WHAT HAPPENS IF WE SWITCH THE DEFAULT LANGUAGE OF A WEBSITE?

Te Taka Keegan, Sally Jo Cunningham

Computer Science Department, University of Waikato,Hamilton, New Zealand

Keywords: User Interfaces, Usability, Digital Libraries, Multi-lingual web sites

Abstract: In this paper we investigate the effect of the default interface language setting on a bilingual website. Log file analysis is undertaken to determine usage patterns of the Niupepa digital library (a collection of historic Māori language newspapers) when the default interface language is switched between Māori and English in alternate weeks. Activity is grouped into active user sessions, which are further analysed to determine methods of access and searching patterns. The results clearly show that changing the default language of a website will affect the ways in which users access information.

1 INTRODUCTION

The bilingual website that we chose to investigate is called the Niupepa Collection. It is being served by the Greenstone software (Witten and Bainbridge 2002) of the New Zealand Digital Library (NZDL) at: www.nzdl.org/niupepa. Niupepa is a collection of historic Māori newspapers published between 1842 and 1933. It is a large source of historic texts, almost 18,000 newspapers pages. The newspapers are available in a full text

format and in two facsimile forms: a low resolution image that downloads quickly for previewing, and a high resolution image that takes longer to download but is readable on screen (Apperley et al 2002).

70% of the documents are written in Māori, 27% are written bilingually in both Māori and English, and 3% written in English only. The collection is a rich source of Māori language texts in an environment where there is a dearth of Māori language resources. The default language of the collection is normally set to Māori.

Figure 1: Home Page of the Niupepa Website

Approximately 14% of New Zealand's total resident population are Māori and about 1 in 4 of these are able to converse in Māori (Te Puni Kōkiri 2003). When we consider that 65% of Māori have never accessed the Internet, the potential users of a Māori language interface is perhaps 1-2% (40,000-80,000) of New Zealand's population. In contrast, potential English-speaking New Zealand users of the website are approximately 51% (2,040,000) (Te Puni Kōkiri 2001). By setting the default language of the Niupepa Collection to Māori we are clearly going against the preferences of the majority of potential users. Research undertaken by Jones (Jones et al 2000) suggests that users of a digital library system rarely amend the default settings for options with regard to query types and result displays.

Through the use of log file analysis this research seeks to find answers to what differences will occur in user behaviour if the default language of an interface is alternated between Māori and English.

2 LIMITATIONS OF LOG FILE ANALYSIS

As a method of gathering data on user actions log file analysis has its shortcomings– the primary one being the effect of web caches. Web caches sit between web servers and clients and will serve repeated requests to the client without having to bother the original server. This saves time and reduces network traffic. However the issue is that the original server does not receive the request and consequently it is not be recorded in its web log file.

There are two main types of web caches: a browser cache which is handled by a user's browser software and a proxy cache which is configured within a network. Both types of caches will prevent repeated requests from a single user appearing in web log files. However a network cache will also mask repeated requests from different users within its network.

There are other limitations of log file analysis including false hits due to web robot activity, false hits due to server upgrades and maintenance; and the inability to accurately delimit user sessions.

3 GATHERING THE DATA

The NZDL website (www.nzdl.org) makes available over 40 different collections in various formats. All user activity is logged. Every request or 'hit' is recorded along with information such as the page requested, the language used in the interface, the time of the request, the type of request, the previous action, the IP address of the requestor and the various preferences that are set.

The NZDL site is mirrored with a site located at the University of Lethbridge in Alberta, Canada. The New Zealand site, located at the University of Waikato, serves the collection to Web requests from within New Zealand. The Lethbridge mirror site is responsible for serving the collection to Web requests from outside of New Zealand. The data collected in this analysis is from the University of Waikato site only, and thus only reflects usage within New Zealand.

We chose to analyse a four week period running from 8.40am Monday 5 July 2004 to 8:40am Monday 2 August 2004. In the first week we changed the Niupepa default language setting to English. The second week we changed it back to Māori. The third week it was in English and the fourth week we changed it back to Māori again. The raw NZDL log file was collected for this time period and the hits relating to the Niupepa collection were extracted.

