Exploring Entity-Centric Methods in the UK Government Web Archive

Philip Webster[†], Paul Clough[†], Gianluca Demartini[†], Tom Storrar[‡], Sonia Ranade[‡], Graham Seaman[‡]
[†]University of Sheffield, Sheffield UK, [‡]UK National Archives, Kew
{philip.webster,p.d.clough,g.demartini}@sheffield.ac.uk
{Tom.Storrar,Sonia.Ranade,Graham.Seaman}@nationalarchives.gsi.gov.uk

ABSTRACT

Being able to explore large digital collections effectively is of interest to both academics and practitioners alike. The need to go beyond the provision of keyword-driven functionality to features that support exploration and discovery is widely recognised. In addition, providers are seeking to support more diverse groups of users with varying information needs and tasks. Increasing amounts of cultural heritage are being stored in web archives that present unique challenges as a form of digital cultural heritage. This paper describes a collaboration between the University of Sheffield and the UK National Archives to investigate entity-based methods for exploring the UK Government Web Archive.

Keywords

Cultural Heritage, Web Archives, Entity-Based Access

1. INTRODUCTION

There is a clear need for cultural heritage institutions (museums, libraries and archives) to provide systems that go beyond keyword-based search and support more diverse information seeking behaviours, such as browsing and exploration [1, 2, 3]. Whitelaw [2] points out that many digital cultural collections are only accessible via keyword search and suggests that users can often feel constrained by the search box as it limits their ability to browse and explore the collection. He calls for the design of more "generous interfaces" that provide richer browsable user experiences and allow the scale and complexity of digital cultural heritage collections to be revealed, especially to non-specialist users. Similar calls are being made to support exploration by enabling navigation, interpretation and use of items in digital collections and aiding users' analytical and sense-making processes more generally [4, 5].

Web archives are an area of digital cultural heritage gaining increasing attention from researchers. One of the key challenges of such collections is the sheer volume of content [6]. In this paper we describe a recently instigated collaborative research project between the UK National Archives[1] and the Information School (University of Sheffield) to investigate the use of entity-based methods for supporting user's exploration the UK Government Web Archives. In particular we are focusing on issues of scaling up approaches for entity extraction and disambiguation. There is a need to assist users with navigating the content of large digital archives and help them to better understand how resources are interconnected over different dimensions, such as time, entities and events. Ultimately the users of the Government Web archive will benefit from improved features for exploration and discovery. Section 2 describes related work; Section 3 provides background to the UK Government Web Archive; and Section 4 describes planned research, including research aims and challenges.

[1]http://www.nationalarchives.gov.uk/

In Proceedings of 1st International Workshop on Accessing Cultural Heritage at Scale (ACHS'16), June 22, 2016, Newark, NJ, USA.

2. RELATED WORK

2.1 Access to digital cultural heritage

Increasingly cultural heritage portals are encouraging user participation by offering people opportunities to interact with content, for example encouraging them to tag resources, making recommendations to other users and personalisation [7]. Users served by providers of cultural heritage commonly can range from expert user groups (e.g., scholars and curators) accessing the content for professional purposes through to novices engaging with materials for leisure purposes and enjoyment [8, 9]. There may also be a wide range of users in between, such as students or hobbyists, accessing cultural heritage to learn and discover [10].

Johnson [3] argues that users attempting to access online cultural heritage resources face at least three challenges: (1) knowing where to look; (2) knowing what to say; and (3) making sense of archival material (e.g., interpreting and forming connections between items). This is particularly pertinent for non-expert users who often lack the knowledge and skills to engage with cultural heritage resources [11, 12]. Common search tasks by users of digital cultural heritage include fact-finding and those of a more exploratory or information gathering nature. Fact-finding and known-item tasks tend to revolve around search, whilst information gathering tasks lend themselves more to browsing and exploration. There is a clear need for cultural heritage institutions to provide systems that go beyond keyword-based search functions [2]; something that is also recognised by the wider search community [13].

