and stored until they choose to return to the path. A button is available on each subsequent page the user visits which will return them to the narrative when they are ready to continue.

In addition to the linear path through a lesson, a narrative can also have a degree of difficulty which alters the type of content that comprises a user's path. This is to allow for the fact that certain users may not wish to cover parts of the lesson either because it does not interest them or because it does not present them with anything new. For example, a novice researcher who is studying Ireland in the 17th century may not be aware of the state of the country at that time. They will therefore potentially require a more detailed lesson in order to compensate for this lack of knowledge. An adept researcher who is familiar with 17th century Ireland, but would like to learn more does not necessarily require as much information as the novice and so may be able to take a briefer route along a similar lesson plan.

The narrative structures are implemented and stored as XML files on the server machine. The lessons themselves are designed by designated experts. The contents of the XML file dictate which resources are to appear in the lesson, how they are to be ordered and how they are to be presented (text, entity visualisation, etc).

2.2 Personalised Search

A good search engine will return a ranked list of results with artefacts being ordered according to their relevance to the query. This approach is known to be extremely effective, but it does not attempt to actively help the user achieve their goals.

When an individual performs a single search across the corpus, they are often looking for something specific. However, if they are conducting multiple searches, then it is likely that they are attempting to research a more general topic. The CULTURA personalised search is attempting to identify when a user is researching a topic, what it is they are trying to research and thus return results which are more relevant to the user's interests.

As a user executes queries, the personalised search module attempts to identify terms which are common to each of the requests. If a term is appearing with a high degree of frequency, then this is indicative that the user is trying to research a broad topic to which this entity is relevant. When the module identifies a theme that is common to the searches, then in addition to performing the user's original query, the module can silently generate new searches by appending these relevant terms to the initial query string. The results of the multiple searches are then interwoven to produce a single ranked list which is presented to the user.

For example, consider a scenario where a user conducts a search for the Irish rebel Phelim O'Neill across a collection of depositions. The user model indicates that the researcher has been gathering information related to Louth prior to this search. Based on this, it would appear that the person's interests are more deeply rooted in County Louth than anywhere else in the country. The personalised search will thus return a list of all depositions which reference Pheilim O'Neill, but it will also silently conduct a search for depositions which reference both Pheilim O'Neill and County Louth. The results of this silent search will be assessed and interwoven with those of the initial query before the results are shown to the user.

In this manner the personalised search module attempts to actively and dynamically assist the user in their search for information rather than mechanically responding to static queries.

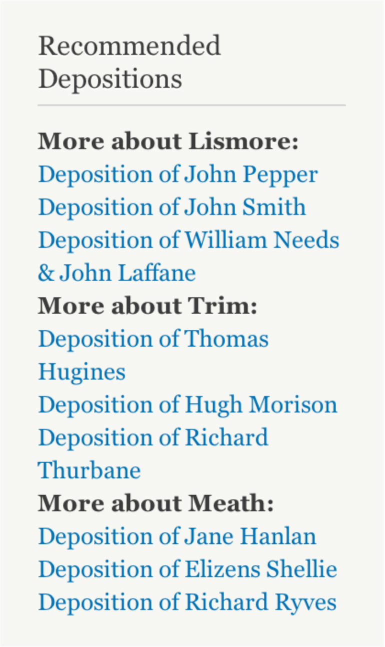
2.3 Recommender Module

As mentioned previously, CULTURA attempts to encourage users to actively explore and investigate the data that it curates. The recommender module is one of the primary means by which this is achieved. This service presents the user with lists of various entities which appear to be relevant to the topics which they are exhibiting an interest in. This entices the user into following new threads of information through the repository. In addition to encouraging active explorers, it also has a role in easing researchers from the guide phase to the explore phase by recommending deviations from the lesson plan.

In a similar manner to the personalised search, the recommender module attempts to identify entities and themes in which the user is expressing an interest. It can derive these terms from the user model which is constructed as the user interfaces with the system. The recommender then attempts to locate entities which are related to these identified terms. This is achieved through the use entity relationship extraction which is performed in CULTURA using IBM's LanguageWare [3].

As all the entities which exist in the corpus are known, CULTURA attempts to identify and link entities which are related. The recommender module uses these links to locate new artefacts which may be of interest to the user. The recommender module's suggestions are presented as a list in a statically located block alongside the content (Figure 2). In addition to presenting the user with the recommender module's suggestions, the recommender block also indicates why a particular entity is being suggested to them. This is particularly important due to the means by which the user interfaces with their user model (see Section 3).

As an example of how the recommender module may be of assistance, consider a user who has been reading many documents which describe the town of Trim in the 17th century. The recommender notices that Trim is a common term in many of the artefacts that the user is viewing. It examines the network of entities which are linked to Trim and finds that Hugh Morison gave a deposition which contained information about this town. It therefore suggests that the user read the deposition of Hugh Morison in order to get more information about this place of interest (see Figure 2).



Recommended Depositions

More about Lismore:
Deposition of John Pepper
Deposition of John Smith
Deposition of William Needs & John Laffane

More about Trim:
Deposition of Thomas Hugines
Deposition of Hugh Morison
Deposition of Richard Thurbane

More about Meath:
Deposition of Jane Hanlan
Deposition of Elizens Shellie
Deposition of Richard Ryves

Figure 2: The recommender block showing suggestions

3. User Model

Much of the functionality described in Section 2 is facilitated by CULTURA's comprehensive user model. The user model reflects the system's interpretation of the researcher and their goals.

The model is constructed dynamically as the user interacts with the environment and maintained for the lifetime of the user's account. Information such as pages the user visits, content annotated, artefacts bookmarked, searches performed, visualisations viewed and commonly observed entities are logged. Out of this data the user model extracts the common entities which were present in the user's searches. Recurring instances of a particular place, person etc. are considered indicative of an interest in that particular artefact. The more often an entity is observed, the greater the user's curiosity is believed to be. From this, conjectures can be made with regards to the user's long term interests, short term interests and level of expertise with respect to the site's content.

Due to the fallible nature of computers, it is possible that the model which the system generates will not be an accurate representation of the user. Facilities are therefore provided which allow an individual to tweak their model so that it is more in line with their actual intentions. This forms the basis of the reflect phase.

At present, modification is performed by means of an interactive tag cloud (Figure 3) where the model's interpretation of a researcher's interests is presented as a list of weighted terms within the cloud. Users can select entities which interest them and increase their weights so that they have a larger bearing on the recommendations that they receive. Alternatively, the user can reduce the weight of a term so that it has a negligible effect on their experience.

The user model is an invaluable resource for the provision of a unique, personal experience for the individual researcher. If well-structured, the model allows the system to offer meaningful information during the suggest phase. Furthermore, by allowing the user to modify their representation as they require, accurate personalisation can be assured.

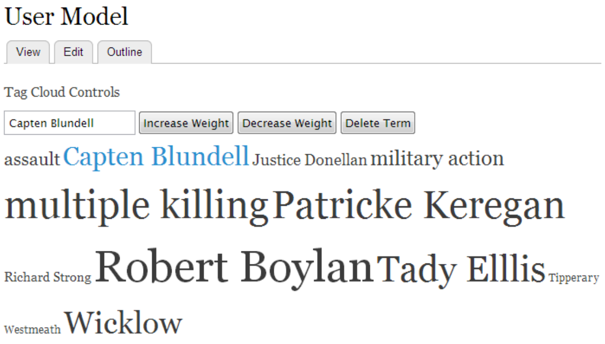


Figure 3: Tag cloud for interacting with the user model

4. Deployment

At present the personalisation services described in this document are being trialled as part of the CULTURA environment. In its current form, CULTURA aims to provide adaptive and personalized access to two historical collections – the 1641 depositions [4] and IPSA collection [5].

The 1641 depositions are a collection of hand written manuscripts, mostly legal in nature, which provide a unique insight into the cultural and political state of Ireland in the 17th century. The content of these documents is entirely textual. However, due to the non-standardised state of the English language in that century, the data is noisy with many inconsistencies in spelling and referencing of entities. The collection challenges the personalisation tools due to the inconsistent nature of the language it contains.

The IPSA (Imaginum Patavinae Scientiae Archivum) manuscripts are a collection of illustrated documents from Italy which describe the various properties of herbs and plants dating from as far back as the 14th century. Each record is comprised of hand drawn images of various floras, but contains little or no text. For investigators, the interest in a particular artefact is influenced by the content of the image, rather than textual data. Thus these manuscripts present a different challenge for the personalisation tools to the 1641 depositions.

The personalisation services have been trialled on several occasions during the development of the system and continue to be adapted to reflect the feedback from the users [6]. Volunteers who tested the system ranged widely in levels of expertise from the complete novice to relative expert. Among the groups involved were a group of secondary school students from Lancaster and a number of researchers from Padua, Italy.

As development of the tools continues, more user trials will be conducted to determine the usefulness and accuracy of the services.

The results of trials are being gathered, managed and examined using CULTURA's own evaluation engine – Equalia [7].

5. Conclusion

The personalisation tools which have been developed as part of the CULTURA project show great promise for improving user interaction with cultural collections. Although many of the tools are still in development, early user trials indicate that researchers of all levels of expertise are deriving benefits from their provision.

For the remainder of the project, these tools and services will continue to be enhanced in order to maximise the advantages experienced by the users. As mentioned in Section 1, ultimately the goal is to provide a system which is both accessible and equally useful to researchers at all levels of expertise.

6. ACKNOWLEDGMENTS

The work reported has been funded by the Seventh Framework Programme of the European Commission, Area "Digital Libraries and Digital Preservation" (ICT-2009.4.1)

7. REFERENCES

[1] The CULTURA Project: Supporting Next Generation Interaction with Digital Cultural Heritage Collections. C. Hampson, M. Agosti, N. Orio, E. Bailey, S. Lawless, O. Conlan, V. Wade Ding, W. and Marchionini, G. 1997. *A Study on Video Browsing Strategies*. Technical Report. University of Maryland at College Park.

[2] The Narrative Approach to Personalisation, Owen Conlan, Athanasios Staikopoulos, Cormac Hampson, Seamus Lawless & Ian O'Keeffe, New Review of Hypermedia and Multimedia [In Press]

[3] LanguageWare, IBM. Retrieved 06 2013, from http://www-01.ibm.com/software/globalization/topics/languageware/

[4] 1641 Depositions, Trinity College Dublin. Retrieved 06 2013, from http://1641.tcd.ie/index.php

[5] IPSA – Imaginum Patavinae Scientiae Archivum, Universita Degli Studi Di Padova. Retrieved 06 2013, from http://www.ipsa-project.org/

[6] "CULTURA: Supporting Enhanced Exploration of Cultural Archives through Personalisation", Bailey, E., Lawless, S., O'Connor, A., Sweetnam, S., Conlan, O., Hampson, C. and Wade, V. In the Proceedings of the 2nd International Conference on Humanities, Society and Culture, ICHSC 2012, Hong Kong, China, 2012.

[7] An Evaluation System for Digital Libraries, Alexander Nussbaumer, Eva-Catherine Hillemann, Christina M. Steiner, Dietrich Albert. TPDL 2012: 414-419

The CULTURA Evaluation Model: An Approach Responding to the Evaluation Needs of an Innovative Research Environment

Christina M. Steiner,
Eva-C. Hillemann,
Alexander Nussbaumer,
Dietrich Albert
Knowledge Technologies Institute,
Graz University of Technology
Inffeldgasse 13/V, 8010 Graz, Austria
{christina.steiner, eva.hillemann,
alexander.nussbaumer,
dietrich.albert}@tugraz.at

Mark S. Sweetnam
Department of History, Trinity College
Dublin
Arts Building
Dublin 2, Ireland
sweetnam@tcd.ie

Cormac Hampson,
Owen Conlan
KDEG, School of Computer Science
and Statistics, Trinity College Dublin
O'Reilly Institute
Dublin 2, Ireland
{cormac.hampson,
owen.conlan}@cs.tcd.ie

ABSTRACT

This paper presents the evaluation approach taken for an innovative research environment for digital cultural heritage collections in the CULTURA project. The integration of novel services of information retrieval to support exploration and (re)search of digital artefacts in this research environment, as well as the intended corpus agnosticism and diversity of target users posed additional challenges to evaluation. Starting from a methodology for evaluating digital libraries an evaluation model was established that captures the qualities specific to the objectives of the CULTURA environment, and that builds a common ground for empirical evaluations. A case study illustrates how the model was translated into a concrete evaluation procedure. The obtained outcomes indicate a positive user perception of the CULTURA environment and provide valuable information for further development.

Keywords

Cultural heritage, research environment, evaluation model, evaluation qualities, empirical study.

1. INTRODUCTION

Information retrieval technologies open up a range of possibilities for providing support in exploring, searching, and researching cultural heritage artefacts. Examples are automatic indexing and searching methods [9], normalisation of spelling variations in historical documents [10], extraction of entities representing persons, locations, events etc. from documents [3], the processing and mapping of dates and time intervals [8], or the application of social and influencer network analysis to digital collections. These new technologies and their integration and application in cultural heritage research environments and electronic information services require appropriate evaluation methodologies and thus pose specific requirements and challenges to evaluators. The present paper describes a comprehensive evaluation model accounting for that, which has been developed in the context of the CULTURA project[1]. The project aims at delivering a corpus agnostic research environment integrating a range of innovative services that guide, assist, and empower users' interaction with cultural heritage artefacts and take into account the diverse needs of different user groups. The novel methodological and technical approaches integrated, as well as the intended reusability of the technology with different collections and the diversity of users addressed as target audience challenges evaluation with respect to the use of sound and suitable methods across contrasting digital collections and diverse communities and users. Through the development of an evaluation model serving as a common ground for different evaluation studies, an appropriate level of comparability and generalizability of evaluation results can be maintained. The evaluation model is presented and a case study conducted with historians is outlined to illustrate how the investigation of the different evaluation axes and qualities of the model provide service-specific insights on the quality of the CULTURA environment and meaningful information for further development.

2. THE CULTURA PROJECT & SYSTEM

The interdisciplinary field of digital humanities is concerned with the intersection of computer science, knowledge management and a wide range of humanities disciplines. Recent large-scale digitisation initiatives have made many important cultural heritage collections available online. This makes them accessible to the global research community and interested public for the first time. However, simple "one size fits all" web access is, in many cases, not appropriate in the digital humanities, due to the size and complexity of the artefacts. Furthermore, different types of users need varying levels of support, and every individual user has their own particular interests and priorities. Personalised and adaptive systems are thus important in helping users gain optimum engagement with these new digital humanities assets. Improved quality of access to cultural collections is a key objective of the CULTURA project [7]. Moreover, CULTURA supports a wide spectrum of users, ranging from members of the general public with specific interests, to users who may have a deep engagement with the cultural artefacts, such as professional and trainee researchers. To this end, CULTURA is delivering a corpus agnostic environment, with a suite of services to provide the supports and features required for such a diverse range of users.

