In Situ Text Summarisation for Museum Visitors

Timothy Baldwin[a], Patrick Ye[a b], Fabian Bohnert[b], and Ingrid Zukerman[b]

[a] Department of Computer Science and Software Engineering
The University of Melbourne, Australia
tb@ldwin.net ye.patrick@gmail.com
[b] Faculty of Information Technology
Monash University, Australia
{fabian.bohnert,ingrid.zukerman}@monash.edu

Abstract. This paper presents an experiment on in situ summarisation in a museum context. We implement a range of standard summarisation algorithms, and use them to generate summaries for individual exhibit areas in a museum, intended for in situ delivery to a museum visitor on a mobile device. Personalisation is relative to a visitor's preference for summary length, the visitor's relative interest in a given exhibit topic, as well as (optionally) the summary history. We find that the best-performing summarisation strategy is the Centroid algorithm, and that content diversification and customisation of summary length have a significant impact on user ratings of summary quality.

1 Introduction

With the increasing saturation of mobile technology, museums and other cultural heritage institutions are increasingly looking to deliver content to visitors via their personal mobile device. This has led to a move away from a traditional mode of content delivery via static information on placards in the museum space, to interactive applications on mobile devices supporting path finding, social networking and personalised content delivery (Burnette *et al.*, 2011; Filippini-Fantoni *et al.*, 2011).

This paper explores the feasibility and utility of in situ personalised content delivery in a museum context, focusing on document summarisation. Museums provide a compelling context for in situ summarisation, as visitors often wish to access key information relevant to their immediate surroundings, but want to avoid information overload. Personalised summarisation is an integral component of museum content delivery, as exhibits are typically associated with vast amounts of curated information, predominantly in textual form. This can range from simple tabular information such as the date of acquisition of an exhibit, to full-length research articles published by museum curators/researchers relating to the exhibit. Personalised summarisation offers the possibility to present the most salient facets of information to a visitor, according to their interests and preferences. The pragmatic choice of a mobile device such as a smart phone to deliver the content poses challenges in terms of the amount of content that can be effectively presented to the visitor (Yang and Wang, 2003; Otterbacher *et al.*, 2006).

Personalised summarisation should ideally be coupled with tracking/geolocating technology to be situation aware (Bohnert *et al.*, 2008; Bohnert and Zukerman, 2009; Bickersteth and Ainsley, 2011) and take place interactively (Callaway *et al.*, 2005). In principle, any evaluation should take place in situ as part of an actual museum visit. However, in this preliminary research, we present the results of a web-based user study targeted at members of Melbourne Museum (Melbourne,

** This research was supported in part by grant no. DP0770931 from the Australian Research Council. The authors thank Carolyn Meehan and her team from Museum Victoria for their assistance.

Australia), based on a fixed path through the museum. To partly overcome this limitation, we elicit the participants' interest in an exhibit topic, and generate a summary which takes into account this interest level. In this way, we examine the impact of interest level on summary length and different summarisation strategies, as a guide for future research.

Our contributions are: (1) we deploy a range of extractive summarisation methods over a fixed path through a museum, focusing on generating summaries for individual exhibit areas, personalised in length according to visitor preferences and interest levels, and diversification of summary content; (2) we carry out a medium-scale user study over the generated summaries, to determine the relative utility of the summaries and the effectiveness of the various strategies trialled; and (3) we present the results of a web-based museum visitor questionnaire on opportunities for in situ personalisation in a museum environment.

2 User Study

The goal of this research is to investigate whether automatic text summarisation techniques can be harnessed for in situ personalised content delivery in a museum. We explore this in a two-part web-based survey, where we: (1) presented participants with a questionnaire regarding their interest in mobile device-based personalised content delivery in a museum; and (2) asked participants to rate automatically generated summaries for individual exhibit areas in the museum space. The survey was advertised to members of Melbourne Museum via the official e-newsletter. It was completed outside the museum, but the participant group can reasonably be expected to be very familiar with the museum. In total, we received 34 valid responses to the survey, which form the basis of the results in this paper.