2.2 Searching web archives

Increasing amounts of cultural heritage are being stored in web archives. However, similar to cultural heritage more generally, "unlocking the potential of web archives requires tools that support exploration and discovery of captured content" [6] (p. 851). The development of such tools typically requires the use of natural language processing. For example, the ARCOMEM project investigates the extraction and enrichment of entities, topics, opinions and events over time [14].

Providing access to web archives presents a variety of challenges. One issue is coping with scale as typically collections run into billions of documents. In light of this, Lin et al. [6] describe a Big Data architecture suitable for large-scale infrastructure. Another issue is performance. Inverted indexes are considered to be an essential technique for the provision of timely search results [15]. Indexes are widely used to reference archive data held in ARC (ARChive) or WARC[2] (Web ARChive) file formats [16]. Indexes are held in memory for performance reasons [15], but the increase in the amount of archived data makes this increasingly difficult. Thus, research has also investigated techniques for reducing the size of the index, such as de-duplication [17].

Another key challenge is how to provide effective access to web archives, especially as users typically expect a user experience similar to live-web search engines [18]. Search log analysis performed by the Portuguese National Archive shows that web archive users typically have a brief relationship with archival search [19]. A typical session consists of either a fulltext search request or a URL search request. Fulltext search accounts for around two thirds of queries with around 60% of fulltext sessions lasting only 1 minute. Costa et al. [19] also show that 85% of fulltext sessions contain up to 3 queries, with 44% of queries modifying existing queries. The difficulties that users have with Web archives are made clearer still by the results of user studies. Users of Web archives do not often search historical versions of archived Web pages [19]. When archived documents are accessed via a search interface temporal restrictions are infrequently used as users often seek the oldest documents in the archive only rather than discovering how documents have evolved over time [19].

Dealing with particular structural characteristics is another challenge, such as time. For example, Berberich et al. [20] describe approaches for providing text search over temporally versioned document collections, such as web archives. This is achieved through adapting an inverted index to support temporal search. Their 'time-travel text search' approach supports the exploration of digital collections over time by enabling the evolutionary history of document collections, such as Wikipedia or the Web, to be indexed and exposed.

.

2.3 Entity-based information access

Identifying named entities (e.g., people, organisations and locations) has been the focus of research for many years. More recent efforts have attempted to link entities to linked data resources and knowledge bases, such as DBPedia[3], Freebase[4] and YAGO2 [21]. A major challenge is disambiguation: identifying the correct entity among a number of entities with the same name. Many techniques exist for entity extraction, disambiguation and linking them to linked data resources [22][5]. Researchers have investigated entity-based methods for specific domains. For example, relevant to this paper Van Hooland et al. describe the use of entity recognition and disambiguation methods for cultural heritage collections [23]. Named entities are commonly used to analyse and provide access to web archives [14]. However, adapting general methods to specific domains and applications remains an enduring challenge.

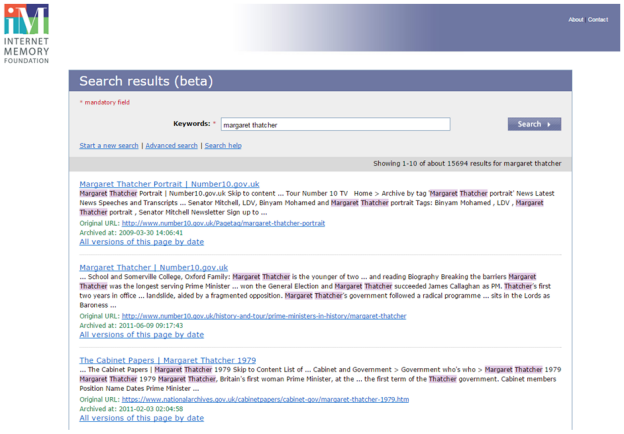

Figure 1: Example results page for the UKGWA (query="Margaret Thatcher")

3. UK GOVERNMENT WEB ARCHIVE

The UK National Archives seeks to collect and secure the future of the public record in all its forms and to make it as accessible as possible. One of the largest digital collections is the UK Government Web Archive[6] (UKGWA). The UKGWA is the preservation and access solution for digital content that is published online by UK Government. It is free to use and is one of the largest and most heavily used web archives in the world receiving approximately 15 million page views per month. The collection is comprised of the contents of over 3,000 websites and social media channels, including 2.5 billion web pages dating from 1996 to the present.

