University Student Use of the Wikipedia

Andrew Trotman
Department of Computer Science
University of Otago
Dunedin, New Zealand
andrew@cs.otago.ac.nz

David Alexander
Department of Computer Science
University of Otago
Dunedin, New Zealand
dalexand@cs.otago.ac.nz

Abstract: *The 2008 proxy log covering all student access to the Wikipedia from the University of Otago is analysed. The log covers 17,635 student users for all 366 days in the year, amounting to over 577,973 user sessions. The analysis shows the Wikipedia is used every hour of the day, but seasonally. Use is low between semesters, rising steadily throughout the semester until it peaks at around exam time. The analysis of the articles that are retrieved as well as an analysis of which links are clicked shows that the Wikipedia is used for study-related purposes. Medical documents are popular reflecting the specialty of the university. The mean Wikipedia session length is about a minute and a half and consists of about three clicks.*

The click graph the users generated is compared to the link graph in the Wikipedia. In about 14% of the user sessions the user has chosen a sub-optimal path from the start of their session to the final document they view. In 33% the path is better than optimal suggesting that users prefer to search than to follow the link-graph. When they do click, they click links in the running text (93.6%) and rarely on "See Also" links (6.4%), but this bias disappears when the frequency of these types of links' occurrence is corrected for.

Several recommendations for changes to the link discovery methodology are made. These changes include using highly viewed articles from the log as test data and using user clicks as user judgements.

Keywords: Information Retrieval, Link Discovery.

1. Introduction

Keeping the link structure up-to-date in a large hypertext collection is difficult. When a new document is added to the collection it is necessary to link from that document to the collection and from the collection to that document. When a document is deleted all links from the collection to the document must be removed. Finally, when a document changes, new links must be added and old links deleted. Deleting links is a mechanical process, but recommending links for new or changing documents is problematic and is an active field of research known as Link Discovery.

Milne & Witten [11] use machine learning to learn links for documents to be added to the Wikipedia. INEX has the Link-the-Wiki track [3] in which the task is to analyse a document (also from the Wikipedia) and to construct an ordered list of links from which a user can choose; Geva [1] and Jenkinson et al. [7] provide the best solutions.

The recent INEX study by Huang et al. [5] raises questions about the validity of the methods of assessment that had been used with all previous solutions to the Link Discovery problem, and therefore the validity of the solutions themselves.

The prior INEX protocol was as follows: A dump of the Wikipedia is taken. From that dump a single document is extracted (the *orphan*). All links between the orphan and the collection are removed. The task is to recommend links for the orphan. Performance is measured relative to the links that were originally in the orphan.

Huang et al. introduced a new protocol to INEX, based on the Cranfield methodology. In this protocol, INEX participants' runs were pooled and manually assessed. Importantly, the links in the original Wikipedia articles were added to the pool. Most importantly, the Wikipedia articles themselves were scored against the pool. Unexpectedly, the Wikipedia articles performed no better than the best submitted runs.

This result suggests that there are many links in the Wikipedia that are not considered relevant to the topic of the articles. The nature of those non-relevant links is not known, but could be studied by analysing the INEX assessments.

This approach would shed light on the nature of relevant and irrelevant links in the Wikipedia and could be used both to help recommend new links and to remove bad links. But a link that is *relevant to the content* of the page may not be *relevant to the information need* of the user. To find *useful* links it is necessary to study how users use links. This raises our research question: *How do users use the Wikipedia link structure?*

To answer this question we studied the log of the University of Otago student web proxy, which all student users of the University computing facilities must pass through, for the 2008 calendar year. From the log we extracted all references to the Wikipedia.

Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 4 December 2009.