[1] http://www.cultura-strep.eu/

These services include recommenders, where links to relevant resources, based on the current document's entities (people, places etc.) and the user's overall interests, are displayed alongside the resource. Other features enabled by CULTURA include the creation of guided lessons. Importantly, CULTURA stores a detailed model of user actions within its environment, so by monitoring changes in this model it is possible to adapt the lesson to user interests, as well as improving recommendations. Annotations are another key service that CULTURA offers, and they can be used for private notes, group collaboration, or for teaching aids. All these features are offered by CULTURA on top of a keyword search facility, a text normaliser service, an entity based browser and social network visualisations, which provide complementary ways of exploring and understanding a cultural heritage collection. Due to the service-based architecture the suite of services can be extended iteratively over time, allowing new features to be offered to existing and future collections.

In order to validate the CULTURA environment, two major collections have been selected - the 1641 Depositions[2], held in Trinity College Dublin, Ireland and the IPSA Illuminated Manuscript Collection[3], which is distributed between a number of museums and universities around the world. In terms of the case study application presented in section 4, the instance of CULTURA with the 1641 Depositions collection was used.

3. THE CULTURA EVALUATION MODEL

CULTURA incorporates a range of different services, which necessitate specific consideration in evaluation to get comprehensive outcomes on the quality of the system. The interaction Triptych model [5][15] has been used as a starting point for a conceptual analysis of the components and aspects of CULTURA. This model distinguishes three main components: system, content, and user. Between these components the axes and qualities of evaluation can be identified: performance (system-content axis), usefulness (content-user axis), and usability (system-user axis). The model was extended for CULTURA (see Figure 1) to address the qualities specific to the research environment and its services and to form the theoretical basis for evaluation studies. In the following the evaluation qualities covered by the model are presented.

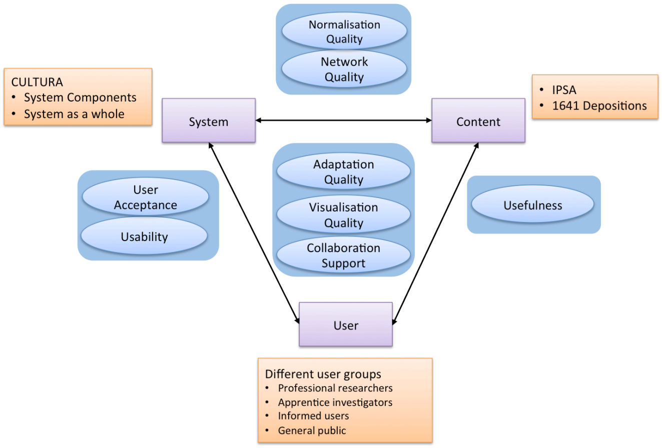


Figure 1. The CULTURA evaluation model.

Usefulness of content refers to the interaction between content and user: Is the content relevant and suitable for the user? This relates to the question whether the digital collection supports the personal user needs and/or the needs of the user group. A certain level of content usefulness is necessary for a meaningful evaluation of the other evaluation qualities.

Usability refers to the interaction between system and user: Does the system allow users to effectively, efficiently, and satisfactorily accomplish their tasks? This relates to whether the communication and interaction between user and system are smooth and whether the system is easy to use and learn. It also includes aspects of the learnability, navigation, and complexity of the system. This evaluation quality is often considered using the ISO standard as a reference for collecting evaluation data (e.g. [6]).

User acceptance has been considered on the system-user axis of the evaluation model in addition to usability: Do users consider the research environment and its services acceptable? Users may not necessarily have a positive attitude towards the system, even if it is technologically sound. Commonly, the following user acceptance aspects are distinguished [4]: ease of use (related to usability aspects), usefulness (of the system – to be distinguished from usefulness of content), and behavioural intentions to use.

Adaptation quality refers to the interaction between system, content, and user: Is the adaptation provided by the CULTURA system appropriate and useful? This relates to users' perceived benefit of system adaptation/recommenders received (user-centred viewpoint) [13]. It can also be related to layered evaluation of adaptation [2], examining whether user variables are correctly inferred and whether adaptation decisions are appropriately taken.

Visualisation quality also refers to the interaction between all three components, system, content, and user: How do users feel about the visualisations provided by the research environment? In the context of CULTURA social and influencer network visualisations of collection contents and user communities are applied. Visualisation quality relates to users' perceived benefit of the visualisations provided (helpfulness, insights gained etc.).

Collaboration support is another quality at the centre of the evaluation model, relating to the collaboration between the users of a research environment. It refers to the extent/quality to which users feel supported by the system in getting in contact with each other, and in exchanging information about the collection content.

Performance: *Normalisation quality and network quality*. The aspect of performance (system-content axis) is usually not directly visible to the users and often difficult to evaluate via user feedback. In CULTURA the performance aspect is operationalized as normalisation and network quality. Normalisation quality refers to text normalisation as well as entity extraction from text, i.e. to the quality and accuracy of the output of these processes. Network quality refers to whether relations between the entities of digital artefacts are accurately technically presented in the visualisations. Network quality thus investigates the accuracy of the data visualisations and the occurrence of inconsistencies between the entity data and the visualisation.

4. EMPRICAL CASE STUDY

To demonstrate the applicability and application of the evaluation model, this section presents a small-scale evaluation of the CULTURA environment for the 1641 Depositions collection with professional researchers. The study used an evaluation method that was set up in alignment with the evaluation model.

[2] http://1641.tcd.ie/

[3] http://www.ipsa-project.org/

4.1 Method

Evaluation instruments defined in line with the evaluation model were: an online survey covering items or scales on all evaluation qualities, semi-structured interviews, as well interaction logs as quantitative data complementing participants' self-reports. Thirteen professional researchers in history took part in the study. Log data was available for the whole sample, while only seven persons (6 male, 1 female) completed the survey. Participants were on average 39 years old, with a range from 28 to 47 years.

Participants were introduced to the CULTURA environment and its functionality. Subsequently, they had the possibility to use the system in their own time. Interaction data was recorded for the whole duration of usage. Users visited on average 15 pages while interacting with the system. Noticeably, the users were unlikely to 'play' with the system, but tended to ask for demonstrations of the different services. After working with the system, participants completed the online survey and took part in an interview.

4.2 Results

The *usefulness of content*, i.e. of the 1641 Depositions collection, was assessed very high, with M = 6.64 (SD = 0.94) on a scale ranging from 1-7, as it would be expectable for a user group with explicit expertise and interest in the digital collection in question.

The standard *usability assessment* [1] yielded an average score of 68.21 (SD = 19.18), indicating good usability (possible score range 0-100). Participants did not have a highly consistent perception of the system's usability, though, which might be due to a variable level of comfort with technology, in general, as it could be identified in the interviews. A number of participants highlighted the need for a guided tour introducing the system's features and how to use them.

User acceptance, assessed with an instrument adopted from prior research [14], was positive on all aspects (on a 1-7 scale, in each case), with the best result for behaviour intention (M = 6.0, SD = 1.83), arguing for participants' willingness and interest in actually using the system. The perceived usefulness of the CULTURA environment was also very good (M = 5.89, SD = 1.81), the score on ease of use (M = 5.11, SD = 1.88) was good.

For *adaptation quality* and *visualisation quality* subscores on estimated usage, usability, and perceived benefit, as well overall scores were calculated (see Figure 2). As can be seen a similar pattern can be identified for both qualities, with visualisations scoring generally slightly better than recommenders. Users indicated rather scarce usage of the recommenders and visualisations (recommenders M = 3.12, SD = 2.04; visualisations M = 3.57, SD = 2.15). Log data reinforces this finding: while 5 people did not use the visualisation service at all, the other 8 users visited 1 to on maximum 5 visualisations (M = 1.46, SD = 1.56 for N = 13); recommendations were on average used only 0.46 (SD = 1.66) times, and part of the users completely waived them. Overall quality scores, as well as usability and perceived benefit scores were assessed with medium quality in both cases, except for the benefit of visualisations, which scored more positively. Confirming this, in the interviews researchers expressed an appreciation of the possibilities offered for new insights into the Depositions by visualisations, but also pointed to the need for more flexible visualisations. Interviews also confirmed the usefulness of recommendations for exploring the collection – especially when not intimately familiar with the content, which explains why participants did not extensively use them themselves. In addition, users stressed the need for transparency – they wanted to know not only what is recommended, but also why they are seeing a particular recommendation.

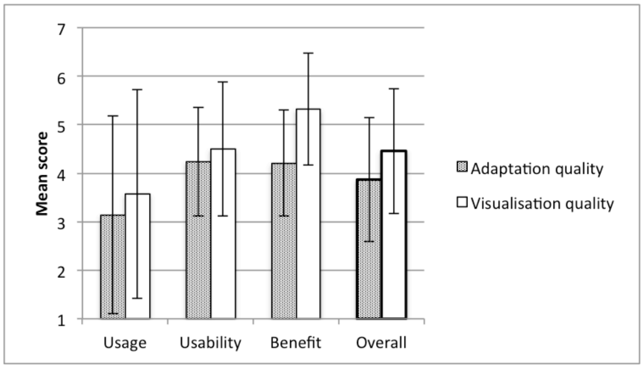


Figure 2: Results (mean scores with SD) on adaptation quality and visualisation quality.

Collaboration support was perceived as good (M = 5.36, SD = 0.99); researchers felt the system may assist them in collaborating with others. The assessment of the annotation service, which goes beyond pure collaboration features, was even better (M = 6.29, SD = 1.25). However, although users' were amenable to the idea of annotations, they did not take or share any annotations themselves.

Responses on normalised search were very positive, thus indicating an excellent user-centred assessment of *normalisation quality* (M = 6.64, SD = 0.64). Open comments highlighted that this feature is highly valuable and the majority of users had also made use this feature during the trial. Interviews, however, uncovered that users had concerns with respect to the accuracy of information extraction. In the lists of automatically extracted entities that appeared alongside the transcription of each deposition, the occurrence of errors – even if they were isolated – tended to have a devastating effect on users' confidence in the system. Addressing these concerns was of importance to their adoption of the research environment.

5. CONCLUSION

This paper introduced the evaluation approach developed and applied in the CULTURA project to respond to the evaluation needs of an innovative research environment for digital cultural heritage collections. Through the alignment of all evaluation tasks to the common evaluation model, general comparability of results is maintained. This is especially important since the CULTURA system is intended as a corpus agnostic research environment usable with different digital collections and addressing a broad range of user groups along the dimension of expertise. The final aim for evaluation is therefore to prove the benefit of the CULTURA environment independent of a specific collection and type of user. A comparison and consolidation of results over user groups and collections allows finding out about the benefits and issues that are of general interest and the overall quality of the research environment and its integrated services.

Evaluation studies other than the one presented herein have involved task-based evaluations (e.g. [11]), where users are requested to work on a predefined task while trialling the system. Such a task-based procedure enables a more detailed investigation of evaluation qualities, like adaptation quality or collaboration support, by drawing conclusions from the actual use of the system. This kind of procedure was unsuitable in the present case, though, since professional researchers wished to explore the system themselves without being forced to work on a given task. Moreover, other user studies on the CULTURA environment have also involved the comparison with the original web application of the respective digital collection or with the baseline version of the CULTURA system, without intelligent services.

Although participants in our case study acknowledged the general usefulness of the adaptive recommenders, their scarce actual usage highlights another important issue: the recommender service is rather intended for users with no or low prior knowledge in the collection than for expert researchers, who do not need or even do not want to have any guidance. This points up that the qualities of the evaluation model need to be investigated with appropriate groups of users, corresponding to the target audience of the services underlying this quality.

The results obtained from the presented study were consolidated with results obtained from other and larger scale user trials to derive implications for further development. Changes already implemented in the meantime were appreciated in more recent user trials and were singled out as being especially valuable in terms of building users' comfort with and confidence in the CULTURA environment.

The presented evaluation model is considered to have high potential for reuse in other research environments. It provides a valuable starting point for identifying the axes and topics of interest in other evaluation contexts and for specifying the actual evaluation design and evaluation instruments to be applied. The evaluation model is also used as a basis for the development of an evaluation service in the CULTURA project [12], aiming at supporting evaluators in planning, carrying out, and analysing evaluations. Through explicitly specifying the quality model underlying an evaluation, data collection can be systematized and automated reports based on the mapping to evaluation qualities can be derived. Future work will focus on triangulating data gathered via different modes (i.e. explicit retrospective or on-line user feedback, and non-invasive log and sensor data) by using the evaluation model as a reference base.

6. ACKNOWLEDGMENTS

The work reported has been partially supported by the CULTURA project, as part of the Seventh Framework Programme of the European Commission, Area "Digital Libraries and Digital Preservation" (ICT-2009.4.1), grant agreement no. 269973, and could not be realised without the close collaboration between all CULTURA partners.

7. REFERENCES

[1] Brooke, J. 1996. SUS: a „quick and dirty" usability scale. In P.W. Jordan, B. Thomas, B.A. Weerdmeester & a. L. McClelland (Eds.) *Usability evaluation in industry* (pp. 189-194). Taylor & Francis, London.

[2] Brusilovsky P., Karagiannidis C., and Sampson D. 2004. Layered evaluation of adaptive learning systems. *International Journal of Continuing Engineering Education and Life-Long Learning* 14, 402-421.

[3] Carmel, D., Zwerdling, N., and Yogev, S. 2012. Entity oriented search and exploration for cultural heritage collections. *World Wide Web 2012 European project track. Lion, France.*

[4] Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. 1989. User acceptance of computer technology: A comparison of two theoretical models. *Management Science* 35(8), 982-1003.

[5] Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, C.-P., Kovács, L., Landoni, M., Micsik, A., Papatheodorou, Peters, C., and Solvberg, I. 2007. Evaluation of digital libraries. *International Journal on Digital Libraries* 8, 21-38.

[6] Gediga, G., Hamborg, .-C., and Düntsch, I. 1999. The IsoMetrics usability inventory: An operationalisation of ISO 9242-10. *Behaviour and Information Technology* 18, 151-164.