From work undertaken by the National Archives to identify and analyse users of the UKGWA the various user groups include academics (e.g., researchers), National Archives staff, Central Government staff, professional services (e.g., law firms) and the general public. The purposes for using the archive are varied and include family history research by the general public, locating government reports (e.g., by school teachers), finding information about government procedures (e.g., computing tax), viewing the evolution of websites over time, and locating previous versions of current documents.

The current interface to the UKGWA provides traditional keyword-based search functionalities. Figure 1 shows an ex-

[2]http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml
[3]http://wiki.dbpedia.org/
[4]https://www.freebase.com/
[5]See also: http://jhoff.de/wp-content/papercite-data/pdf/hoffart-2015wk.pdf
[6]http://www.nationalarchives.gov.uk/webarchive/

ample of results presented for the query "Margaret Thatcher". The aim of our collaborative research project is to develop interfaces that better support exploration and browsing through the use of named entities extracted from the content. Entities, such as people, places, organisations and events can be extracted from the archive and linked to form a network that users can explore in addition to navigating the content directly.

4. PLANNED RESEARCH

4.1 Research goals

The aim of our research project is to investigate entity-centric methods for supporting users as they navigate the UK Government Web Archive. This would allow users to explore the archive based on entities (e.g., people, locations, events, etc.), as well as allowing connection with existing linked data resources, such as DBPedia and Freebase, and knowledge bases such as YAGO2. The project also provides opportunities to investigate the success of applying alternative methods to domain-specific collections.

The UK National Archives have already explored large-scale entity extraction and linking and this work will utilise existing annotations and ontological resources [24], in addition to exploring new methods and especially focusing on approaches that can be scaled to large collections. This research project consists of two main strands: (1) investigating entity-centric techniques for entity extraction, disambiguation and linking; and (2) investigating how entity-based networks can best support user browsing and exploration. Use cases developed in prior studies that go beyond known-item search tasks will help inform the development of prototype systems (e.g., a user wanting to view the evolution of web pages over time). Originality of the work will include: (1) investigating the effectiveness of various entity-centric techniques for a large digital archive; (2) investigating the preferred way of surfacing the entity network to users; and (3) developing suitable evaluation methodologies to establish the success of entity-centric approaches for navigating digital archives.

4.2 Challenges

The challenges faced in the research are similar to those faced by web archives in general[7] and include:

Scale: as with any web archive we face the problem of applying Natural Language Processing techniques, developing interactive systems and providing storage solutions at scale. The UKGWA contains around 80TB of data, but some of this is duplicated. Even so, the non-duplicated portions of the archive present a potential overhead of many months of computational effort to process the corpus to extract and disambiguate sets of relevant entities. As a result, recent advances in GPU acceleration of NLP algorithms and database management systems will be explored to maximise processing efficiency and to enable deeper analyses that would otherwise be difficult or impossible due to time constraints. GPU acceleration has been the focus of recent studies in NLP [25].

Parallelisation: the size of the corpus does make one aspect of the processing much simpler – the parallelisation of such a collection is trivial as it consists of a large number of documents that can be processed completely independently of each other. This allows greater flexibility when designing a system to process the data - options for such a dataset include clusters of machines, multicore CPUs, and the use of massively parallel processing using GPU cores. However, questions around how to make parallel abound.