2.1 Questionnaire on Personalised Content Delivery

Survey participants were provided with a brief description of the project, asked a few questions on demographics and their museum visiting patterns, and then asked specific questions about personalised content delivery.

The broad findings of the poll are as follows: (1) over two thirds of participants are very likely or likely to actively interact with a mobile device in the museum for personalised content delivery, and would consent to being tracked for user modelling purposes; (2) two thirds or more of the participants are likely to make use of images, videos or sound bites on the mobile device, but participants are considerably less likely to use (museum-related) games or (external) websites; and (3) most participants are at least likely to use recommendations across a range of content types during a visit, excluding recommendations of games. These survey findings show that there is broad support and scope for personalised content delivery within the museum context.

Participants were also provided with the facility to provide comments. Specific reasons cited for being interested in personalised content delivery (with or without tracking) were: the desire for personalised learning, difficulties in accessing placards in crowded exhibitions, and frustration with "unfathomable" static displays. Primary reasons for participants not being interested in mobile devices were their unsuitability for smaller children (especially in a group setting), the feeling that mobile devices would be a distraction, and also concerns over data privacy and security. There were also comments suggesting that a personalised post-visit follow-up on topics of interest was superior to in situ content delivery.

In sum, participants were generally supportive of in situ personalised content delivery of various types, lending support to the basic premise of this research.

2.2 Rating of Summaries

The second stage of the web survey was to rate the quality of automatically generated summaries. Participants were first presented with three manually generated summaries of varying length (short $\approx$ 50 words; mid-length $\approx$ 75 words; long $\approx$ 100 words) for the Dinosaur Walk exhibit area in

Melbourne Museum, and asked to state their preference for summary length relative to the pre-prepared summaries. This was intended to reduce the effect of summary length bias on summary preference. Next, participants were given the following instructions:

> *Imagine yourself in front of a series of exhibits at Melbourne Museum. First you will be asked about your relative interest level in an exhibit. Assuming a non-zero response (i.e. you have some interest in the exhibit area), you will be presented with three descriptions of the exhibit, and asked to rate each one. Your rating should reflect your response to the quality of the content as well as the language used.*

The participants were then shown four exhibit areas in Melbourne Museum. For each exhibit area, they were given an indication of its location on a map, and a photo of the exhibit area. The rating of interest level in each exhibit area was on a scale of 0 ("Not interested at all") to 3 ("Extremely interested"). If a participant indicated a non-zero interest level in an exhibit area, they were provided with three summaries and asked to rate each on a scale of 1 ("Horrible") to 5 ("Excellent"); if the participant indicated no interest, they were taken directly to the next exhibit area without being shown any summaries. Summary lengths were tailored to both the indicated preference for summary length, and the relative interest level in the exhibit area (Section 3.4).

All participants were shown the same four exhibit areas in the same order. The first two exhibit areas were from the same gallery ("Bugs Alive") with a strong thematic connection relating to insects. The next two exhibit areas were deliberately selected from galleries that have little connection with any of the other three exhibit areas (relating to deep sea life and horse racing, respectively). This was done to explore the impact of exhibit area "theming" (i.e., thematic relation) on content differentiation, as described in Section 3.2 (with the first pairing of exhibit areas being themed, and the second two pairings being "unthemed").

3 Automatic Summarisation

The setting where we apply automatic summarisation is characterised by the following properties:

- **Per-exhibit sequenced summaries:** a stand-alone summary is generated for each exhibit area, for presentation to the visitor in situ, in sequence of the order in which the exhibit areas are visited; each summary is personalised to the summary length preference and interest level of the participant, and optionally based on "content diversification" over preceding summaries.
- **Short summaries:** the summaries need to be relatively short, in order to display them effectively on the small screen of a mobile device.
- **Single primary document:** the amount of text associated with a given exhibit area varies, but is generally of the order of slightly over 1000 words, in a single primary document authored by museum curators; this is complemented with a secondary document from Wikipedia which is anywhere from 600 to slightly over 3000 words in length.