[7] Hampson, C., Agosti, M., Orio, N., Bailey, E., Lawless, S., Conlan, O., and Wade, V. 2012. The CULTURA project: supporting next generation interaction with digital cultural heritage collections. In *Proceedings of the 4th International Euromed Conference, Limassol, Cyprus* (pp. 668-675). Springer, Heidelberg.

[8] Kauppinen, T., Mantegari, G., Paakkarinen, P., Kuittinen, H., Hyvönen, E., and Bandini, S. 2010. Determining relevance of imprecise temporal intervals for cultural heritage information retrieval. *International Journal of Human-Computer Studies* 68, 549-560.

[9] Koolen, M., Kamps, J., and Keijzer, V.d. 2009. Information retrieval in cultural heritage. *Interdisciplinary Science Reviews* 2-3, 268-284.

[10] Lawless, S., Hampson, C., Mitankin, P., and Gerdjikov, S. 2013. *Normalisation in historical text collections*. Accepted for publication at Digital Humanities, University of Nebraska-Lincoln.

[11] Lin, J. 2007. Is question answering better than information retrieval? A task-based evaluation framework for question series. In *Proceedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting* (pp. 212–219). Association for Computational Linguistics, Rochester.

[12] Nussbaumer, A., Hillemann, E.-C., Steiner, C.M., and Albert, D. 2012. An evaluation system for digital libraries. In Zaphiris, P., Buchanan, G., Rasmussen, E., and Loizides, F. (Eds.), *Theory and practice of digital libraries. Second International Conference, TPDL 2012. LNCS vol. 7489* (pp. 414-419). Springer, Berlin.

[13] Steiner, C.M. and Albert, D. (2012). Tailor-made or unfledged? Evaluating the quality of adaptive eLearning. In Psaromiligkos, A. Spyridakos, & S. Retalis (Eds.), *Evaluation in e-learning* (pp. 111-143). Nova Science, New York.

[14] Thong, J.Y.L., Hong, W., and Tam, K.-Y. 2002. Understanding user acceptance of digital libraries: what are the roles of interface characteristics, organizational context, and individual differences? *International Journal of Human-Computer Studies* 57, 215-242.

[15] Tsakonas, G. and Papatheodorou, G. 2006. Analysing and evaluating usefulness and usability in electronic information services. *Journal of Information Science* 32, 400-419

A Social Network Study of Evliyâ Çelebi's *The Book of Travels-Seyahatnâme*

Ceyhun Karbeyaz*
Bilkent University
Department of Computer Engineering
Ankara 06800, Turkey
karbeyaz@bilkent.edu.tr

Ethem F. Can†
Bilkent University
Department of Computer Engineering
Ankara 06800, Turkey
ethemfatih.can@gmail.com

Fazli Can
Bilkent University
Department of Computer Engineering
Ankara 06800, Turkey
canf@cs.bilkent.edu.tr

Mehmet Kalpakli
Bilkent University
Department of History
Ankara 06800, Turkey
kalpakli@bilkent.edu.tr

ABSTRACT

Evliyâ Çelebi, an Ottoman writer, scholar and world traveler, visited most of the territories and also some of the neighboring countries of the Ottoman Empire in the 17^{th} century. He took notes about his trips and wrote a 10-volume book called *Seyahatnâme (The Book of Travels)*. In this paper, we present two methods for constructing social networks by using textual data and apply it to *Seyahatnâme-Bitlis Section* from book IV and check if the constructed networks hold social network properties. The first social network construction method is based on proximity of co-occurence of names. The second method is based on 2-pair associations obtained by association rule mining by using sliding text blocks as transactions. The social networks obtained by these two methods are validated using a Monte Carlo approach by comparing them with the social network created by a scholar-historian.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: [Data mining]

General Terms

Experimentation, Verification

Keywords

Social networks, data mining, historical documents

1. INTRODUCTION

Evliyâ Çelebi; a 17^{th} century Ottoman writer, scholar, and world traveler (born on 1611, died circa 1682); visited most of the territories and also some of the neighboring countries in Africa, Asia and Europe of the Ottoman Empire over a period of 40 years. His work *Seyahatnâme (The Book of Travels)* is known by its distinguished style and detailed descriptions of people and places that he visited during his long journeys [4]. A preliminary version of this study can be seen in [10].

Seyahatnâme is a long masterpiece that is composed of ten books each containing information about different locations and several prominent people of its time. This nature of the work is appealing for social network studies which aim to identify relationships among people. He described the people and the incidents he came across in Bitlis in book IV and V. In our study we aim to automatically identify relationships among the people who appear in the text. These identified relationships between the important historical characters would open new research avenues [12]. For this purpose we use the text in transcribed form [9] and perform several experiments.

The contributions of this study can be summarized as follows. We present two different methods for constructing social networks from textual data and apply them to *Seyahatnâme-Bitlis Section* to obtain a social network that represent relationships among people. The first social network construction method we present is based on proximity of co-occurence of names. The second method is based on 2-pair associations obtained by association rule mining [1]. We use the social network created by a human expert as the ground truth and assess the effectiveness of the methods by comparing the generated network structure with that of the ground truth. We use a Monte Carlo approach for validating the automatically constructed social networks and show that the social network structures obtained by our methods are significantly different from random. We also analyze

*Current Address: Department of Electrical and Computer Engineering, Northeastern University, 02115 USA

†Current Address: Department of Computer Science, University of Massachusetts Amherst, 01003 USA

ENRICH 2013 Workshop, August 1, 2013, Dublin, Ireland.

the manually and automatically generated networks to see if they contain the social network properties.

2. RELATED WORK

One of the methods we present in this study is based on association rules, which are the derived relations between the items of a dataset. The notion was first introduced by Agrawal et al. [1] and later the notion of assocation rules are used in many studies. In this method, let I = $\{i_1, i_2, ..., i_n\}$ be a set of items with size n and T = $\{t_1, t_2 ,..., t_m\}$ is set of transactions (in market data analysis a transaction involves the group of items purchased together) with size m. Then an association rule is shown as $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \oslash$. The aim of the association rules is to find all pairs of items within all T that have support S and confidence C values greater than user defined amounts. S is a threshold to extract the frequent itemsets from transactions whereas C is the ratio of support for items X and Y occuring together to support of item X. The best-known algorithm to mine association rules is Apriori [2].

Brin et al. [3] use association rule mining to find frequent itemsets over market data. They show that their algorithm provides better performance with respect to the Apriori algorithm while finding large itemsets. Raeder and Chawla [13] applied association rules over construction and analysis of a social network from market basket data. Their main goal was to search for meaningful relationships over the formed social network. They conclude that the highly rated product communities have a significant relationship with a clear purpose in the social network.

3. METHODS: ProxiBM and RuleBM

In this paper, we present two social network construction methods. These are the text proximity-based method (ProxiBM) and the association rule-based method (RuleBM). Both methods are based on co-occurrence of names in close proximity within a text block.

For determing text blocks with a meaningful cohesive context we use two approaches. In ProxiBM we use the paragraph information provided in the transcribed text. We manually identified and tagged 164 paragraphs. Each paragraph is used as a block. However, in the original work of Evliyâ Çelebi there are no explicit paragraphs. Furthermore, manual paragraph tagging is difficult. This provides a motivation for automatic identification of blocks, hence in RuleBM we employ a sliding text window approach and use a certain number of consecutive words as a block. The obtained blocks partially overlap. Since blocks are overlapping and obtained from consecutive words we intuitively expect that they will at least have a (partially) cohesive context. The sliding blocking approach is explained in Figure 1.

In ProxiBM, edges for the undirected graph of social network are derived by creating a link between every character that appear in the same paragraph within a close word proximity (using a threshold). The proximity threshold between any two names is varied between 5 words to 500 words in steps. This approach is inspired by the use of term closeness as an indicator of document relevance [6] and by an earlier study that uses a similar approach in social network analysis [14].

In RuleBM we use the sliding text window for blocking and treat each block as a transaction. The names that appear in a block correspond to shopping items; in this way, we are able to extract association rules [1]. By using the Apriori algorithm[2] we derive the 2-item pairs from the transactions by using a support threshold value. (Note that in RuleBM association rules that involve more than two character names are too few to use.) In RuleBM we use the 2-pair association rules as relational edges of the social network. We employ different support threshold values and repeat the blocking operation for different block and sliding window sizes in order to find the best performing parameters.

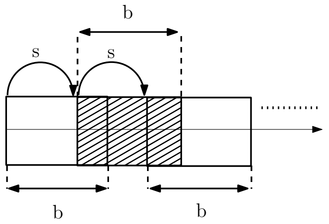


Figure 1: Sliding window-based blocking: l: total text length, b: block size ($0 < b \le l$), s: step size, nb: number blocks, $nb = 1 + \lceil \frac{(l-b)}{s} \rceil$ for $0 < s \le b$, $nb = 1 + \lfloor \frac{(l-b)}{s} \rfloor$ for $s > b$.

4. MEASURING EFFECTIVENESS

For measuring the effectiveness of the methods in predicting the correct links we use the social network created by Prof. Kalpaklı, a scholar-historian. The agreement between automatically constructed social networks and the manually constructed (actual) social network is measured by *precision*, *recall*, and the *F-measure* [11] (their formulas are given below). In *precision* we calculate what fraction of of automatically generated links are correct. In *recall* we calculate what fraction of actual links is identified. *F-measure* is a harmonic mean of these two measures [11]. They all assume a value between 0 (worst case: no match at all) and 1 (best case: perfect match). While constructing the links only the people of that time are considered, i.e., names of people who were not alive, prophet names, different names of god etc. are excluded.

$$Precision = \frac{No.\ of\ matching\ edges}{No.\ of\ edges\ obtained\ by\ method} \tag{1}$$

$$Recall = \frac{No.\ of\ matching\ edges}{No.\ of\ edges\ of\ manually\ constructed\ network} \tag{2}$$

$$F = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

5. EXPERIMENTAL RESULTS

5.1 Generation and Validation of Social Networks

In the experiments we first measure the accuracy of automatically generated social network by measuring their similarity with the ground truth data using *precision*, *recall*, and *F-measure*. In the experiments, both methods are tested in various conditions in order to find their best matching (most similar) configuration to the ground truth. For ProxiBM we use various proximity threshold values. For RuleBM we report the results for block sizes 500 and 1,000 words (with block size smaller than 500 and larger than 1,000 we obtain low effectiveness values and we do not report them here). Table 1 shows the *precision*, *recall* and *F-measure* results of ProxiBM for different proximity threshold values. The best configuration for this method with and *F-measure* value of 0.59 is observed when proximity threshold is 25 words.

Table 1: Performance results of ProxiBM over paragraphs for different proximity threshold (Θ) values in terms of no. of words.

Measure	$\Theta = 5$	$\Theta = 10$	$\Theta = 25$	$\Theta = 50$	$\Theta = 100$	$\Theta = 250$	$\Theta = 500$
Precision	0.47	0.52	0.54	0.49	0.48	0.46	0.45
Recall	0.16	0.39	0.66	0.70	0.70	0.71	0.71
F-measure	0.24	0.44	0.59	0.58	0.57	0.56	0.56

A similar experiment is done for RuleBM that is based on assocation rule mining. For this purpose, exactly the same blocking operation is done by considering sliding window sizes. And again the experiment is performed for block sizes 500 words and 1,000 words. Association rules are derived from these blocks for different support thresholds ranging from 5% to 20%. The best configuration for this method with and *F-measure* value of 0.29 is observed when blocksize: 500 words, stepsize: 300 words and support threshold: 5%.

Another set of experiments are conducted that aim to understand if the automatically generated social networks are significantly different from random. For this purpose Monte Carlo experiments are performed [8]. Monte Carlo experiments define a reference population and provide a baseline distribution [8, pp. 161-162]. Note that all randomly created social networks that are used in our tests have the same network properties with the automatically generated network that is being evaluated (e.g., the same number of nodes and the same average degree distribution within the nodes where degree of a node is the number of incoming edges to that node). In order to achieve this the Erdős-Renyi random network generation algorithm is used [5]. In all Monte Carlo experiments, we generate a random version of the social network which is being evaluated 1,000 times and measure the average *F-measure* values. The reported *F-measure* values are obtained from the averages obtained for 1,000 *precision* and *recall* observations. The comparison of average *F-measure* results for the proposed methods versus Monte Carlo average *F-measure* results for different experimental conditions are provided in Figure 2 and Figure 3 for ProxiBM and RuleBM, respectively. The plots show that both methods with proper parameters generate networks which are significantly different from random.

5.2 Social Network Analysis

In this section, we first show that the constructed social networks have the small world property. The social networks, which have small world concept, have short average path length and relatively high clustering coefficient [14]. In Table 2, we provide the network properties the average path length l and the clustering coefficient C of the social networks for ProxiBM and RuleBM with their best configuration and also for the manually created ground truth network. For comparison, we also provide clustering coefficient C_{rand} of the random graphs that have the same average degree and size of each network. The results show that the networks hold the small world property.

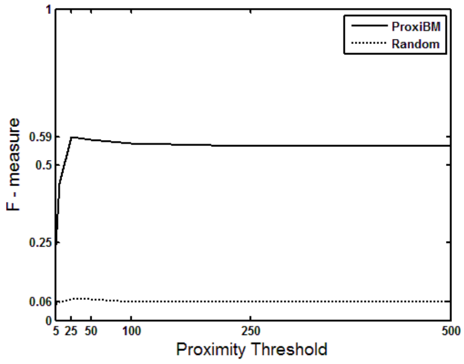


Figure 2: ProxiBM results vs. Random (Monte Carlo) results for different proximity threshold values.

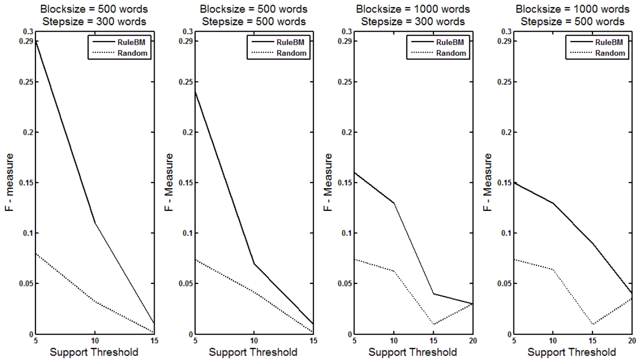


Figure 3: RuleBM results vs. Random (Monte Carlo) results for different blocksize, stepsize and support threshold values.