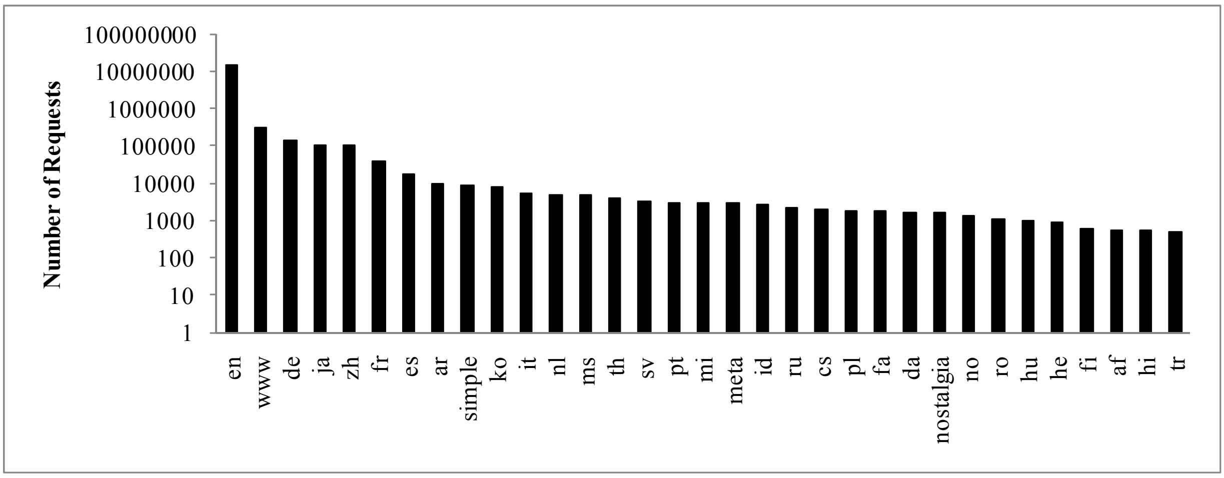

Figure 1: Frequency of use of the versions of the Wikipedia seen more than 500 times in the log. English is the preferred language followed by German, Japanese, Chinese, French, Spanish and so on. The subdomains for language versions of Wikipedia are ISO 639 codes.

Before studying the link-clicking behaviour displayed in the log, we performed a number of preliminary analyses in order to better understand the data, and its applicability to our goal of improving the link structure of Wikipedia. These included examining the request frequency at different times of day and times of year, calculating the length of user sessions, and finding the most commonly-requested pages. The results of these analyses are presented in Sections 3.1 and 3.2

In Section 3.3, the link-clicking behaviour seen in the log is analysed, with particular focus on the question of whether or not the current link graph is being used efficiently. This question is addressed in two ways. The first is to determine the proportion of links clicked on in each article, and to look for patterns in the types of links clicked. The second is to determine whether or not users are reaching their destinations by following links, and if so, whether or not they are doing so in the most efficient way possible.

2. Prior Work

Prior IR research on logs has focused on search engine log analysis. Zhang & Moffat [14], for example, present an analysis of the MSN log while Spink et al. [13] present an analysis of an Excite log.

Internet use by students has previously been studied; however such studies are typically conducted through surveys, for example Metzger et al. [9].

Proxy log use has been limited. Kamps et al. [8] used a (3 month long) New Zealand high school proxy log to validate INEX 2007 results. Their analysis is short. They state: the number of queries; the number of unique queries; the number of clicks in the Wikipedia; the number of queries with Wikipedia clicks; and the number of unique queries with Wikipedia clicks.

There is a growing body of work in link recommendation. Early work [10, 11] conducted outside INEX considers the problem of generating a set of links. INEX considers link discovery to be a recommender task and consequently systems generate a ranked list of results. Geva's solution [1] at INEX is to match the titles of Wikipedia documents against the text of a document, preferring longer titles if several overlap. The Jenkinson et al. solution [7] is based on Itakura & Clarke [6]. They generate a list of all anchors used in the collection along with a list of all documents that are targeted by each anchor text. They rank anchor texts on the frequency with which they occur as links, as a proportion of their overall frequency. They then search for these in the new document and recommend links based on the above frequency. The two approaches perform comparably.

3. Analysis

In this section an analysis of the proxy log is given. The global statistics are presented followed by an analysis of the sessions. Finally the use of the hypertext links is given.

3.1. Global Statistics

The proxy log covers the period from 1st January 2008 to 31st December 2008. It covers 366 days because 2008 was a leap year. The proxy configuration at the university consists of a set of proxies each logging and fulfilling user requests. There were a total of 6 proxy servers and so the analysis is over 2,196 source log files. One of these files (from 30 April 2008) was lost and so the analysis is short by one sixth on that date.