The secondary Wikipedia document is determined by manual alignment for each exhibit area. It is intended to be the document of best fit within Wikipedia, which can vary from a Wikipedia page dedicated to the exact museum artefact to an article which is only thematically related (e.g., an article on gold mining, to represent a historical diorama of a particular gold mine). Our motivation in including these secondary documents is to reduce data sparseness in the sentence ranking, without allowing the content of the Wikipedia article to be included in the summary, as the article often diverges significantly from the context of presentation of the exhibit area.

Our summarisation task is differentiated from conventional document summarisation in: (1) the segmentation into discrete exhibit areas, and generation of short individuated summaries per exhibit area for in situ delivery; (2) the relative sparsity of text (multi-document summarisation is

typically based on at least 10 documents); and (3) the interaction between the physical movements of the visitor and summary generation, in terms of the sequentiality of the exhibit area summaries (mirroring the physical path through the museum), and potential interplay between the summaries generated for different exhibit areas. A primary interest in this research is the determination of the utility of established multi-document summarisation algorithms for our novel summarisation task.

In the following sections, we outline the summarisation algorithms trialled in this research, describe how we personalise summary length, and outline a simple method for content diversification, to avoid repetition in the summaries for individual exhibit areas.

3.1 Summarisation Algorithms

For our experiments, we implemented five standard extractive summarisation algorithms from the literature (Radev *et al.*, 2002):

- First-N sentence algorithm (FIRSTN): select the first N sentences of each document.
- Lead-based algorithm (LEAD): select the first N sentences of each paragraph.
- Centroid algorithm (CENTROID): cluster the document collection using a variant of TF·IDF, and rank sentences through a weighted sum of token weights based on the cluster centroid, a sentence positional weight, and similarity with the first sentence (Radev *et al.*, 2000).
- LexRank algorithm (LEXRANK): cluster the document collection, and rank sentences using a variant of PageRank (Brin and Page, 1998) over the component words (Erkan and Radev, 2004).
- Manifold-ranking algorithm (MANIFOLD): score each sentence based on a manifold-ranking process, and rerank sentences based on a diversity penalty (Wan *et al.*, 2007).

Each algorithm was run over the primary and secondary document for a given exhibit area. All sentences were ranked, but only sentences from the primary document (that authored by Melbourne Museum) were candidates for selection in the final summary. In this sense, the task we are performing is not, strictly speaking, multi-document summarisation so much as document condensation, using a secondary document (and optionally summary context) to bias the sentence selection. Any findings on the relative successes of the summarisation algorithms should be interpreted accordingly.

3.2 Content Diversification

The task of generating in situ content for museum visitors on an exhibit-by-exhibit basis has an inherent segment granularity (summarise one exhibit area at a time) and chronological order, neither of which is found in generic multi-document summarisation tasks. Exhibit areas vary greatly in similarity (Grieser *et al.*, 2011), and for closely-related exhibit areas (e.g., those found in the same gallery, as is the case with our "themed" exhibit area pairing), there is significant potential for generating overlapping content. Due to the strict chronological ordering, we can of course access content previously delivered to the visitor, and personalise the summary to ensure "content diversity", akin to result diversification in personalised web search (Radlinski and Dumais, 2006). That is, we can reduce redundancy in the information presented to the museum visitor by explicitly dispreferring sentences similar to those the visitor has already seen.

We adopt the following rather simple approach: for preceding exhibit area pairs, we bias the sentence ranking by including the summary for the preceding exhibit area as an extra secondary document. Similarly to the Wikipedia document, sentences from this summary are included in the sentence ranking, but cannot be included in the final summary. This has the effect of demoting sentences for the current exhibit area which are very similar to those for the preceding exhibit area, hence leading to diversification.