Table 2: Characteristics of social networks for both methods with the best configuration and ground truth compared to random networks that have the same average degrees <*k*> and size (no. of nodes in network) values.

Network	*Size*	*No.of Edges*	$<k>$	l	C	C_{rand}
Ground Truth	71	321	9.04	1.90	0.93	0.15
ProxiBM	88	395	8.98	3.34	0.83	0.11
RuleBM	82	515	12.56	2.32	0.86	0.14

We also examine the social networks about node degree distribution. $P(k)$ is the distribution function and it denotes the probability that a randomly selected node has k edges. The distribution function of social networks with small world property has a power-law degree distribution whereas in the randomly generated networks with the same network properties it has a Poisson distribution [14]. Power law distribution is denoted as follows.

$$P(k) \sim k^{-\gamma} \tag{4}$$

In the above formula γ is called the scaling factor. Networks with power law distribution are called scale-free networks. In Power-law degree distribution, there are numerous nodes that have small degrees whereas few nodes have high degrees [14]. Degree distribution of the characters with the highest degrees for the actual network and the networks that are created by RuleBM and ProxiBM are listed in Table 3. The table shows that for decreasing degree k values, frequency of nodes with the same degrees increases. The degree distributions of all constructed networks for *Bitlis Section* follow a power law distribution rather than a Poisson distribution. The scaling factor γ of the distribution is calculated as 1.79 for actual network, 1.75 for ProxiBM and 1.71 for RuleBM best performing networks.

Table 3: Degree distribution of characters with the highest degrees for best configurations of ProxiBM and RuleBM along with actual social network for *Seyahâtname-Bitlis Section*.

Actual Network		ProxiBM		RuleBM	
Character	Degree	Character	Degree	Character	Degree
Abdâl Hân	69	Ziyâeddin Beğ	20	Abdâl Hân	66
Ziyâeddin Beğ	20	Selmân	20	Beşaret Ağa	40
Şeref Beğ	18	Beşaret Ağa	19	Şeref Beğ	32
Beşaret Ağa	17	Şemseddîn	16	Hüsrev Paşa	30
Haydar Ağa	15	Hasan Beğ	16	Haydar Ağa	27
Cündevân	13	Haydar Ağa	14	Zâl Paşa	27
Salmân-u Buhtî Ağa	13	Şeref Beğ	13	Salmân-u Buhtî Ağa	27
Racoy Ağa	13	Maktûl Haydar Kethudâ	13	Ziyâeddin Beğ	26
Bedir Beğ	13	Racoy Ağa	13	Şeref Beğ	23
Şemseddin Beğ	13	Seyfi Ağa	13	Âlemşâh Beğ	23
Âlemşâh Beğ	13	Bedir Beğ	13	Yaşar Beğ	23
Kerrârkulu Beğ	13	Siyâvuş	13	Cündevân	23
Yaşar Beğ	13	Kâzım Sührâb	13	Racoy Ağa	23
Seyfi Ağa	13	Salmân-u Buhtî Ağa	12	Seyfi Ağa	23
Vildân	12	Kevkeban	12	Süleymân Hân	23

6. CONCLUSION AND FUTURE WORK

In this work we introduce two methods for constructing social networks: ProxiBM and RuleBM by using textual data and apply it to the Bitlis Section of Seyahatnâme. We also check if the constructed networks hold social network properties. Both methods generate meaningful social networks which are significantly different from random and substantially similar to the social network manually constructed by a scholar-historian. The experimental results show that the networks created by ProxiBM show a higher similarity to the manually created social network than those of RuleBM. However, the disadvantage of ProxiBM is that it requires more focused (cohesive) blocks obtained from paragraphs. On the other hand, RuleBM is more flexible since it simply exploits blocks obtained from a sliding text window.

It is possible to obtain a better performance with RuleBM if we use contextually meaningful sliding text blocks: Our preliminary experiments with sliding paragraph-based text blocks provide evidence in that direction. For the construction of such cohesive units we may use an automatic text segmentation method [7]. Obtaining blocks automatically is also important for ProxiBM since natural cohesive units may not be readily available and their manual tagging is usually impractical. Furthermore, the results of these two methods can be combined in some elaborate data fusion techniques.

7. ACKNOWLEDGEMENTS.

This work is partially supported by the Scientific and Technical Research Council of Turkey (TÜBITAK) under the grant number 109E006. Any opinions, findings and conclusions or recommendations expressed in this article belong to the authors and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Database mining: A performance perspective. *IEEE Transactions on Knowledge and Data Engineering*, 5:914–925, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[3] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. *SIGMOD Rec.*, 26:255–264, June 1997.

[4] R. Dankoff. *Evliyâ Çelebi in Bitlis: The relevant section of the Seyahatnâme*. E. J. Brill, Netherlands, 1990.

[5] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[6] D. Hawking and P. Thistlewaite. Relevance weighting using distance between term occurrences. http://david-hawking.net/pubs/HawkingT96.pdf, 1996.

[7] M. A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[8] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[9] S. A. Kahraman and Y. Dağlı. *Evliyâ Çelebi Seyahatnâmesi 4. Kitap*. Yapı Kredi Yayınları, İstanbul, 2003.

[10] C. Karbeyaz, E. F. Can, F. Can, and M. Kalpakli. A content-based social network study of Evliyâ Çelebi's *Seyahatnâme-Bitlis Section*. In E. Gelenbe, R. Lent, and G. Sakellari, editors, *Computer and Information Sciences II*, pages 271–275. Springer London, 2011.

[11] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[12] J. Preiser-Kapeller. Calculating Byzantium? Social network analysis and complexity sciences as tools for the exploration of medieval social dynamics. http://www.oeaw.ac.at/byzanz/repository/Preiser_ WorkingPapers_ Calculating_ I.pdf, 2010.

[13] T. Raeder and N. V. Chawla. Modeling a store's product space as a social network. In *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*, pages 164–169, Washington, DC, USA, 2009. IEEE Computer Society.

[14] A. Özgür and H. Bingöl. Social network of co-occurrence in news articles. In C. Aykanat, T. Dayar, and I. Körpeoglu, editors, *Computer and Information Sciences - ISCIS 2004*, volume 3280 of *Lecture Notes in Computer Science*, pages 688–695. Springer Berlin Heidelberg, 2004.

Publishing Social Sciences Datasets as Linked Data: a Political Violence Case Study

Rob Brennan
Knowledge and Data Engineering Group
Trinity College Dublin, Ireland
rob.brennan@cs.tcd.ie

Kevin C. Feeney
Knowledge and Data Engineering Group
Trinity College Dublin, Ireland
kevin.feeney@cs.tcd.ie

Odhrán Gavin
Knowledge and Data Engineering Group
Trinity College Dublin, Ireland
gavinod@cs.tcd.ie

ABSTRACT

This paper discusses the design, application and generalization of a Linked Data vocabulary to describe historical events of political violence. The vocabulary was designed to capture the United States political violence 1795-2010 dataset created by Prof. Peter Turchin and has been generalized to support a semi-automated data collection process suitable for the creation of a complimentary dataset of political violence events in the UK and Ireland. Both datasets will be published as managed linked data that is inter-connected with other web-based datasets such as DBpedia, a computer-readable version of Wikipedia. The lifecycle of the datasets will be actively managed with tool support for further harvesting, evolution and consistency checking. The harvesting tool, data harvesting process, political violence vocabulary and US political violence dataset were connected to our existing linked data management platform, DaCura.This political violence vocabulary described herein has been validated by application to a real-world dataset and publication use-cases. Our data harvesting process is potentially applicable to a wide range of social science or historical research activities that focus on generating structured data-sets or annotations of human-readable corpora. The publication of the US political violence dataset as linked data is a contribution towards the emerging fields of Digital Humanities and Linked Science. This paper describes a new linked data vocabulary for political violence events, provides insights into the processes of creating a new vocabulary for social science datasets. It also illustrates the potential benefits of publishing social science or other cultural heritage datasets as linked data.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: User / Machine Systems – *Human information processing*. H.2.1 [**Database Management**]: Logical Design – *Data models, Schema and sub-schema*. H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Retrieval models, Selection process*.

General Terms

Management, Documentation, Design, Experimentation, Human Factors, Standardization.

Keywords

Linked Data, Vocabularies, RDF, Schema Design, Cliodynamics, Data Curation.

1. INTRODUCTION

The collection and curation of structured data-sets from unstructured and semi-structured sources is a common requirement for research both in social sciences and more general cultural heritage projects [1]. Linked Open Data (LOD) approaches to online data publishing are based upon RDF and semantic-web technologies such as RDFS, OWL and SPARQL. These should, in theory, be a very attractive solution for harvesting, curating and publishing structured social science or humanities data-sets.

In this paper, we describe a case-study of an approach to migrating a social-science dataset to an LOD platform. The dataset in question is the United States political violence 1795-2010 dataset created by Prof. Peter Turchin in the course of his research into Cliodynamics [2]. The dataset was originally distributed as an Excel spreadsheet, consisting of 1828 event records, each of which had several properties associated with it. This process formed a test-case of the DaCura system which we have been developing in Trinity College Dublin [3]. That system is designed to provide easy-to-use tool support for non-expert users to allow them to easily harvest data from web-based sources into an RDF based triple-store. It furthermore provides support for the management of that data-set over time with a focus on supporting constrained schema evolution.

The focus on this paper is on the process by which we designed the LOD schema from the original dataset spreadsheet. The schema is represented as an RDFS (RDF Schema) vocabulary. In designing this schema we had the following goals:

1. Re-use, wherever possible, existing LOD vocabularies to represent the events and their properties in the data set.

2. Provide support within the schema for the process by which the data is collected and not just the final data format. Thus, for example, a requirement is that we can capture candidate events in our dataset which may need to be approved for inclusion in the final dataset by a domain expert.

3. Design the schema in such a way that it would integrate well with our DaCura platform. DaCura provides several features such as the ability to generate simple web-based widgets to represent dataset instances. To take full advantage of this facility certain properties must be present in the schema.

In designing our schema, we attempted to describe entities in a general and extensible way while minimizing the overall complexity of the schemata upon which we were relying. Rather than trying to define everything in an entirely general way, we attempted to steer a pragmatic middle-ground between generality and specificity and only introduced more general schema in situations where we could envisage future situations in which we might take advantage of this generality.

2. SYSTEM DESIGN

This section discusses the development of the political violence vocabulary, a formal process for harvesting political violence events from a historical corpus, our harvesting tool and finally the online repository for political violence datasets.

2.1 Political Violence Vocabulary Design

There were five distinct activities involved in the vocabulary design: survey of other vocabularies; examination of the original US dataset; consideration of the requirements for the UK and Ireland dataset; the semantic uplift process and creating interlinks to other linked data datasets. Each of these activities is discussed below.

2.1.1 Survey of Other RDF Vocabularies

One of the key features of vocabularies based on RDF (Resource Description Framework) is that they can easily be combined to produce larger models. RDF-based systems do not depend on the existence of a single, canonical ontology into which every vocabulary or specialized ontology must fit. This frees vocabulary designers to create domain or application specific designs but it also creates a proliferation of overlapping vocabularies published on the web. In recent years the Linked Data community [4] has focused on reuse of a few well-known vocabularies such as the Dublin Core metadata for describing documents. This has the beneficial outcome of reducing the requirements for applications that consume linked data, as terms defined by these common vocabularies appear again and again in datasets published on the web.

2.1.2 Evaluate and Analyze the Example Dataset

The United States Political Violence (USPV) dataset was initially compiled in order to assist research into the dynamics of political instability in the United States [2]. It was compiled from a number of sources and was published as a spreadsheet consisting of 1,828 reports of incidents of violence, recording date, category, motivation, fatalities, location, source, a description of the event, and research-specific coding. In conjunction with the appendix to [2], historical research was undertaken in order to formulate precise definitions of the types of political violence events in the dataset, as described by the category and motivation fields. Our vocabulary was designed to ensure that all information contained within the published dataset could be captured without loss.

Two features of the dataset particularly informed design choices in the vocabulary. The presence of duplicate reports in the dataset led to the decision to differentiate between reports and events. The presence of reports marked with question marks to indicate uncertainty, led us to decide to include the capability to report levels of uncertainty about reports of political violence.

2.1.3 Generalisation to UK and Ireland Dataset

Historical knowledge of the period 1785-2007 was used to determine the suitability of the vocabulary, based on the USPV dataset, for the United Kingdom and Ireland Political Violence (UKIPV) dataset. In most cases, vocabulary terms used to describe political violence in the United States were also appropriate to describe political violence events in the United Kingdom and Ireland. However, due to historical differences between the two regions, a small number of terms describing motivations required changes in order to capture the characteristics of political violence for the UKIPV dataset more accurately.

2.1.4 Semantic Uplift

We define semantic uplift as the process of converting non-RDF data, for example the original US political violence spreadsheet, into an RDF-based knowledge representation such as a set of RDF triples describing the individual events according to the Political Violence vocabulary. Semantic uplift is often ignored in favor of focusing on schema modeling tasks. However it has an important impact on the vocabulary design process. Converting events into RDF exercises the vocabulary and exposes flaws or weaknesses. In our case the semantic uplift process was written as a PHP script that processed a CSV (comma separated value) representation of the spreadsheet.

2.1.5 Creating Links to other Linked Data Datasets

One of the major motivations for publishing the political violence datasets as (RDF-based) linked data is to enable combination of the data with other datasets already available on the web. In theory once the dataset is published as RDF on the web it is available to all RDF-consuming applications. However this can place onerous requirements on those applications if a new vocabulary is used and no interlinks are created between the political violence dataset and already existing datasets. In general this means that generic, browsing-oriented applications are able to display the data but that more sophisticated use cases such as mash-ups of the data are less likely.

At the dataset consumption level, enabling discovery is a topic addressed by several ongoing research efforts such as the Data Hub / CKAN by the Open Knowledge foundation and the Sindice semantic web index by DERI [5]. At the vocabulary level it is possible to reuse common vocabularies such as Dublin Core that are often used in linked data datasets. At the dataset level it is possible to include interlinks to instances in other datasets. For example when recording the location of an event as the US state of Ohio it may be preferable to record this as the instance of that concept defined by the Dbpedia or Geonames datasets. Thus a "dbpediaLocation" property is defined in the PV vocabulary which enables us to directly embed references to instances of the DBpedia concept "Place".

2.2 A Data Harvesting Process

The manual process of extracting US political violence events from the historical record was described by Turchin in his analysis of that data-set [2]. However for this work it was necessary to formalize and document the harvesting process model with six goals in mind:

1. Establishing the requirements placed by the collection process on the political violence vocabulary in terms of what concepts need to be modeled.