User interaction: designing effective features and interfaces for users to support entity-based exploration of the archive. In particular, surfacing large networks of entities will present challenges in making them accessible and usable, particularly for non-experts. Developing techniques that engage users and allow the surfacing of interesting content from the archive in accessible ways will require careful thought [26].

Data: not only is the scale of data a challenge, its heterogeneous nature is also problematic as it requires handling web pages, PDF files and many other formats commonly found in web archives. Many PDF files contain only scanned images of textual content that will require pre-processing stages, including OCR and error correction. Web pages require boilerplate removal prior to entity extraction. Historical proprietary word processor file formats are another potential source of technical issue. Open Source libraries will be used wherever possible, but a processing pipeline will be developed to manage handling diverse data formats and associated processing steps. Such a pipeline would need to demonstrate both vertical and horizontal scalability in order to be relevant to archive-scale use cases.

Structure: capturing and surfacing structural properties and characteristics of the archive data, such as time, will be important. This will help users to contextualise and navigate web content and support user tasks such as viewing the evolution of archive content and entities over time.

Domain: an enduring challenge in entity-based methods adapting techniques and resources (e.g., knowledge bases and vocabularies) to specific domains. For example, in the UKGWA there will likely exist many entities that may occur frequently (e.g., political leaders of less significant parties) that cannot be found and therefore linked to in general resources, such as YAGO2 or DBPedia.

Longevity: of practical importance to TNA is long term support of entity knowledge bases. For example, Freebase has been bought by Google and on 21 August 2016 the Freebase API will be shutdown. On the other hand WikiData, run by Wikimedia, may be more stable over time. However, longevity of knowledge bases used for entity extraction and disambiguation is still an issue with deploying entity-based solutions in archival contexts that needs attention.

5. CONCLUSIONS

This paper describes a collaborative project to investigate the use of entity-centric methods to support exploration and discovery within the UK Government Web Archive. Aspects of the work will involve investigating methods that can operate at scale for the identification, disambiguation and linking of named entities, as well as developing effective interfaces and functionalities to support the wide range of users accessing the archive with varying information needs, goals and tasks. There is a clear need to transfer exploratory research into practice and this project provides a unique opportunity to assist with this transfer.

[7] http://www.netpreserve.org/sites/default/files/resources/2011_06_IIPC_WebArchives-TheFutures.pdf

6. ACKNOWLEDGEMENTS

Work partially supported by the UK Arts and Humanities Council (AHRC) and the UK National Archives.

7. REFERENCES

[1] Benjamins, V.R., Contreras, J., Blázquez, M., Dodero, J.M., Garcia, A., Navas, E., Hernandez, F., Wert, C.: Cultural Heritage and the Semantic Web. In: The Semantic Web: Research and Applications: First European Semantic Web Symposium, ESWS 2004 Heraklion, Crete, Greece, May 10-12, 2004. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg (2004) 433–444

[2] Whitelaw, M.: Generous interfaces for digital cultural collections. Digital Humanities Quarterly **9** (2015)

[3] Johnson, A.: Users, use and context: Supporting interaction between users and digital archives. In Craven, L., ed.: What are Archives? Cultural and Theoretical Perspectives: A Reader. (Ashgate Publishing Ltd)

[4] van den Akker, C., van Nuland, A., van der Meij, L., van Erp, M., Legêne, S., Aroyo, L., Schreiber, G.: From information delivery to interpretation support: Evaluating cultural heritage access on the web. In: Proceedings of the 5th Annual ACM Web Science Conference. WebSci '13, New York, NY, USA, ACM (2013) 431–440

[5] Koolen, M., Kamps, J.: Searching cultural heritage data: Does structure help expert searchers? In: Adaptivity, Personalization and Fusion of Heterogeneous Information. RIAO '10 (2010) 152–155

[6] Lin, J., Gholami, M., Rao, J.: Infrastructure for supporting exploration and discovery in web archives. In: Proceedings of the 23rd International Conference on World Wide Web. WWW '14 Companion, New York, NY, USA, ACM (2014) 851–856