Preferred length	Interest Level		
	1	2	3
Short	25	50	75
Medium	50	75	100
Long	75	100	125

Table 1: Summary lengths in number of words

3.3 Pronoun Filtering

Many sentences in the museum documents associated with a given exhibit area contain personal or possessive pronouns, which may not be resolvable out of context, or may resolve to an unintended antecedent. To avoid this, we considered the following strategies: (1) remove all sentences containing pronouns from the sentence ranking step, but include them in the clustering step (to avoid exacerbating the effects of data sparseness); and (2) allow sentences containing pronouns, but recursively include the preceding sentence from the original document if a selected sentence includes a pronoun.

3.4 Personalisation and Summary Length

In prior work, Berkovsky *et al.* (2008) found a strong correlation between summary length and interest level, i.e., the more interested a user is in a topic, the greater the likelihood they will prefer a longer summary. To explore this effect further, we first asked participants to state their overall preference for summary length (short, medium or long), prior to presenting them with any exhibit area summaries. For each exhibit area, we generated summaries of varying length for each of three interest levels (1, 2 or 3) and the three summary length preferences, as indicated in Table 1. For each algorithm — optionally in combination with content diversification and pronoun filtering — we fashioned a summary of each of the indicated lengths by monotonically adding sentences from the sentence ranking determined by the summarisation strategy, and selecting the summary which best approximated the required summary length.

When presenting a participant with the summaries for a given exhibit area, we show them a summary of the requested length, in addition to a shorter and longer summary, to determine the relative impact of variance in length on their summary ratings. For example, if a visitor's summary length preference were "mid-length" and they indicated an interest level of 1 ("Have a tiny bit of interest") in a given exhibit area, a summary of length of 50 words would be selected; a summary of length 25 and a summary of length 75 words would then be added (for a total of three summaries per exhibit area).

Our motivation for varying the summary lengths in this manner was to explore the interaction between summary length and perceived quality for different summarisation strategies, hoping to validate the findings of Berkovsky *et al.* (2008). Naturally, in a fully deployed in situ summarisation system, we would hope to dynamically learn the level of user interest (Bohnert and Zukerman, 2009) and customise the summarisation length accordingly. In our current research, we simply hope to establish the need for user interest prediction for the purposes of summary length personalisation.

3.5 Summary Selection

To recap, we generate summaries based on five summarisation algorithms, optional pronoun exclusion, and optional content diversification, for a total of 20 basic summarisation configurations over nine possible summary lengths. These are evaluated over four separate exhibit areas, for each of which we present three summaries of differing length; one pairing of the four exhibit areas is themed, and two are unthemed. All summaries were pre-generated to minimise time lag in the

trial. Due to content diversification being conditioned on the previous summary, the total number of pre-generated summaries was $180 + 180^2 + 180^3 + 180^4 = 1,055,624,580$.

To expose each participant to as many summarisation configurations as possible, over their visit, we performed random selection without replacement over the 20 summarisation configurations. Additionally, for each exhibit area, we randomly varied the order in which the "correct" length summary vs. the short and long summaries were presented to the visitor.

4 Results

4.1 Clustering of Summarisation Configurations

To determine the relative differentiation in content between the pre-generated summaries for each exhibit area, we performed a pairwise summary comparison using the ROUGE-2 metric (Lin and Hovy, 2003). For each pairing of the 20 summarisation configurations, we averaged across the different summary lengths and exhibit areas to generate an overall similarity. Based on these similarity values, we clustered the summarisation configurations using "oblivious" hierarchical agglomerative clustering over the three attributes of summarisation algorithm (binarised into the individual algorithms), pronoun filtering and content diversification. That is, we calculated the single attribute which leads to the (weighted) purest partitioning of the data at each level of the dendrogram in a bottom-up fashion.

Overall, the greatest differentiating factor was the choice of summarisation algorithm, followed by the inclusion/exclusion of sentences with pronouns, and finally, content diversification. Comparing the individual algorithms, LEAD was the most different to the other four algorithms, and CENTROID and FIRSTN generally produced very similar summaries (all irrespective of the pronoun and content differentiation settings). MANIFOLD without pronouns produced summaries similar to CENTROID, while MANIFOLD with pronouns was more differentiated.