2. Establishing the possible actors or roles in the data collection process.

3. Specializing the process to consider the requirements placed on it by the UKIPV dataset sources.

4. Reviewing the process with respect to the possible activities where automation could both be beneficial and could leverage the advantages of having a formal vocabulary describing the data being extracted.

5. Linking the data collection process to our previous work on DaCura, a managed linked data curation platform [3].

6. Determining the experimental process by which we would gather data to validate the utility of our tool support for data collection, validation, publication and management of the datasets.

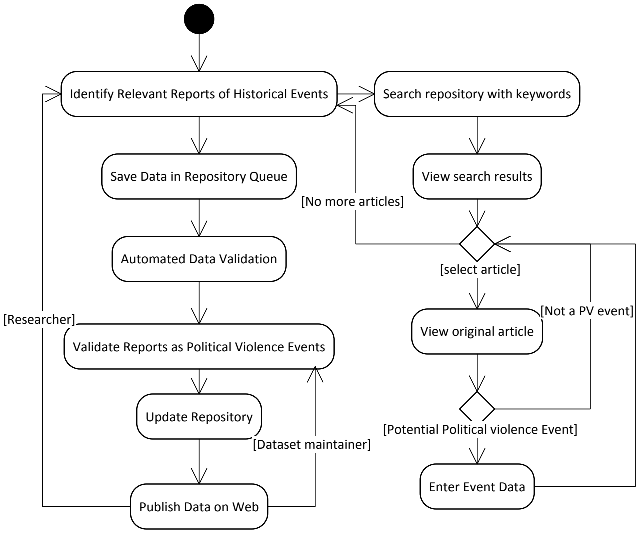

Identify Relevant Reports of Historical Events
Search repository with keywords
Save Data in Repository Queue
View search results
[No more articles]
Automated Data Validation
[select article]
[Researcher]
View original article
[Not a PV event]
Validate Reports as Political Violence Events
Update Repository
[Potential Political violence Event]
[Dataset maintainer]
Enter Event Data
Publish Data on Web

Figure 1. The political violence data harvesting process.

Figure 1 illustrates the data harvesting process as a UML activity diagram. The left hand side of the figure focuses on the overall data lifecycle. In this lifecycle, events are identified in the repository by a researcher, then data is validated as conformant to the vocabulary by the DaCura platform, then a dataset maintainer examines the report data to validate whether or not it should be recorded as a new political violence event (or a duplicate, or out of scope, etc) and finally the data is updated as linked data on the web. This is an iterative process that can continue as long as there is event data to be found or maintained.

The right-hand side of the figure shows details of the event report extraction from the online newspaper repository. First a set of search keywords are used to retrieve a set of articles that are candidates for event reports. The researcher then views each article in turn to evaluate it against the requirements for inclusion in the dataset as a report. If it is to be included then data about the article and the underlying event as reported is recorded. This event report data is then placed into the overall process flow on the left-hand side of the figure.

2.3 Political Violence Vocabulary

The approximate structure of the political violence vocabulary is represented as a UML class diagram in figure 2. This is an approximation because the RDF semantics do not exactly align with the object oriented modeling assumptions of UML. There are three main classes defined: the historical Event and its two subclasses, the Report and the Political Violence Event.

In addition to all the classes used to model the properties of events (on the right of the figure and discussed further below) the vocabulary makes use of the Open Annotation Data model [7] vocabulary to enable researchers or other consumers of the data-set to annotate individual dataset elements.

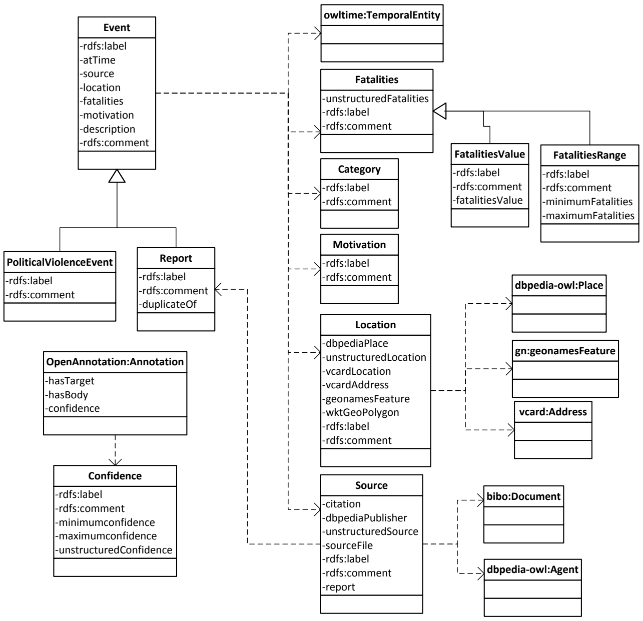


Figure 2. UML class diagram illustrating vocabulary structure.

2.3.1 Vocabulary Terms

The basic building-block of the dataset is our concept of an event, which is defined as any individual historical event. Based on the dataset and requirements, events were further subdivided into two classes, political violence events and reports. A report refers to a source's record of an event, e.g. a newspaper article. A political violence event refers to the event itself. In general, political violence events are referred to by one or more reports. This division reproduces both the existence of duplicate records of events in the original USPV spreadsheet, and the occurrence of multiple reports of individual historical events in the historical source material for the UKIPV dataset.

2.3.2 Categories

The category class identifies what form the political violence event takes. In the USPV and work based on it [2], most events are categorized into one of four categories – assassination, terrorism, lynching, and riot – based on the number of perpetrators and victims. There are a number of other categories which describe less commonly-occurring political violence events. The most common of these is rampage, which refers to events such as school and workplace shootings. The remaining categories describe uncommon events or are excluded from the analysis, and are included to fully capture the USPV dataset.

2.3.3 Motivations

The motivation class describes the reasons political violence event occurred. Events may have multiple motivations if they have numerous or complex causes.

2.4 Links to other DataSets

The value of datasets is expanded if it is possible to easily combine them with other datasets already published on the web. Hence this vocabulary contains multiple connection points to three important linked data datasets: (1) DBpedia, the RDF version of Wikipedia [8], (2) Geonames, a geographical database accessible through RDF and (3) vCard a vocabulary for representing people and organizations in RDF that is reused in many open datasets. These links are created by creating properties in the PV vocabulary that reference the other datasets.

2.5 Integration with DaCura software

The DaCura system is designed to improve the manageability of RDF datasets over time by imposing a set of constraints on RDF schemas and updates to RDF datasets above and beyond those that are mandated by RDF and RDFS standards themselves. For example, it requires that properties must have labels specified and requires that classes cannot be removed from a schema if there are instances of those classes in the dataset. It also defines naming conventions for RDF URLs. The combined effect of these constraints is to allow schema and dataset evolution while maintaining the consistency of the dataset over time.

3. RELATED WORK

Vocabulary and ontology design is an evolving subject area as the actual deployment of Semantic Web technologies and Linked Data is immature. The focus of theoretical and practical design concerns have rarely overlapped. A major venue for this debate is the annual Workshop on Ontology Patterns [9]. However Dodds and Davis [10] give a concrete set of examples for designs that are based on Linked Data use cases and were influential on this paper.

Shaw et al. [6] provide an overview of current ontologies for representing events in RDF and show the common attributes of event representations and how the differing modelling approaches tackle each aspect. In addition they provide a “Linked Open Data Event Model” (LODE) that encapsulates the common attributes in other representations but concentrates. This is a laudable and useful outcome but it was found to be lacking for our application to political violence datasets in two main respects. First, it assumes that these factual aspects represent some form of “consensus reality” whereas in harvesting data from the London Times archive it is often found that newspaper reports over time can be inconsistent or contain incorrect factual assertions. Second, it uses the DOLCE+DnS Ultralite [11] upper ontology for several property value types and we didn’t want to be constrained to using such an abstract and complex description of our dataset because of the resultant complexity in querying the dataset.

4. CONCLUSION AND FUTURE WORK

In this paper we have examined the process of generating a vocabulary to support extraction of political violence event data from online historical sources. The ontology is flexible enough to capture the original US political violence dataset while still supporting the needs of the proposed UK and Ireland political violence dataset. It is potentially suitable for collecting political violence event data from other sources. Using this vocabulary, we have created a set of tools which allow for harvesting and collation of political violence events. These tools will be used to construct the UK and Ireland political violence dataset. They will also underpin the experimental process examining the utility of tool support for collecting and managing linked data datasets.

Future work will involve extending the functionality of the data extraction toolset. Currently, candidate political violence reports are selected via a small set of searches chosen to offer acceptable and consistent precision and recall. We intend to provide users with the facility to suggest potentially useful search terms after data retrieval, in order to improve the precision and/or recall of the results. Another planned feature is to implement a domain expert (historian or social scientist) moderator queue.

5. ACKNOWLEDGEMENTS

The authors wish to thank Prof. Declan O’Sullivan and Prof. Peter Turchin for their encouragement and feedback on this work.

6. REFERENCES

[1] Hampson, C., Agosti, M., Orio, N., Bailey, E., Lawless, S., Conlan, O., and Wade, V. 2012. The CULTURA Project: Supporting Next Generation Interaction with Digital Cultural Heritage Collections. In progress in Cultural Heritage Preservation - 4th International Conference (EuroMed 2012). 668-675. Lecture Notes in Computer Science (LNCS) 7616, Springer, Heidelberg, Germany.

[2] Turchin, P. 2012. Dynamics of political instability in the United States, 1780–2010. Journal of Peace Research, 49(4). 577-591. DOI:10.1177/0022343312442078

[3] Tai, W., Feeney, K., Brennan, R., and O'Sullivan, D. 2012. Manageable Dataset Curation for Linked Data. 18th International Conference on Knowledge Engineering and Knowledge Management, (EKAW, 8 - 12 October, Galway, Ireland, 2012).

[4] Bizer, C., Heath, T., and Berners-Lee, T. 2009. Linked Data - The Story So Far, International Journal on Semantic Web and Information Systems (IJSWIS), 5 (3). 1–22.

[5] Käfer, T., Umbrich, J., Hogan, A. and Polleres, A. 2012. Towards a Dynamic Linked Data Observatory. WWW2012 Workshop: Linked Data on the Web (LDOW2012, Lyon, France, 16 April, 2012).

[6] Shaw, R., Troncy R., and Hardman L. 2009. LODE: Linking Open Descriptions of Events. In Gómez-Pérez A., Yong, Y., and Ying, D. (eds.), Proceedings of the 4th Asian Conference on The Semantic Web (ASWC '09), Springer-Verlag, Berlin, Heidelberg, 153-167. DOI=http://dx.doi.org/10.1007/978-3-642-10871-6_11

[7] Sanderson, R., Ciccarese, P., and Van de Sompel, H. (eds.) 2013. Open Annotation Data Model. Community Draft, 08 February 2013. Retrieved June 9, 2013 from W3C: http://www.openannotation.org/spec/core/20130208/

[8] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. 2009. DBpedia - A crystallization point for the Web of Data. Web Semantics. Science, Services and Agents on the World Wide Web, 7(3), 154-165.

[9] Blomqvist, E., Gangemi, A., Hammar, K., and Suárez-Figueroa, M.C.(Eds.) 2012. Proceedings of the 3rd Workshop on Ontology Patterns (11th International Semantic Web Conference 2012 (ISWC 2012), Boston, USA, November 12, 2012.)

[10] Dodds, L., and Davis, I. 2012. Linked Data Patterns, A pattern catalogue for modelling, publishing, and consuming Linked Data, 2012-05-31, Retrieved June 6, 2013, from: http://patterns.dataincubator.org/book/

[11] Powell, A., Nilsson, M., Naeve, A., Johnston, P., and Baker, T. 2007. DCMI Abstract Model. DCMI Recommendation, 2007.

GALATEAS D2W: A Multi-lingual Disambiguation to Wikipedia Web Service

Deirdre Lungley
CIMeC, University of Trento
Rovereto, Italy

Marco Trevisan
CELI S.R.L.
Torino, Italy

Vien Nguyen
University of Lugano
Italy

Maha Althobaiti
CSEE, University of Essex
Colchester, ESSEX, U.K.

Massimo Poesio
CIMeC, University of Trento
Rovereto, Italy

ABSTRACT

The motivation for entity extraction within a digital cultural collection is the enrichment potential of such a tool – useful in this context for such tasks as metadata generation and query log analysis. The use of Disambiguation to Wikipedia as our particular entity extraction tool is motivated by its generalisable nature and its suitability to noisy text. The particular methodolgy we use does not avail of specific natural language tools and therefore can be applied to other languages with minimal adaptation. This has allowed us to develop a multi-lingual Disambiguation to Wikipedia tool which we have deployed as a web service for the use of the community.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms

Algorithms,Languages, Experimentation

Keywords

Disambiguation to Wikipedia, Entity recognition

1. INTRODUCTION

Information Retrieval within the digital cultural heritage context must contend with often inately "noisy" resources: poor spelling and punctuation, obsolete word forms and abbreviated forms. This can be the case in both text-based resources and in the metadata of image-based resources. This provides an often unsurmountable challenge to traditional natural language processing techniques, e.g., traditional Named Entity Recognition (NER). However, the enrichment potential, within this context, of such a technique, e.g., metadata generation, makes it a challenge worth pursuing.

Most work on NER has focused on (a subset of) the ACE entity types [3], and in particular on **PER** (person), **LOC** (location), **ORG** (organisation). These types are very important for news but not as central in other domains: for example, in a digital art collection, types such as **WORK-OF-ART** are as important, as are finer-grained specialisations of the standard types such as **ARTIST** or **photographer**. One solution to the problem involves developing techniques for rapid domain adaptation. But such techniques are of limited use for a web service as adaptation to a domain is impossible.

The approach we adopted was to develop a service for **Disambiguation to Wikipedia**: linking text to the appropriate Wikipedia page, from which the appropriate type can then be extracted [5, 2, 6, 8]. Figure 1 shows an example of D2W. Given a text "John McCarthy, 'great man' of computer science, wins major award.", a D2W system is expected to detect the text segment "John McCarthy" and link to the correct Wikipedia page *John_Mc Carthy_(computer_scientist)*[1], instead of other *John McCarthys*, e.g., the ambassador, the senator or the linguist.

This D2W approach offers substantial coverage of entities of most domains and in multiple languages. It also allows us to contend with the noisy text environment of digital cultural collections – Wikipedia, a large encyclopedic collection from the web, allows us to extract structured knowledge which is also noisy. Our work builds on previous work in this area [4, 6], the novel aspects being our evaluation in a noisy environment and the multi-lingual adaptations.