The raw Niupepa data was then further filtered to remove hits of unwanted origin. These included incorrect language argument (27), undefined IP address of requestor (213), web crawler and web robot hits (33) and hits from the local research team (18). This left a total of 14,416 hits, of which 7724 were in the weeks that the default language was set to English and 6692 were in the weeks when the default language was set to Māori (Graph 1).

Clearly there are significantly more hits (15.4% more) when the default language of the website is set to the more commonly spoken English language. It is also clear to see, and perhaps unsurprising to note, that the number of hits in English increases when the default language is set to English and the number of hits in Māori increases when the default language is set to Māori.

4 DEFINING THE SESSIONS

To further analyse what the users of the web site were doing, the hits recorded in the log files were grouped into sessions. Cookies were used to define these sessions. When users connect to the website a cookie is created on their machines which holds information that includes the IP address of the machine connecting and the time that the cookie was created. This information is recorded with each hit as the z argument of a hit and so a sessions is simply a group of hits with the same z argument within a given time frame. Web analysis software usually defines a session as a series of hits such that the time

Filtered Niupepa Hits
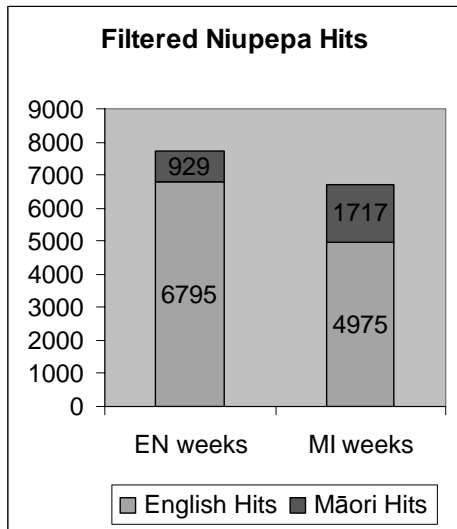
Graph 1: Total filtered hits for the weeks under analysis

gap between any two successive hits is less than 30 minutes. However because Niupepa users may spend long periods reading single newspaper pages is was more appropriate to extend the minimum time length between hits to 60 minutes.

Once the log file of hits was grouped into sessions, the types of sessions were then arranged based on the length of the session and what the user accessed. This gave three types of sessions:

o *Single hit*: these could be a request where the user has clicked on the site and decided not to look any further, or who could be users that have not enabled cookies.
o *Exploratory only sessions:* multi-hit sessions where the user only accessed the home page, the help page, and/or the preferences page. No documents in the collection were accessed and no searches were undertaken.
o *Extended sessions:* multi-hit sessions where documents of the Niupepa collection where accessed and/or searches were undertaken.

From Graph 2 we can see there is an increase (23.4%) from the number of single hits in the EN weeks (367) to the number of hits in the MI weeks (453). This can be expected because the majority of users logging onto the website are not Māori literate, so when they first log onto the site and see that it is in Māori they are perhaps not prepared to continue. By contrast the number of extended sessions and exploratory sessions were quite similar for both English and Māori default language settings of the website. This suggests that active users of the website were intent on using the website no matter which default language the website is set to.