4.2 User Study

As stated in Section 2, we received a total of 34 valid responses to the web survey. In terms of the summary ratings, this amounted to up to 12 rated summaries per participant,[1] and a total of 357 summary ratings. This breaks down into two overlapping subsets: 62 summary ratings for the "themed" exhibit area pairing, and 326 ratings for the two "unthemed" exhibit area pairings.

The different factors that potentially impact on the summary ratings are as follows:

- Actual summary length (*act_length* $\in \{-1, 0, 1\}$): the selected summary length, based on the summary length preference, exhibit area interest level, and the shorter ("-1") and longer length ("1") variations over the selected length.
- Summarisation algorithm (*algorithm* $\in \{$LEAD, FIRSTN, CENTROID, LEXRANK, MANIFOLD$\}$): the choice of the summarisation algorithm.
- Interest level (*indicated* $\in \{1, 2, 3\}$): the participant's indicated interest level in the exhibit.
- Summary length preference (*pref_length* $\in \{1, 2, 3\}$): the participant's indicated preferred summary length ("1" = short; "2" = mid-length; "3" = long).
- Content diversification (*diversification* $\in \{0, 1\}$): is content diversification used in the summarisation ("1") or not ("0")?
- Pronoun filtering (*pronoun* $\in \{0, 1\}$): is pronoun filtering used in the summarisation ("1") or not ("0")?

To investigate the interaction between the various factors and the summarisation ratings, as well as the interactions between the factors, we perform a factorial ANOVA (Tabachnick and Fidell,

[1] Participants were not presented with summaries for exhibit areas where they indicated they had no interest. Additionally, it was possible for participants to submit an incomplete set of summary ratings for a given exhibit area.

Combination of factors	All (p-value)	Themed (p-value)	Unthemed (p-value)
{*algorithm,diversification,indicated*}	**0.00**	NaN	**0.00**
{*pref_length*}	**0.00**	0.62	**0.00**
{*algorithm,act_length,indicated,pronoun*}	**0.01**	NaN	**0.03**
{*indicated*}	**0.01**	**0.05**	**0.05**
{*act_length,pronoun*}	**0.01**	0.64	**0.02**
{*diversification*}	**0.02**	0.54	**0.00**
{*algorithm,pref_length,indicated*}	**0.04**	0.94	**0.03**
{*algorithm,pref_length,diversification,indicated,pronoun*}	**0.05**	NaN	**0.05**
{*algorithm,pref_length,act_length,diversification,indicated*}	**0.05**	NaN	**0.03**
{*algorithm,pref_length,indicated,pronoun*}	0.06	NaN	0.06
{*algorithm,pref_length*}	0.07	0.47	**0.04**
{*algorithm,act_length,diversification,indicated*}	0.07	NaN	0.31
{*pref_length,act_length,indicated*}	0.07	0.21	**0.04**
{*pronoun*}	0.07	0.38	0.07
{*algorithm,pref_length,act_length,pronoun*}	0.07	NaN	**0.02**

Table 2: Top-15 combinations of factors for "All" exhibit areas, based on factorial ANOVA, compared to the subset of "Themed" exhibit areas and the subset of "Unthemed" exhibit areas (**boldfacing** indicates a statistically significant combination of factors, $p \leq 0.05$)

2006) for each combination of factors over the participants' ratings. In this, we perform separate factorial ANOVAs for each of: (1) the overall set of exhibit areas ("All"); (2) the themed exhibit areas ("Themed"); and (3) the unthemed exhibit areas ("Unthemed"), to investigate whether thematic relations between exhibit areas have any effect on the relative impact of the factor combinations. Table 2 shows the top-15 factor combinations for "All" exhibit areas (i.e., the factor combinations with the lowest p values). Not all feature combinations are represented in the "Themed" data, leading to gaps in the table (labelled as "NaN"). It is worth noting that there are no statistically significant ($p \leq 0.05$) factor combinations for "Themed" and "Unthemed" outside the top-15 factor combinations for "All" presented in the table.