The following section summarises previous work in this area. Section 3 details the methodology we employed to extract statistics from the structural information of Wikipedia and how these statistics are used for disambiguation. Section 4 details the steps taken to create a multi-lingual web service. We conclude with Section 5 which details an evaluation of our methodology and the future directions of this work.

[1] We use the title *John_McCarthy_(computer_scientist)* to refer to the full address *http://en.wikipedia.org/wiki/John_Mc Carthy_(computer_scientist)*.

2. PREVIOUS LITERATURE

A method for named entity disambiguation based on Wikipedia was presented in [2]. They first extract resources and context for each entity from the entire Wikipedia collection, then use NER in combination with other heuristics to identify named entity boundaries in articles. Finally, they employ a vector space model which includes context and categories for each entity for the disambiguation process. The approach works very well with high disambiguation accuracy. However, their use of many heuristics and NER means the method is difficult to adapt to other languages as well as to other content types such as noisy text.

A general approach for D2W is proposed by [6]. First, they process the entire Wikipedia and collect a set of incoming/outgoing links for each page. They employ a statistical method for detecting links by gathering all *n-grams* in the document and retaining those whose probability exceeds a threshold. For entity disambiguation they use machine learning with a few features, such as the commonness of a surface form, its relative relatedness in the surrounding context and the balance of these two features. Our methodology builds on this approach, adapting it for a multi-lingual environment and evaluating it on noisier text.

Local and global features are combined in an approach to the D2W task in [8]. They implement their approach using traditional *bag-of-words* and *TF-IDF* measures to calculate semantic relatedness. However, their use of many natural language specific tools, e.g., NER, chunking and part-of-speech tagging makes their method difficult to adapt to noisy text and to other languages.

Previous approaches to *D2W* differ with respect to the following aspects: 1. the corpora they address; 2. the type of the text expression they target to link; 3. the way they define and use the disambiguation context for each entity. For instance, some methods focus on linking only named entities, such as [1, 2]. The method of [2] defines the disambiguation context by using some heuristics such as entities mentioned in the first paragraph and those for which the corresponding pages refer back to the target entity. [6] utilise entities which have no ambiguous names as local context and also to compute semantic relatedness. A different method is observed in [8] where they first train a local disambiguation system and then use the prediction score of that as disambiguation context.

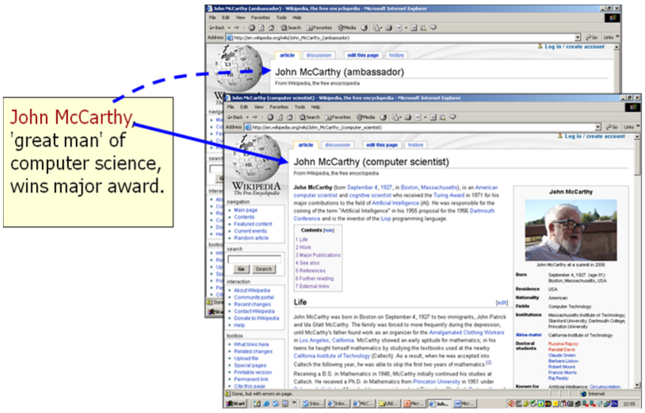

John McCarthy, 'great man' of computer science, wins major award.
John McCarthy (ambassador)
John McCarthy (computer scientist)
WIKIPEDIA
John McCarthy

Figure 1: Disambiguation to Wikipedia

3. EXTRACTING WIKIPEDIA STRUCTURAL INFORMATION

Our approach involves two steps: first extract a number of statistical measurements from a Wikipedia dump, and then leverage these metrics in the disambiguation phase.

3.1 Parsing Wikipedia Dump

The wikipedia dump used is pages-articles.xml, published in July 2011, which contains all current edition Wikipedia articles, templates, media/file descriptions and primary metapages. Wikipedia parsing enables us to build necessary dictionaries and structures. All category pages, disambiguation pages, help pages, 'list_of' pages and pages referring to templates and Wikipedia itself are excluded from the parsing process. Consequently, only the most relevant pages with textual content are utilised.

In the first parse, the system scans Wikipedia articles and constructs the *redirection pairs* set (i.e., one article contains a redirect link to the actual article for that entity), and a list of all Wikipedia article titles. An example of a redirection pair is *Leonardo_da_Vinci* and *Da_Vinci*. *Leonardo_da_Vinci* is the full name of Da Vinci. Thus, the article entitled *Da_Vinci* just points to the actual article with the title *Leonardo_da_Vinci* [7].

The second parse builds a list of links, i.e., *surface form*, *target article* for each link, and the number of times one surface form is linked to the target article (incoming and outgoing links for each article are computed). We use the term 'surface form' to indicate the mention of an entity that already has a corresponding Wikipedia article and the term 'target article' to indicate the Wikipedia article that the surface form is linked to. Within the set of links, the redirected article related to the link is changed to the actual article by using the redirection pairs collected in the first parse. For example, if the title *Da_Vinci* appears in one link, we will change it to *Leonardo_da_Vinci*.

The third parse involves parsing individual Wikipedia pages to construct the: ID, title, set of categories, set of templates and set of links for each article.

Furthermore, the set of Wikipedia titles and surface forms are preprocessed byway of case-folding. So, at the end of this parsing phase, we have obtained a set of dictionaries and structures. We use the dictionaries of titles, surface forms and files (e.g, *File:Mona Lisa, by Leonardo da Vinci, from C2RMF retouched.jpg*) to match the textual content and detect entity boundaries. The set of links is used to compute statistical measures, necessary for the disambiguation phase.

3.2 Computing Statistical Measures

Links embedded in Wikipedia articles provide millions of manually defined examples to learn from, i.e., for every surface form in an article a Wikipedia editor has manually selected the correct target article that represents the meaning of the anchored text. We derive the following statistical measures from the dictionaries and structures, which we have extracted from Wikipedia.

- Keyphraseness: This is the probability that a word or a phrase will link to a Wikipedia article. Thus, to identify important words and phrases, we follow the methodology of [4] in which all word n-grams are extracted and the probability of each n-gram is computed, as follows:

$$keyphraseness(s) = \frac{count(s_{link})}{count(s)}$$

Here, s is a surface form (anchor text), $count(s_{link})$ is the number of Wikipedia articles in which s appears as a link, and *count(s)* is the number of Wikipedia articles in which s appears.

- Commonness: This is the probability of a specific Wikipedia article t to function as the target of a link with a word or a phrase s.

$$Commonness(s,t) = \frac{P(t|s)}{P(s)}$$

Here, $P(t|s)$ is the number of times s appears as a link to t, and *P(s)* is the number of times s appears as a link.

- Relatedness: This metric allows us to measure the semantic similarity of two terms [9]. We consider each term a representative Wikipedia article. For example, the term *wood* is represented by the Wikipedia page *http://en.wikipedia.org/wiki/Wood.* Pages that link to both terms suggest relatedness.

$$Relatedness(a,b) = \frac{log(max(|A|,|B|)-log(|A \cap B|))}{log(|W|)-log(min(|A|,|B|))}$$

Where a and b are the two articles of interest, A and B are the sets of all articles that link to a and b respectively. W is the entire Wikipedia.

3.3 Disambiguation Method

Our disambiguation method is made up of two steps. First, all surface forms with their potential candidates (target articles) are detected in given documents. Next, a scoring method is used to select a candidate sense, either our baseline or a machine learning method.

The identification of candidates follows the methodology employed in [4], where all word n-grams are extracted and the keyphraseness of each n-gram is computed. The keyphraseness determines the probability that each n-gram word will be a candidate to link to a Wikipedia article. Only n-grams whose keyphraseness exceed a certain threshold remain. Following preliminary experiments, we found that the best mention detection performance is accomplished with keyphraseness = 0.01.

In the next step, we employ commonness in order to identify potential candidate links. For each surface form we use the top 10 candidates (target articles) with the highest commonness. The entity disambiguation problem can be converted to a ranking problem, i.e., the link corresponding to a surface form s is defined as the one with highest score:

$$\hat{t} = \arg\max_{t_i} score(s, t_i)$$

In this formula, $score(s, t_i)$ is an appropriate scoring function.

As a baseline we use solely commonness, which is the fraction of times the title t is the target page for a surface form s. This single feature is a very reliable indicator of the correct disambiguation [8].

3.3.1 Machine Learning Method

This method uses three types of features: the commonness of each candidate link, its average relatedness in the surrounding context of the current document, and a feature which balances these two statistical measures.

The relatedness of each candidate sense is the weighted average of its relatedness to each context article as proposed by [6]:

$$score(s,t) = \frac{\sum_{c \epsilon C} relatedness(t,c)}{|C|} \times commonness(s,t)$$

where $c \epsilon C$ are the context articles of t. Only unambiguous context articles (surface forms) are considered, i.e., those that have only one related Wikipedia article. Their unambiguous nature makes them more helpful in defining the particular context.

The final feature balances commonness and relatedness, by summing the weights that were previously assigned for each unambiguous surface form, as proposed by [6].

4. BUILDING A MULTI-LINGUAL D2W WEB SERVICE

Our baseline approach has been adopted, with a few modifications, to produce seven wikifiers in seven different languages: English, Italian, French, German, Dutch, Polish and Arabic. To build each language model we acquired the language-specific Wikipedia dump and list of 'stop words'. Multi-lingual mappings of Wikipedia internal link keywords, e.g., 'category', 'help', 'file', 'disambiguation', 'template' and 'list of' were also required by the parser. UTF-8 encoding allowed for the representation of the broad range of characters required.

4.1 Web Service Technology

The D2W system has been deployed as a Web service into Linguagrid[2], a platform for the controlled distribution of NLP Web services. Software as a Service (SaaS), such as Web services, are attractive to the system integrator because they delegate the burden of the optimisation of hardware resources to the service provider. For the system provider, SaaS is also attractive since it reduces the cost of delivering and maintaining the system for multiple users and also allows him to have finer control over how the system is used. Usage statistics can also be collected and analysed.

The D2W system has been adapted and optimised to ensure it integrated effortlessly and performed efficiently as a Web service in Linguagrid. The Web service framework we used to encapsulate the D2W system into a Web service is Apache CXF. The D2W Web service exposes an API (WSDL schema) that uses data structures based on the Morphosyntactic Annotation Framework (MAF) parts, an ISO standard (ISO/DIS 24611)[3]. This WSDL schema has been developed in the context of the EUROPEANA[4] project. According to this schema, the information returned by the D2W web service consists of a set of annotations, each annotation associating the URI of a Wikipedia article to a substring of the input text. This URI can easily be used

[2] http://www.linguagrid.org

[3] http://www.iso.org/iso/catalogue_detail.htm?csnumber=51934

[4] http://www.europeana.eu

within client software to retrieve linked data from ontologies and thesauri.

In order to reduce the memory requirements of GATE[5], the D2W Web service relies on a customised GATE gazetteer that reads data from a database instead of from memory. This customisation has allowed us to have a single system supporting multiple languages, at the cost of reducing the processing speed. In the context of the GALATEAS[6] project, the D2W system has been used to extract named entities from 1.5M short text queries, each 14 characters long on average (total 21M characters). The system achieved a throughput of almost 600 characters per second running on a dedicated multicore server, using a limited amount of main memory (400 MB).

5. RESULTS AND DISCUSSION

The evaluation of our D2W methodology, detailed in [7], involved 1000 queries from the Bridgeman Art Library (BAL)[7] – a noisy dataset containing spelling errors, malformed sentences, etc.. The annotators were asked to link the first five nominal mentions of each co-reference chain to Wikipedia. Table 1 details these results to highlight the potential of this methodology. The first five rows report the results for our baseline method, according to the number of disambiguation candidates generated. The last row shows the results for the machine learning method.

No. of Candidates	**Recognised mention(s)**	**F-measure**
Candidate 1	853	64.77
Candidate 2	1022	71.59
Candidate 3	1092	75.42
Candidate 4	1134	77.18
Candidate 5	1157	78.32
All features	n/a	69.32

Table 1: Results with 1000 BAL queries

A second evaluation, also detailed in [7], details the results obtained with standard datasets: ACQUAINT – a subset of this newswire text corpus which is annotated to mimic the hyperlink structure in Wikipedia, and a dataset constructed from 10,000 paragraphs from Wikipedia itself. These results, detailed in Table 2 prove the potential of the machine learnt method. Milne-Witten (2008) refers to the results reported in [6], while Ratinov-Roth refers (2011) to those in [8].

System	**ACQUAINT**	**Wikipedia**
Our M/L D2W	86.16	84.37
Milne-Witten (2008)	83.61	80.31
Ratinov-Roth (2011)	84.52	90.20

Table 2: Results with standard datasets

The novelty of our work lies in our evaluation on a noisy dataset and the conversion of this methodology to a Web service, available as a resource to the digital cultural heritage community [8]. Although the multi-lingual version of the D2W Web service has been tested and provided similar results as the original English version, a thorough evaluation remains for future work.

6. ACKNOWLEDGMENTS

This work has been supported by the GALATEAS project (http://www.galateas.eu/ – CTP-ICT PSP-2009-3-25430) funded by the European Union under the ICT PSP program.

7. REFERENCES

[1] R. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Procesings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, Trento, Italy, 2006.

[2] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[3] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassell, and R. Weischedel. The automatic content extraction (ACE) program–tasks, data, and evaluation. In *Proc. of LREC*, 2000.

[4] O. Medelyan, I. H. Witten, and D. Milne. Topic indexing with wikipedia. In *Proceedings of the AAAI WikiAI workshop*, 2008.

[5] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pages 233–242. ACM, 2007.

[6] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 509–518. ACM, 2008.

[7] T.-V. T. Nguyen and M. Poesio. Entity disambiguation and linking over queries using encyclopedic knowledge. In *Proceedings of the 6th workshop on Analytics for Noisy Unstructured Text Data*, AND '12, December 2012.

[8] L. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384. Association for Computational Linguistics, 2011.

[9] I. Witten and D. Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30, 2008.