Graph 1 showed that the total number of hits was lower in the Māori weeks than in the English weeks. However Graph 2 showed that the number of sessions in the Māori weeks was higher than in the English weeks. This implies that the difference rests in the activity that the users are undertaking when they are accessing the web site. The single hit and exploratory sessions are not as important as they do
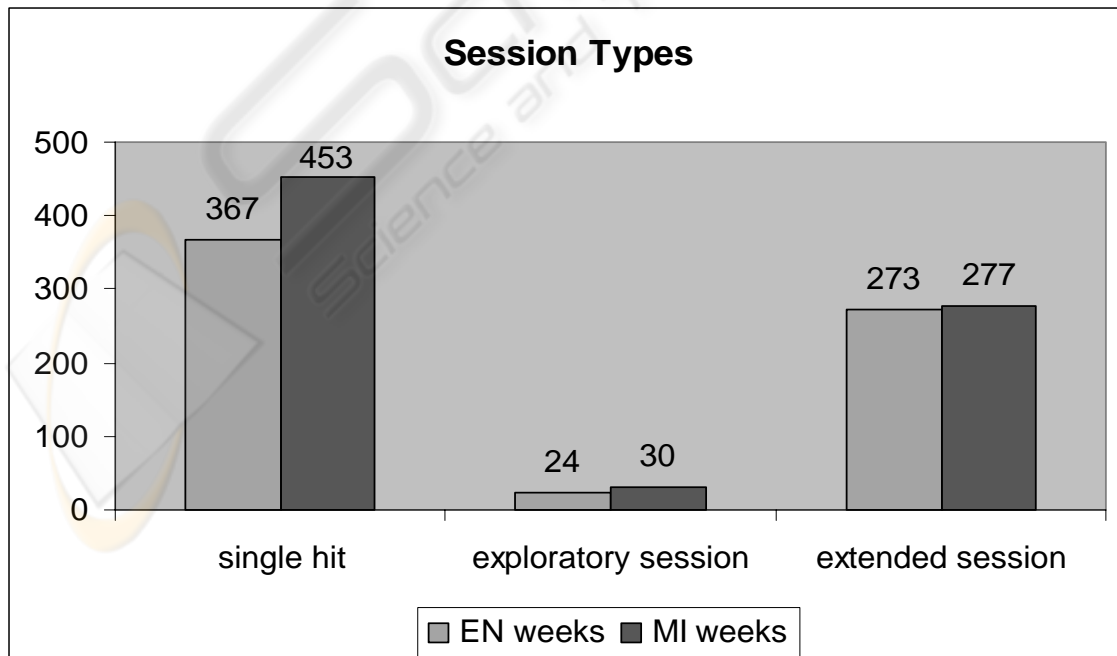
Session Types

Graph 2: Sessions Types for the weeks under analysis

not represent activity from users who are actively using the website. Although steps have been taken to remove hits from known web crawlers and known web robots we still cannot be sure that hits generated by non human activity have been completely removed from the statistics. However by definition an extended session involves undertaking searches and/or browsing the collection; consequently we can be confident that these hits are a result of human activity. The next stage of the analysis involves having a closer look at the extended sessions.

5 EXTENDED SESSION ANALYSIS

The extended sessions were further subdivided into language categories. Initially we decided that there were just two categories: those sessions that were conducted in English, and those sessions that were conducted in Māori. However when we looked at the data closely we realized that there was a third type of session: the bilingual session, in which users conducted their interaction with both English and Māori interfaces.

A Māori language session was defined as a session that was conducted in Māori at least 80% of the time and did not involve more than two user interface language switches. An English language session was defined as a session that was conducted in English at least 80% of the time and did not involve more than two user interface language switches. A bilingual language session was a session that involved three or more user interface language switches and/or sessions that spent at least 20% of the activity in both languages.

Table 1: Summary Statistics for English Sessions.

English Extended Sessions		
	EN wks	MI wks
total sessions:	210	172
min hits/session:	2	2
max hits/session:	286	136
mean hits/session:	29.4	25.3
median hits/session:	12.5	16
Std Dev hits:	41.5	26.2
shortest (min):	<1	<1
longest (min):	233	191
mean (min):	22.3	22.3
median (min):	7.5	8.5
Std Dev (min):	36.2	31.5

When we consider the English Extended Sessions and compare the weeks that the default language is set to English (EN weeks) with the

weeks that the default language is set to Māori (MI weeks), (Table 1) we can see a number of differences. There are more sessions, 210 as opposed to 172, when the default language is set to English. The mean number of hits in the English default weeks is larger, the standard deviation is larger, and as the median is lower in the English default weeks which indicates large sessions is at the upper extreme of the data. Overall these statistics suggest that sessions in English involve more activity over a similar time period when the default language of the interface is set to English.