Looking first at the individual factors, we see that *indicated* is the only factor that universally affects a participant's rating (for all of "All", "Themed" and "Unthemed"). This suggests that, despite the instructions provided to the participants, they found it hard to judge the intrinsic quality of the summaries independently of their relative interest level in the exhibit area. This effect was particularly notable for the CENTROID algorithm (Figure 1).

Content diversification (*diversification*) has a significant impact on "All" and "Unthemed" summaries, but no impact on "Themed" ones ($p = 0.54$ in isolation). Analysis of the data indicates that this is due to diversification negatively affecting unthemed exhibit area pairings (which account for the majority of the data in "All"), but having negligible impact on themed exhibit area pairings. Manual analysis of the summaries for the unthemed exhibit area pairs indicates that this negative impact may be due to *diversification* artificially removing lead sentences (because of the boilerplate structure of the curated documents); the effect was felt less for themed exhibit area pairs, as even if the lead sentence for the second exhibit was dropped, the similarity in content with what had already been presented to the participant meant that the summary was still coherent. It is disappointing that diversification has no impact on the visitors' assessment of themed exhibit summaries, despite the fact that it has a noticeable effect on the summaries themselves, by incorporating extra content (and removing content that overlaps with that provided for the preceding exhibit area, including the lead sentence). However, as participants rated the quality of the summaries, rather than the cumulative novel content delivered over the series of the summaries,

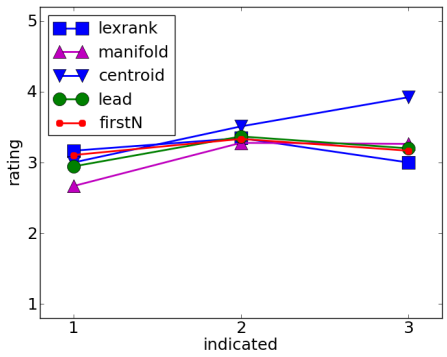

Figure 1: Graph of *indicated* vs. user rating for the different summarisation algorithms

Algorithm	Marginal means
LEAD	3.10 ± 0.51
FIRSTN	3.00 ± 0.42
CENTROID	3.50 ± 0.38
LEXRANK	2.89 ± 0.48
MANIFOLD	3.46 ± 0.53

Table 3: Estimated marginal means for individual summarisation algorithms, with 95% confidence interval

this difference was not reflected in the ratings.

Summary length preference (*pref_length*) has a significant impact on "All" and "Unthemed", but not "Themed". Analysis of the data shows that for "Themed", the choice of algorithm has a highly variable effect on a participant's rating, particularly for participants who indicated they preferred longer summaries (with CENTROID performing strongly), while with the other two datasets, the results were more consistent. Whether this is an effect of data sparseness for the "Themed" data or can be replicated over other pairings of themed exhibit areas is left for future work.

None of *pronoun*, *act_length* and *algorithm* had a significant impact on the participants' ratings. The fact that *act_length* had no impact on results is surprising given the high impact of *indicated*, suggesting that the participants' ratings were biased heavily by their relative interest level and don't reflect subtle variations in summary length. It would be interesting to carry out follow-up experiments with greater differentiation in summary length, and a clearer separation between the rating of exhibit area interest level and intrinsic summary quality.

Table 3 shows the breakdown of estimated marginal means across algorithms (based on "All"). It is evident that CENTROID and MANIFOLD rate better on average than the other methods, but not statistically significantly ($p > 0.05$). Despite strong results in other contexts (Erkan and Radev, 2004), LEXRANK was, surprisingly, the worst performer.