[5] http://gate.ac.uk/

[6] http://www.galateas.eu

[7] http://www.bridgemanart.com/

[8] http://ws.linguagrid.org/d2w

Generating Automatic Keywords for Conversational Speech ASR Transcripts

Hohyon Ryu
Twitter, Inc.
San Francisco, CA 94103
hohyonryu@twitter.com

Matthew Lease
School of Information
University of Texas at Austin
ml@ischool.utexas.edu

ABSTRACT

While a plethora of *conversational speech* has been recorded and archived for over a century, it has not been easily accessible due to many technical challenges vs. text and *rehearsed speech* to be addressed before conversational archives can be effectively searched and used. In this paper, we describe two language modeling methods for automatically assigning keywords to automatic speech recognition (ASR) transcripts, to benefit search and browsing of conversational speech archives. Experiments performed with the English CLEF CL-SR MALACH collection of oral history interviews. In comparison to a prior baseline generating 20 keywords per conversation segment, we use 1/20th the training data yet improve Recall@20 in matching manual keywords. However, while indexing of manual keywords yields improved search accuracy, indexing automatic keywords (ours or the baseline) fails to improve search accuracy, evidencing the need for additional research.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*

General Terms

Algorithms, Experimentation, Performance

Keywords

Spontaneous Conversational Speech, Language Modeling, Nearest Neighbor Classifier

1. INTRODUCTION

While *spoken document retrieval* was claimed to be a solved problem [3] in TREC over a decade ago, the data considered was only clean *prepared speech*, aka "rehearsed" or "read" speech, such as broadcast news and political speeches. In contrast, *spontaneous conversational speech* (SCS) (e.g., phone calls or voicemail, meetings, classroom discussion, talk shows, interviews, cocktail parties, etc.), which has been widely collected and archived for over a century, has mostly remained in its raw form posing many challenges for effective information retrieval (IR) today [11, 16, 10, 13].

With SCS, automatic speech recognition (ASR) performs much worse than with prepared speech due to many factors: wider speaker variation, non-native speakers who may "code-switch" between languages, emotional speech, noisy environments, and lower quality microphones [5, 4]. Speech may also use specialized vocabulary not part of common discourse and ASR vocabularies. In addition, speech exhibits cognitive processing effects and speaking errors, including mumbles, partial-words, filler words, repetitions, and corrections. The following ASR transcript provides an example of ASR quality in the MALACH oral history collection [11] (excerpted from VHF34774-159541.027, see Section 3.1):

> do you recall the first time you were beaten yes forty forty five what was that did you what it was not the the the the the cement or do you it was very hard for me was at home uh one of the people live like that you know and that was not far from the city where you couldn't you was punished at all is that the that the from the i could go back to the were couldn't buy any friends as a uhhuh polish and what did you so when you get the uhhuh lucky superficial watches who did such a you and if you could have uhhuh and what and that was what was your what and that was all in and polish jews remember his name in the face of all slammed hide the fact that you...

When relevance judgments were being collected for the English MALACH IR test collection, ASR transcripts were not yet available. Instead, judging relied on manual summaries and keywords, with occasional reference to the audio. Metadata for this same conversation segment is show below:

> Auschwitz II-Birkenau (Poland : Death Camp) | Poland 1944 | kapos, Jewish | kapos, Polish | brutal treatment in the camps | beatings

For the search topic "Birkenau daily life", "Birkenau" never appears in the ASR text but does appear in the manual keywords, and this segment was indeed judged as relevant to this search topic. How is a user or search system to recognize this segment's relevance without labor-intensive manual curation to produce such keywords? To address this, prior work has investigated automatic keyword generation [16,

ACM SigIR '2013 Dublin, Ireland

12]. We further explore this idea, building on our prior inferring a text's place and time from implicit lexical cues rather than detecting explicit places or dates [8, 17, 14].

From 2005-2007, the Cross Language Evaluation Forum (CLEF) held a Cross-Language Speech Retrieval (CL-SR) Track [16, 10, 13]. The CL-SR track used part of the Survivors of the Shoah Visual History Foundation oral history archive, and the retrieval tasks were conducted on the ASR text, manual summary, manual keywords, and automatic keywords. We study the CL-SR'07 test collection here.

In [2] and [15], a machine learning approach was proposed for keyword extraction. [1] applied a keyword extraction algorithm designed for written text and showed ASR quality is crucial to keyword extraction performance. [6] extracted keywords using supervised machine learning and linguistic knowledge. [7] further refined [1]'s method taking into account the semantic meanings of the keywords. [9] extracted keywords from a conversational meeting corpus using supervised machine learning approach and bigram expansion.

The most relevant prior work [16, 12] to ours inferred keywords for the same MALACH collection. [12] used the previous segment to provide additional context for the next segment. However, whereas their data-intensive approached utilized 168,584 segments of private training set, we attempt this task using only the roughly 8,000 conversation segments publicly available in the CL-SR'07 collection.

2. METHODOLOGY

Our task is to automatically select the best keywords, from a pre-defined set, to assign to each conversational segment. We describe two approaches to this task below. In both, a language model (LM) estimates a probability of document d being relevant to query q, or $p(d|q)$. Using Bayes Rule, we derive the usual query-likelihood IR formulation as:

$$p(d|q) = \frac{p(q|d)p(d)}{p(q)}$$

$p(q)$ can be ignored since it is constant across documents. $p(d)$ is the document prior. $p(q|d)$ is query likelihood, i.e. the probability of generating query q for document d.

2.1 Pseudo-Document (PD) Language Model

In our first approach, we construct a pseudo-document for each keyword by aggregating all the segments to which the given keyword has been manually assigned. In this way, we create a collection of psuedo-documents, one per keyword, which can be searched. Each conversation segment represents a query, and the top-K ranked pseudo-documents correspond to the most likely K keywords that should be assigned, where K is a parameter. The LM is redefined as $p(k|s)$ for keyword k and conversation segment s via Bayes:

$$p(k|s) = \frac{p(s|k)p(k)}{p(s)}$$

where denominator $p(s)$ is again constant and can be ignored, $p(k)$ defines a keyword prior, and $p(s|k)$ is the likelihood of the segment given a particular keyword.

We investigate expanding the query segment with similar segments to provide contextual information (CI). For segment s with word vector $\vec{w}$, neighbor segment s' is added to s with weight inversely-proportional to its distance from s.

$$\delta = |i_{s'} - i_s|, w'_j = f(d; \mu, \sigma^2) \times tf_{w'}$$

where δ is the distance between s and s', i is the sequential segment index, $tf_{w'}$ is the term frequency of a word w' in the segment s', and f is a Gaussian probability density function:

$$f(x;\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/(2\sigma^2)}$$

As in prior work [16, 12], we also try assigning keywords with two separate models (TM): a temporal-geolocation model and a more general concept model. Each model is trained by partitioning the manual keywords into these two categories. For example, 'Germany 1943' vs. 'literature and writing'.

2.2 Segment Language Model (kNN)

Prior work [16, 12] used k-Nearest neighbor (kNN) approach to find similar segments and their keywords. We explore a similar LM approach here, using the query segment to find similar segments (from other interviews).

$$p(s'|s) = \frac{p(s|s')p(s')}{p(s)}$$

We then aggregate manual keywords assigned to those similar segments in order to select keywords to assign to the query segment. Given the top ranked k most similar segments, each manual keywords assigned to the ith retrieved segment s_i is receives weight $w_{kwd} = k - i + 1$. The n keywords with highest aggregated weight are assigned to s.

As with the earlier psuedo-document approach, the query segment can be expanded with similar segments prior to matching, akin to traditional IR psuedo-relevance feedback.

3. EXPERIMENTS

Search to infer keywords is performed using Galago[1]. With the two model (TM) approach, separate searches are used for temporal-geolocation vs. general concept keywords. 16 concept keywords and 4 temporal-geolocation keywords are assigned to each segment. We evaluate in two ways: automatic vs. manual keywords, and change in CLEF CL-SR search accuracy when we add keyword indexing.

3.1 Test Collection

Field	Description
DOCNO	the interview id + the segment id
INTERVIEWDATA	the names of interviewees
NAME	the names of people mentioned in the segment
ASRTEXT2003 ASRTEXT2004 ASRTEXT2006A ASRTEXT2006B (ASR)	4 versions of ASR texts
SUMMARY (SUM)	manual summary
MANUALKEYWORD	manual keyword
AUTOKEYWORD2004A1 AUTOKEYWORD2004A2	2 versions of automatic keywords

Table 1: CLEF CL-SR Interview Segment Fields.

The CLEF 2007 Cross-Language Speech Retrieval collection (CL-SR) [13] is used in this experiment. CL-SR consists of 272 interviews with Holocaust survivors, witnesses, and rescuers (589 hours of speech) [16]. These interviews are divided into 8,104 segments, including four versions of ASR

[1] www.galagosearch.org

transcripts (we use the best only, ASRTEXT2006B), manual keywords, and two versions of automatic keywords. Table 1 shows the metadata fields for each segment. We exclude two interviews (15 segments) which are missing ASR texts. Short, blank, or corrupted ASRs are also filtered out, leaving 7902 segments.

We observe that 5.6 manual keywords are assigned on average, using 3605 unique keywords. Of the two sets of automatic keywords, we adopt AUTOKEYWORD2004A2 as baseline since it better matches manual keywords. Quality of baseline keywords are shown in Table 2. As the baseline methods assign 20 keywords per segment, we report Recall at 20 keywords (R@20) for comparable evaluation.

Keywords	Precision	Recall	F-Score
AUTOKEYWORD2004A1	0.076	0.289	0.116
AUTOKEYWORD2004A2	0.090	0.326	0.136

Table 2: CL-SR baseline automatic keyword quality.

3.2 Methods

Pseudo-documents. Keywords are assigned to the segments using 10-fold cross-validation. We retrieve the 20 most similar keyword pseudo-documents for each query segment s. The query segment s is simply the segment's ASR transcript. For each keyword pseudo-document k, we not only concatenate the ASR transcripts for the segments to which it is assigned, but we also follow prior work [16, 12] in "cheating" by including manual summaries as well. This allows fair comparison but will ultimately be abandoned in future work. Nevertheless, the task remains quite difficult.

Similar segments. Using Galago, similar segments s' are retrieved for query segment s. Segments s' in the Galago search index includes both ASR text and the manual summary, whereas the query segment includes ASR text only. Experiments vary the number of similar segments used.

3.3 Searching MALACH

Whereas our first evaluation method assesses the system's ability to match manual keywords, our second evaluation measures the benefit of automatic keywords for improving search accuracy. We use 105 CL-SR topics, where queries include all topic fields: title, description, and narrative (e.g., see Table 3). Gold relevance judgments are binary [13]. Stopwords (ST) are removed based on the Indri[2] stop word list, augmented to exclude conversational filled pauses and backchannels such as: "um", "yeah", "uhhuh", and "wow".

Search uses ElasticSearch[3] 0.19.9, based on Lucene[4] 3.5. Retrieval performance, measured by Mean Average Precision (MAP), is compared with keywords we inferred is compared to use of the CL-SR manual or automatic keywords.

4. RESULTS

4.1 Matching Manual Keywords

Table 4 shows results. Critically, note that that BASELINE used 168,584 segments and kNN used about 7,000 segments for training [16, 12], whereas we use only a few thousand segments. We see the kNN approach far outperforms the pseudo-document approach. Recall ST denotes stopwords filtering, CI is context information, TM is the two model approach, and k# denotes the value of parameter k of kNN (Section 2.2). Adding context information led to detrimental query drift and significantly decreased the keyword matching performance. TM had no significant impact.

Topic	Varian Fry
Description	The story of Varian Fry and the Emergency Rescue Committee who saved thousands in Marseille
Narration	Varian Fry, a young American journalist, created an underground operation that smuggled more than 2,000 refugees (including Marc Chagall, Max Ernst, and Andre Breton) out of Vichy France in 1940-1941. The relevant material should contain information about this operation. Any first-hand information of people who have been rescued by Fry is highly relevant

Table 3: A example of the topics.

Experiment	R@20
BASELINE	0.334
PD	0.047
PD(ST)	0.084
PD(ST+CI)	0.029
PD(ST+TM)	0.085
kNN(k10)	0.235
kNN(k10+ST)	0.276
kNN(k200+ST)	**0.369**
kNN(k200+ST+CI)	0.32

Table 4: Keyword matching performance measured for Recall at 20. Only stopword filtered kNN approach with a large k outperformed the baseline.

Figure 1 shows the impact of k of kNN. kNN with stopword filtering (ST) outperforms the baseline at 40.

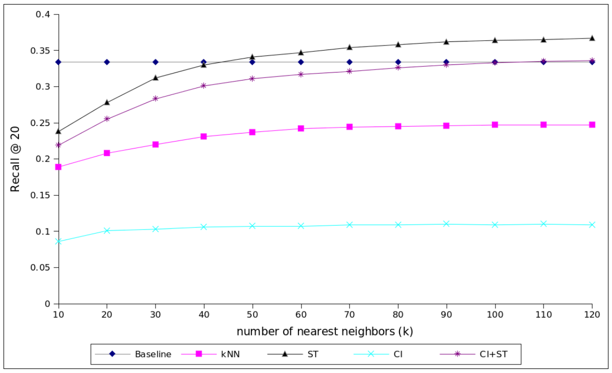


Figure 1: The IR Test Performance of the MANUALKEYWORD and the baselines (AUTOKEYWORDS2004A1 and AUTOKEYWORDS2004A2).

4.2 Using Automatic Keywords in Search

Figure 2 shows the impact of adding keywords for IR performance. We index SUMMARY, ASRTEXT2006B and an additional keyword field that is varied: MANUALKEYWORD, AUTOKEYWORD2004A2 (baseline), and finally

[2] www.lemurproject.org/indri

[3] www.elasticsearch.org

[4] lucene.apache.org

our kNN(k200+ST) generated keywords. We vary the number of keywords added from kNN(k200+ST) from 5 to 20.

Manual keywords improve the IR performance significantly while the automatic keywords have almost no impact. The baseline AUTOKEYWORD2004A2 has zero or negative impact in comparison to the NO KEYWORDS condition. The impact of kNN(k200+ST) on IR performance is negligible.

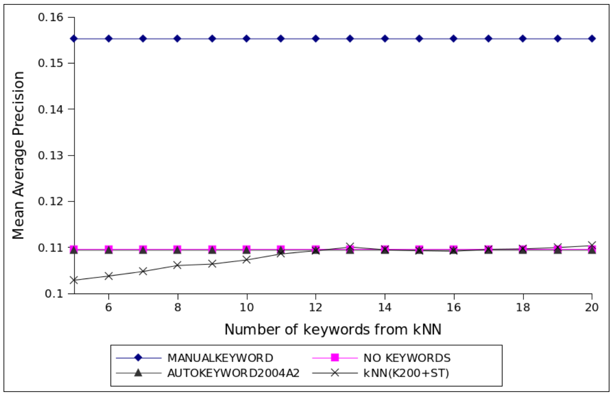


Figure 2: The IR Test Performance of SUMMARY and ASRTEXT2006B for CLEF 2007 topics, descriptions, and narrations when MANUALKEYWORD, AUTOKEYWORDS2004A2, kNN(k200+ST), and no keywords are added.