[7] Ardissono, L., Kuflik, T., Petrelli, D.: Personalization in cultural heritage: The road travelled and the one ahead. User Modeling and User-Adapted Interaction **22** (2012) 73–99

[8] Amin, A.K., Hardman, L., van Ossenbruggen, J.R.: Searching in the cultural heritage domain: Capturing cultural heritage expert information seeking needs. Information Systems [INS] (2007)

[9] Vilar, P., Šauperl, A.: Archival literacy: Different users, different information needs, behaviour and skills. In: Information Literacy. Lifelong Learning and Digital Citizenship in the 21st Century. Springer (2014) 149–159

[10] Skov, M.: Hobby-related information-seeking behaviour of highly dedicated online museum visitors. Information Research **18** (2013)

[11] Skov, M., Ingwersen, P.: Exploring information seeking behaviour in a digital museum context. In: Proceedings of the Second International Symposium on Information Interaction in Context. IIiX '08, New York, NY, USA, ACM (2008) 110–115

[12] Hardman, L., Aroyo, L., Ossenbruggen, J.v., Hyvönen, E.: Using ai to access and experience cultural heritage. IEEE Intelligent Systems **24** (2009) 23–25

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

[14] Risse, T., Demidova, E., Dietze, S., Peters, W., Papailiou, N., Doka, K., Stavrakas, Y., Plachouras, V., Senellart, P., Carpentier, F., et al.: The arcomem architecture for social- and semantic-driven web archiving. Future Internet **6** (2014) 688âĂŞ716

[15] Gomes, D., Nogueira, A., Miranda, J., Costa, M.: Introducing the portuguese web archive initiative. (2009)

[16] Gomes, D., Miranda, J., Costa, M.: A survey on web archiving initiatives. Research and Advanced Technology for Digital Libraries (2011) 408–420

[17] Gomes, D., Costa, M., Cruz, D., Miranda, J., Fontes, S.: Creating a billion-scale searchable web archive. In: Proceedings of the 22nd international conference on World Wide Web companion, International World Wide Web Conferences Steering Committee (2013) 1059–1066

[18] Gomes, D., Cruz, D., Miranda, J.a., Costa, M., Fontes, S.a.: Search the past with the portuguese web archive. In: Proceedings of the 22Nd International Conference on World Wide Web. WWW '13 Companion, New York, NY, USA, ACM (2013) 321–324

[19] Costa, M., Silva, M.J.: Characterizing search behavior in web archives. In: TWAW. (2011) 33–40

[20] Berberich, K., Bedathur, S., Neumann, T., Weikum, G.: A time machine for text search. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '07, New York, NY, USA, ACM (2007) 519–526

[21] Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: Yago2: A spatially and temporally enhanced knowledge base from wikipedia. Artif. Intell. **194** (2013) 28–61

[22] Hoffart, J.: Discovering and disambiguating named entities in text. PhD thesis, University of Saarland, Postfach 151141, 66041 Saarbrucken (2015)

[23] Van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., Van de Walle, R.: Exploring entity recognition and disambiguation for cultural heritage collections. Digital Scholarship in the Humanities **30** (2015) 262–279

[24] Maynard, D., Greenwood, M.A.: Large scale semantic annotation, indexing and search at the national archives. In Chair), N.C.C., Choukri, K., Declerck, T., DoÄ§an, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, European Language Resources Association (ELRA) (2012)

[25] Hall, D., Berg-Kirkpatrick, T., Canny, J., Klein, D.: Sparser, better, faster gpu parsing. In: ACL. (2014)

[26] Merčun, T., Žumer, M., Aalberg, T.: Presenting and Exploring the Complexity of Bibliographic Relationships. In: The Outreach of Digital Libraries: A Globalized Resource Network: 14th International Conference on Asia-Pacific Digital Libraries, ICADL 2012, Taipei, Taiwan, November 12-15, 2012, Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg (2012) 63–66