It is harder to tease out any strong trend from the combinations of factors which show up as having a significant impact on ratings, other perhaps than *algorithm* and *indicated* tending to have a significant impact when combined (including other factors). Also, the impact of *pref_length* tends to be dampened when combined with other factors.

Reflecting back on our original question of whether automatic summarisation methods can be applied to personalised content delivery in a museum domain, the answer would appear to be yes, in that the best-performing methods produced summaries which were rated around 3.5 on average (with 3 indicating "Not bad" and 4 "Good"). Contrary to our clustering results from Section 4.1, the choice of algorithm was found to have no significant effect on summary rating, with *indicated*, *pref_length* and *diversification* having a greater individual impact on summary quality. Finally, the correlation observed by Berkovsky *et al.* (2008) between interest level and preferred summary length was not strongly evident in our results.

5 Related Work

While multi-document summarisation has been a highly active research area, personalised summarisation over single documents has received considerably less attention.

There has been work on personalized document search and summarisation in the medical domain for clinicians and patients, based on semantically-enhanced extraction of snippets and terminology standardisation (McKeown *et al.*, 2001; McKeown *et al.*, 2003). Here, however, the personalisation was at the level of two discrete user profiles, and not truly individualised.

Radev *et al.* (2001) present the design of a search engine which supports recommendation, clustering, and personalised summarisation, but do not include any technical details or evaluation. Goren-Bar and Prete (2005) report on a preliminary situated experiment on content delivery in a museum, but found that the simple static text delivery method was preferred to the adaptive method.

Berkovsky *et al.* (2008) present a method for generating summaries of different length, and demonstrate a correlation between the level of user interest in a topic area, and the preferred summary length. However, their method relies on hand-compiled summaries of expanding length, and no attempt was made to automate the summary generation.

The content differentiation aspect of our work is somewhat related to the TAC 2008 Update Summarization Task (Dang and Owczarzak, 2008), where participants were provided with a set of 10 documents and a pre-prepared summary, in addition to a set of 10 update documents containing new information, from which a summary was to be generated. In the TAC 2008 task, the update summary was for the same topic as the original summary, and the task was specifically to highlight novel content. In our setup, the exhibit areas overlap in content to varying degrees (depending on theming), and the relative importance of differentiation is more subtle (partly because the museum visitor sees the summaries in discrete chunks, in the context of different exhibit areas).

6 Conclusions

We have presented a user study on personalised summarisation for a museum visit. We implemented a range of summarisation algorithms, and used them to generate summaries for individual exhibit areas in a museum, intended for in situ delivery to a museum visitor on a mobile device. Personalisation took the form of adjustment of summary length on the basis of the visitor's indicated interest level in a given exhibit area, as well as (optionally) diversification over previously-delivered summaries. We found that museum visitors are largely supportive of personalised content delivery, and explored the impact of a range of summarisation parameters on visitors' ratings of summaries.

References

Berkovsky, Shlomo, Timothy Baldwin, and Ingrid Zukerman. 2008. Aspect-based personalized text summarization. In *Proceedings of the 5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2008)*, pp. 267–270, Hanover, Germany.

Bickersteth, Julian and Christopher Ainsley. 2011. Mobile phones and visitor tracking. In *Museums and the Web 2011: Proceedings*, `http://conference.archimuse.com/mw2011/papers/mobile_phones_and_visitor_tracking`.

Bohnert, Fabian and Ingrid Zukerman. 2009. Non-intrusive personalisation of the museum experience. In *Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization (UMAP-09)*, pp. 197–209, Trento, Italy.

Bohnert, Fabian, Ingrid Zukerman, Shlomo Berkovsky, Timothy Baldwin, and Elizabeth Sonenberg. 2008. Using interest and transition models to predict visitor locations in museums. *AI Communications*, 21(2–3), 195–202.

Brin, Sergei and Larry Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh World Wide Web Conference*, pp. 107–117, Brisbane, Australia.