Table 2: Summary Statistics for Māori Sessions

Māori Extended Sessions		
	EN wks	MI wks
total sessions:	48	76
min hits/session:	2	2
max hits/session:	76	132
mean hits/session:	13.5	16.7
median hits/session:	8	11.5
Std Dev hits:	14.9	19.3
shortest (min):	<1	<1
longest (min):	103	116
mean (min):	12.4	14.4
median (min):	5.5	5.5
Std Dev (min):	20.0	23.2

A similar pattern occurs for Māori language sessions when the default language is set to Māori (Table 2): there are more Māori sessions, sessions contain more activity (hits), and sessions are conducted over a longer time period.

Table 3: Summary Statistics for Bilingual Sessions

Bilingual Extended Sessions		
	EN wks	MI wks
total sessions:	15	29
min hits/session:	2	3
max hits/session:	182	62
mean hits/session:	31.7	19.2
median hits/session:	8	8
Std Dev hits:	47.5	16.7
shortest (min):	<1	<1
longest (min):	241	117
mean (min):	32.5	14.3
median (min):	8	8
Std Dev (min):	64.0	23.4

The statistics for extended bilingual sessions, as shown in Table 3, also show differences between the English default weeks and the Māori default weeks. Twice as many sessions are defined as bilingual in the Māori default weeks perhaps because a larger number of users are switching more often when the default language is set to Māori. Generally, the extended bilingual session averages in the English default weeks are similar to the English sessions averages in the English default weeks and the bilingual session averages in the Māori default weeks are similar to the Māori sessions averages in the Māori weeks. However as the standard deviations are high and the number of sessions that we are dealing with is low it would be premature to deduct any firm conclusions.

6 ACCESSING DOCUMENTS

The NZDL web log records if the pages accessed were the result of a search, a browse by newspaper publication (series) or a browse by date. The results are grouped by language of the extended session. In Table 4 we can see the document types accessed in the English Extended Sessions. It is apparent that switching the default language doesn't make any difference in how users access documents in the English sessions. However Table 5 shows quite a large difference with Māori Extended Sessions, indicating a large preference (90.9%) to access pages by using a search when the default language is set to English but a much lower preference (69.9%) when the default language is set to Māori. It appears that when the default language is set to Māori, the Māori sessions have a higher possibility of browsing the documents by series or by date. It should be noted however that the number of documents viewed in the Māori sessions is much smaller than the other two sessions. The bilingual extended session preference for accessing pages (Table 6) is similar to the English sessions, with no significant differences when the default language is switched.

Table 4: Types of Documents Accessed in English Sessions

English Extended Sessions		
	EN wks	MI wks
total documents viewed:	4265	3070
pages viewed from search:	81.6%	81.9%
pages viewed from series:	12.0%	12.7%
pages viewed from date:	3.5%	3.0%
other:	2.9%	2.4%

Table 5: Types of Documents Accessed in Māori Sessions

Māori Extended Sessions		
	EN wks	MI wks
total documents viewed:	495	901
pages viewed from search:	90.9%	69.9%
pages viewed from series:	7.1%	13.0%
pages viewed from date:	1.6%	14.9%
other:	0.4%	2.2%

Table 6: Types of Documents Accessed in Bilingual Sessions

Bilingual Extended Sessions		
	EN wks	MI wks
total documents viewed:	4760	3971
pages viewed from search:	82.6%	79.1%
pages viewed from series:	11.5%	12.7%
pages viewed from date:	3.3%	5.7%
other:	2.6%	2.4%

7 SEARCHING

The final analysis undertaken for this paper was an examination of the searching characteristics with each of the three types of sessions to see if there were any differences when the default language was changed. The searching summary for English Extended Sessions is shown in Table 7. It can be seen that more searches are submitted when the default language is set to English; a consequence of having a higher number of English extended sessions in these weeks. The average number of searches per session is slightly higher (5.0 versus 4.4) in the English default weeks, and there is a higher percentage of sessions without any searches (36.2% versus 19.2%) in the English default weeks as well. The average number of terms submitted per search is very similar (1.9 versus 2.2) for both default language settings.