5. DISCUSSION

We compared two language modeling approaches to improve SCS retrieval: generating a psuedo-document for each keyword, and assigning keywords from similar segments. Inferred keywords were also evaluated in terms of change in IR performance using CLEF 2007 CL-SR track IR tasks. For matching manual keywords, we were able to improve Recall@20 despite training on roughly 20x fewer segments than used in prior work. Nevertheless, neither our generated keywords, nor the baseline automatic keywords, led to improved IR accuracy when indexed.

Presently we see that in the majority of the cases, automatic keywords introduce more incorrect than correct keywords. Out of 20 keywords, only about 10% of the automatic keywords are correct, leaving 90% of keywords to confuse the search engine. This is familiar to traditional NLP approaches in that while latent representations offer opportunities to enrich observed terms, errors inferring latent structures can cause more harm than good. Thus, instead of measuring recall, we should really be focusing on improving the precision of keyword extraction to benefit IR accuracy.

As this is a work-in-progress, we have a variety of ideas for refining the modeling approaches from here, and we are also looking into whether the additional training segments used in prior work [16, 12] might be obtainable.

6. REFERENCES

[1] A. Désilets, B. D. Bruijn, and J. Martin. Extracting Keyphrases from Spoken Audio Documents. In A. Coden, E. Brown, and S. Srinivasan, editors, *Information Retrieval Techniques for Speech Applications*, volume 2273 of *Lecture Notes in Computer Science*, pages 36–50. Springer Berlin Heidelberg, 2002.

[2] E. Frank, G. Paynter, and I. Witten. Domain-specific keyphrase extraction. In *6th International Joint Conference on Artificial Intelligence*, 1999.

[3] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees. The TREC spoken document retrieval track: A success story. In *Text Retrieval Conference (TREC) 8*, pages 16–19, 2000.

[4] W. Ghai and M. G. College. Literature Review on Automatic Speech Recognition. *International Journal of Computer Applications*, 41(8):42–50, 2012.

[5] B. Gold, N. Morgan, and D. Ellis. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. Wiley, 2011.

[6] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *The 2003 conference on Empirical methods in natural language processing*, 2003.

[7] D. Inkpen and A. Désilets. Extracting semantically-coherent keyphrases from speech. *Canadian Acoustics*, pages 2–3, 2004.

[8] A. Kumar, M. Lease, and J. Baldridge. Supervised language modeling for temporal resolution of texts. In *Proceeding of the 20th ACM Conference on Information and Knowledge Management (CIKM)*, pages 2069–2072, 2011.

[9] F. Liu and Y. Liu. Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. *Spoken Language Technology Workshop*, pages 181–184, 2008.

[10] D. Oard, J. Wang, G. Jones, and R. White. Overview of the CLEF-2006 cross-language speech retrieval track. In *CLEF 2006*, pages 744–758, 2007.

[11] D. W. Oard, D. Doermann, G. C. Murray, J. Wang, M. Franz, and S. Gustman. Building an Information Retrieval Test Collection for Spontaneous Conversational Speech Categories and Subject Descriptors. In *27th ACM-SIGIR*, pages 41–48, 2004.

[12] J. S. Olsson and D. W. D. Oard. Improving text classification for oral history archives with temporal domain knowledge. In *Proceedings of the SIGIR*, 2007.

[13] P. Pecina, P. Hoffmannova, and G. Jones. Overview of the CLEF-2007 cross-language speech retrieval track. In *Advances in Multilingual and Multimodial Information Retrieval*, pages 1–14. 2008.

[14] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge. Supervised Text-based Geolocation Using Language Models on an Adaptive Grid. *Proc. EMNLP-CoNLL*, pages 1500–1510, 2012.

[15] P. D. Turney. Learning Algorithms for Keyphrase Extraction. *Information Retrieval*, 2(3):303–336, 2000.

[16] R. W. White, D. W. Oard, G. J. Jones, D. Soergel, and X. Huang. Overview of the clef-2005 cross-language speech retrieval track. pages 744–759. Springer, 2006.

[17] B. Wing and J. Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of ACL*, pages 955–964, 2011.

The CULTURA Project:
Supporting Next Generation Interaction with Digital Cultural Heritage Collections

Gary Munnelly[1], Cormac Hampson[1], Séamus Lawless[1], Maristella Agosti[2], Owen Conlan[1]
[1] Trinity College Dublin, Ireland
[2] University of Padua, Italy
munnellg@tcd.ie,cormac.hampson@scss.tcd.ie,seamus.lawless@scss.tcd.ie,
agosti@dei.unipd.it,owen.conlan@scss.tcd.ie

ABSTRACT

This demonstration will present CULTURA [1], a dynamic, customizable web portal which provides a suite of tools designed to empower and assist a variety of users in their exploration of a number of cultural heritage collections.

CULTURA is a three year, FP7-funded project, whose main objective is to pioneer the development of personalized information retrieval and presentation, contextual adaptivity, and social analytics, all in a digital humanities context. A wide array of tools and services are employed to aid and inform the user in their exploration of digital collections.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – Information filtering, query formulation, relevance feedback, retrieval models, search process, selection process. H.3.5 [**Information Storage and Retrieval**]: Online Information Services - Web-based services. K.3.1 [**Computers and Education**]: Computer Uses in Education – Collaborative learning, computer-assisted instruction (CAI), computer-managed instruction (CMI).

General Terms

Algorithms, Management, Measurement, Design, Reliability, Experimentation, Human Factors.

Keywords

Personalisation, adaptive environments, cultural heritage, digital collections, user modelling, assisted learning.

1. CULTURA PORTAL AND DEMO

CULTURA currently supports three digital cultural collections: the 1641 Depositions, a manuscript collection of 17th century witness statements from Trinity College Dublin; the Imaginum Patavinae Scientiae Archivum (IPSA), a medieval collection of illuminated manuscripts of herbal and astrological codices from the University of Padua; and Bureau of Military History, an early 20th century collection of witness statements from the Irish Military Archives. Due to their contrasting modalities, these collections present very different challenges.

CULTURA has been developed using a service-oriented architecture. There are both pre-processing and runtime services which act upon the collections. These services include:

- Text Normalisation
- Entity Extraction
- Entity Oriented Search
- Social Network Analysis and Visualisation
- Annotation
- Personalisation.

Different combinations of these services are appropriate and required for different types of content, and the CULTURA architecture allows the toolset provided by the portal to be tailored to suit each particular collection.

This demonstration will emphasise how the end-user experience delivered is facilitated by both data layer and presentation layer services. Before the data of a collection is exposed to the user, it undergoes preprocessing to extract meaningful information which can be used by both the system and potentially the end user.

In an Early Modern English textual collection, such as the 1641 Depositions, the raw data undergoes a process of normalization, which is intended to introduce consistency into the document language by resolving some of the variant spelling into a more modern form.

Textual collections can then be passed in both original and normalized form to the entity extraction service. Entity extraction is used to identify the named entities within the texts including people, places, dates, etc. These entities provide an important insight into the nature of the text and can be used to guide a user towards resources which are relevant to their research as well as link documents which are thematically similar, such as those which mention the same individuals, places or events.

Social network analysis (SNA) can be conducted on the output of the entity extraction process, or on the metadata of an

SIGIR'13, *ENRICH Workshop* August 1, 2013, Dublin, Ireland.
ACM

image-based collection such as IPSA, in order to help identify and visualise the important individuals involved, and the social networks that exist within a collection.

A user model is maintained for each individual who interacts with the CULTURA environment. Information about a user's browsing history, inferred interests and exhibited level of expertise is persistently stored, thus allowing the environment to personalized the user's experience across several sessions. For example, for a user who is exhibiting a particular interest in content relating to County Wicklow, both by bookmarking documents which relate to it and annotating bodies of text which contain references to it, CULTURA will attempt to establish what aspects of Wicklow are of interest to that user by correlating their user model with the entities extracted from those documents. The relationships determined by entity extraction process can then be used to produce lists of alternative sources which may interest the user.

Simple user interface controls in the CULTURA environment allow users to interact directly with the cultural heritage collections through annotating, sharing, bookmarking and searching. Services can also be called to visualize the social networks contained within a single resource or across the entire collection. Through methods such as this a user can explore the vast range of related entities which chain documents together.

User's seeking a more guided, tutorial based experience of a corpus can avail of "narratives" [2] which guide the user through a collection, explaining the content along the way and providing insight into the nature of the sources. A collection of narrative threads, designed by experts in the domain of the source material, can be offered to the user.

2. ACKNOWLEDGMENTS

The work reported has been partially supported by the CULTURA project, as part of the Seventh Framework Programme of the European Commission, Area "Digital Libraries and Digital Preservation" (ICT-2009.4.1), grant agreement no. 269973. We acknowledge the contribution of all project partners involved in CULTURA (http://www.cultura-strep.eu/).

3. REFERENCES

[1] C. Hampson, M. Agosti, N. Orio, E. Bailey, S. Lawless, O. Conlan and V. Wade (2012). The CULTURA Project: Supporting Next Generation Interaction with Digital Cultural Heritage Collections. In: M. Ioannides, D. Fritsch, J. Leissner, R. Davies, F. Remondino, R. Caffo (Eds), *Progress in Cultural Heritage Preservation, 4th International Conference, EuroMed 2012*, Limassol, Cyprus, LNCS Vol. 7616, Springer, Berlin Heidelberg, 2012, pp. 668-675.

[2] O. Conlan, A. Staikopoulos, C. Hampson, S. Lawless and I. O'Keeffe. The Narrative Approach to Personalisation. *New Review of Hypermedia and Multimedia* [In Press].

Personalized Access to Cultural Heritage Spaces (PATHS)

Paul Clough[1], Paula Goodale[1], Mark Stevenson[2] and Mark Hall[2]
[1]Information School, University of Sheffield (UK)
[2]Department of Computer Science, University of Sheffield (UK)
p.d.clough@sheffield.ac.uk

ABSTRACT

The EU-funded PATHS (Personalized Access to Cultural Heritage) project is investigating ways of assisting users with exploring a large collection of cultural heritage material taken from Europeana, the European aggregator for museums, archives, libraries, and galleries. A prototype system has been developed that includes novel functionality for exploring the collection based on Google map-style interfaces, data-driven taxonomies and supporting the manual creation of guided tours or paths and the use of personalized (and non-personalized) recommendations to promote information discovery.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services - Web-based services

General Terms

Experimentation, Human Factors

Keywords

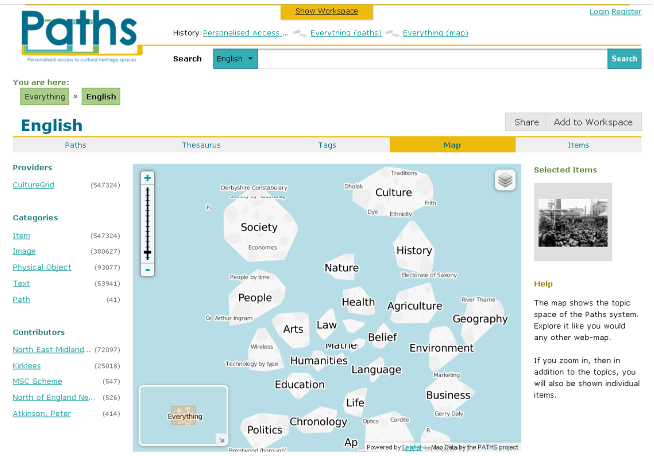
Digital Libraries, Cultural Heritage, Information Access

1. INTRODUCTION

In recent years significant amounts of cultural heritage materials has been made available through online digital library portals, such as Europeana (http://www.europeana.eu), the European aggregator for museums, archives, libraries, and galleries. However, these collections can often be difficult to navigate, especially for those without advanced levels of subject and domain knowledge [1].

The PATHS project [2], which is funded under the FP7 programme of the European Commission, is exploring alternative modes of information access to large cultural heritage collections, such as Europeana. A range of expert and non-expert users from various cultural heritage domains have been involved a user-centred approach to the development of a prototype system [3]. One of the key features of the project has been investigating the design of functionality to support the manual creation of paths or trails through the collection. This has included a workspace feature to store items during exploration of the collection, a path editing feature for arranging the gathered items and forming narratives, and functions for sharing the paths. The resulting paths can be used as a means of navigating items in the collection based on a theme or topic, along with forming tangible learning objects for education purposes. Figure 1 shows an example screenshot from the PATHS system as the user explores the collection using a Google Map-style visualisation.

SIGIR'13, ENRICH Workshop August 1, 2013, Dublin, Ireland. ACM



Providers	
CultureGrid	(547324)
Categories	
Item	(547324)
Image	(380627)
Physical Object	(93077)
Text	(53941)
Path	(41)
Contributors	
North East Midland...	(72097)
Kirklees	(25018)
MSC Scheme	(547)
North of England Ne...	(526)
Atkinson, Peter	(414)

Help

The map shows the topic space of the Paths system. Explore it like you would any other web-map.

If you zoom in, then in addition to the topics, you will also be shown individual items.

Figure 1 – PATHS system showing Google Map-style visualisation for exploring themes in the collection

The PATHS system makes use of state-of-the-art text processing and information retrieval techniques to link similar items within the collection, link to related Wikipedia articles, generate mappings to thesauri and controlled vocabularies to aid navigation, and provide recommendations.

2. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270082. We acknowledge the contribution of all project partners involved in PATHS (http://www.pathsproject.eu).

3. REFERENCES

[1] Skov, M. and Ingwersen, P. (2008). Exploring information seeking behaviour in a digital museum context. In *Proc. 2nd Int. Symposium on Information Interaction in Context* (London, October 14-17, 2008).

[2] Fernie, K. et al. (2012). PATHS: Personalising access to cultural heritage spaces, In *Proc. of 18th Int. Conference on Virtual Systems and Multimedia (VSMM 2012)*, 469-474.

[3] Goodale, P. et al. (2012). User-Centred Design to Support Exploration and Path Creation in Cultural Heritage Collections, In *Proc. of the 2nd European Workshop on Human Computer Interaction and Information Retrieval* (EuroHCIR 2012), 75-78.