Burnette, Allegra, Rich Cherry, Nancy Proctor, and Peter Samis. 2011. Getting on (not under) the Mobile 2.0 bus: Emerging issues in the mobile business model. In *Museums and the Web 2011: Proceedings*. `http://conference.archimuse.com/mw2011/papers/getting_on_not_under_the_mobile_20_bus`.

Callaway, Charles, Tsvi Kuflik, Elena Not, Alessandra Novello, Oliviero Stock, and Massimo Zancanaro. 2005. Personal reporting of a museum visit as an entrypoint to future cultural experience. In *Proceedings of the 2005 International Conference on Intelligent User Interfaces (IUI 2005)*, pp. 275–277, San Diego, USA.

Dang, Hoa Trang and Karolina Owczarzak. 2008. Overview of the TAC 2008 update summarization task. In *Proceedings of 2008 Text Analysis Conference*, pp. 1–16, Gaithersburg, USA.

Erkan, Günes and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22, 457–479.

Filippini-Fantoni, Silvia, Sarah McDaid, and Matthew Cock. 2011. Mobile devices for orientation and way finding: the case of the British Museum multimedia guide. In *Museums and the Web 2011: Proceedings*. `http://conference.archimuse.com/mw2011/papers/mobile_devices_for_wayfinding`.

Goren-Bar, Dina and Michela Prete. 2005. Report on a museum tour report. In *Proceedings of INTETAIN 2005: INtelligent TEchnologies for interactive enterTAINment*, pp. 230–234, Madonna di Campiglio, Italy.

Grieser, Karl, Timothy Baldwin, Fabian Bohnert, and Liz Sonenberg. 2011. Using ontological and document similarity to estimate museum exhibit relatedness. *ACM Journal on Computing and Cultural Heritage*, 3(3), 1–20.

Lin, Chin-Yew and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 3rd International Conference on Human Language Technology Research and 4th Annual Meeting of the NAACL (HLT-NAACL 2003)*, pp. 71–78, Edmonton, Canada.

McKeown, Kathleen R., Shih-Fu Chang, James Cimino, Steven K. Feiner, Carol Friedman, Luis Gravano, Vasileios Hatzivassiloglou, Steven Johnson, Desmond A. Jordan, Judith L. Klavans, André Kushniruk, Vimla Patel, and Simone Teufel. 2001. PERSIVAL, a system for personalized search and summarization over multimedia healthcare information. In *Proceedings of the 2001 Joint Conference on Digital Libraries (JCDL)*, Roanoke, USA.

McKeown, Kathleen R., Noemie Elhadad, and Vasileios Hatzivassiloglou. 2003. Leveraging a common representation for personalized search and summarization in a medical digital library. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 159–170.

Otterbacher, Jahna, Dragomir Radev, and Omer Kareem. 2006. News to go: hierarchical text summarization for mobile devices. In *Proceedings of 29th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pp. 589–596, Seattle, USA.

Radev, Dragomir R., Weiguo Fan, and Zhu Zhang. 2001. Webinessence: A personalized web-based multi-document summarization and recommendation system. In *NAACL Workshop on Automatic Summarization*, Pittsburgh, USA.

Radev, Dragomir R., Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational linguistics*, 28(4), 399–408.

Radev, Dragomir R., Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization*, pp. 21–30, Seattle, USA.

Radlinski, Filip and Susan Dumais. 2006. Improving personalized web search using result diversification. In *Proceedings of 29th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pp. 691–692, Seattle, USA.

Tabachnick, Barbara G. and Linda S. Fidell. 2006. *Using Multivariate Statistics*. Allyn & Bacon, Boston, USA, 5th edition.

Wan, Xiaojun, Jianwu Yang, and Jianguo Xiao. 2007. Manifold-ranking based topic-focused multi-document summarization. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-2007)*, pp. 2903–2908, Hyderabad, India.

Yang, Christopher C. and Fu Lee Wang. 2003. Fractal summarization for mobile devices to access large documents on the web. In *Proceedings of the 12th International Conference on World Wide Web*, pp. 215–224, Budapest, Hungary.