Table 7: Searching Summary for English Sessions

English Extended Sessions		
	EN wks	MI wks
number of searches:	1046	749
average per session:	5.0	4.4
maximum searches:	80	43
minimum searches:	0	0
standard deviation:	3.8	2.2
0 searches:	36.2%	19.2%
1 search:	11.9%	14.0%
2-3 searches:	14.8%	27.9%
4+ searches:	37.1%	39.0%
average search terms:	1.9	2.0

Table 8 presents the searching summary for Māori extended sessions. More searches are submitted when the default language is set to Māori; a consequence of having a higher number of Māori extended sessions in these weeks. The average number of searches per sessions is slightly higher (3.1 versus 2.6) in the Māori default weeks, and there is a slightly higher percentage of sessions without any searches (28.9% versus 27.1%). Again, the average number of terms submitted per search is similar (1.9 – 2.2) for both default language settings.

Table 8: Searching Summary for Māori Sessions

Māori Extended Sessions		
	EN wks	MI wks
number of searches:	126	238
average per session:	2.6	3.1
maximum searches:	19	18
minimum searches:	0	0
standard deviation:	2.0	1.8
0 searches:	27.1%	28.9%
1 search:	25.0%	19.7%
2-3 searches:	29.2%	22.4%
4+ searches:	18.8%	28.9%
average search terms:	1.9	2.0

The searching summary for bilingual extended sessions is displayed in Table 9. It can be seen that there are more than twice as many searches submitted in the Māori default weeks (115) than in the English default weeks (51). This is surprising as Table 6 indicates that there are fewer documents accessed in the Māori default weeks (3971) than in the English default weeks (4760). This, and the fact that the English default weeks have a higher percentage of 0 searches per session (40.0% compared with 27.6%), suggests that the bilingual sessions use more browsing when the default language is set to English and have very effective searches, but when the default language is set to Māori more searches are undertaken and the searches are not as effective.

Table 9: Searching Summary for Bilingual Sessions

Bilingual Extended Sessions		
	EN wks	MI wks
number of searches:	51	115
average per session:	3.4	4.0
maximum searches:	18	14
minimum searches:	0	0
standard deviation:	2.7	1.7
0 searches:	40.0%	27.6%
1 search:	26.7%	17.2%
2-3 searches:	13.3%	20.7%
4+ searches:	20.0%	34.5%
average search terms:	1.5	1.8

8 CONCLUSION

Designers of bilingual websites should be aware that setting the default language of a website strongly favours usage of the website in that language. This will occur despite users having the ability to easily switch the interface language.

An initial look at the data in Table 2 suggests that there is little difference in the two default language settings. However when the actual user data is analysed, the extended session analysis, we can see that the number of sessions in the default language is higher, the sessions are longer and include more user activity, involve the accessing of more pages, and there is less of a reliance on searching to access pages.

We have also discovered a new user type, the bilingual user, who conducts a significant amount of activity in both languages and whose usage characteristics seem to alter depending on the default language setting.

One consideration with these results is the data is drawn from just two 2 week time periods. Analysis over a longer period will produce more conclusive results.

REFERENCES

Apperley M. D., Keegan T. T., Cunningham S. J., Witten, I. H., 2002. Delivering The Māori Newspapers on the Internet in Curnow J, Hopa N, McRae J (ed.s), *Rere Atu Taku Manu! Discovering History Language and Politics In The Māori Language Newspapers*. Auckland University Press. Pages 211-36

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

Te Puni Kōkiri, 2001. *Māori Access to Information Technology*. Te Puni Kōkiri, Wellington, New Zealand.

Te Puni Kōkiri, 2003. *Speakers of Māori within the Māori Population*. Te Puni Kōkiri, Wellington, New Zealand.

Witten, I. H., Bainbridge, D., 2002. *How to Build a Digital Library*. Morgan Kaufmann. San Francisco, CA.