All lines from the log that contained the (case insensitive) word *Wikipedia* were extracted. There were a total of 16,665,418 references in the extracted log, of which 15,696,225 were to the English Wikipedia and 969,193 were to other sites. The references were made by 17,635 students (the university had 20,752 enrolled during 2008). Further fundamental numeric statistics are shown in Table 1.

Table 1: Fundemental statistics of the log

Duration of Log	1/Jan/2008 – 31/Dec/2008
Rows in Log	16,665,418
Rows for English Wikipedia	15,696,225
Users in Log	17,635
Total enrolled students	20,752
Sessions in Log	577,973
Articles Accessed	340,477
Articles in Wikipedia	2,600,000 (approx.)
Wikipedia Subdomains	202 (inc. typos)

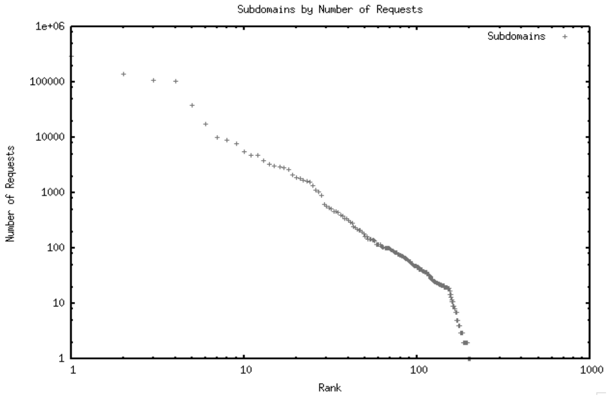

Figure 2: Frequency of use of all versions.

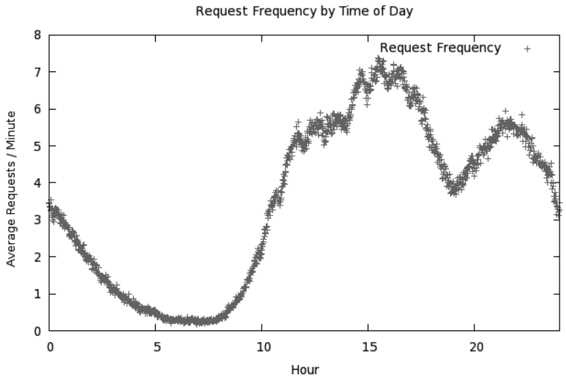

Figure 3: Access to the Wikipedia by time of day.

The Wikipedia exists in many different languages and forms. Each of these versions has its own subdomain of *wikipedia.org*. In the log there are 202 references to different variants (including spelling errors). The most common is the English Wikipedia while the least common (occurring only once) is species.wikipedia.org, the Wikipedia Free Species Directory.

Figure 1 graphs the frequency of use of those variants of the Wikipedia seen in the log more than 500 times. The graph shows that English (subdomain *en*) is the primary language used at Otago, with European and Asian languages also popular. The Māori Wikipedia (subdomain *mi*) was the 17th most popular version, accessed 2,953 times.

All of the subdomains shown in Figure 1 are identified by the ISO 639 codes for their languages, except *www* (an entry point to Wikipedia, having links to the most popular language versions), *simple* (the Simple English Wikipedia, in which articles are written at a level suitable for non-native English speakers or children), *meta.wikimedia.org* (a wiki containing information useful to editors of the various Wikimedia projects), and *nostalgia* (a static copy of a 2001 version of Wikipedia).

Figure 2 shows the request frequency of all subdomains of *wikipedia.org* and *wikimedia.org*. It shows that the subdomains do not completely follow a power-law distribution.

Timestamps in a search engine log are relative to the search engine location. It is therefore not possible to know the user-time at which each query was given. In a proxy log of the type used in this study, however, the user time is the same as the time recorded at the proxy.

Figure 3 shows the mean number of requests per minute at each hour of the day. At midnight there is moderate access steadily falling to low at 5am where access picks up and stabilizes at about 11am. A local peak is seen at 3pm with a dip at dinner-time, picking up at about 7pm and falling again at about 10pm. Student use of the Wikipedia is round-the-clock.

This finding is in line with results seen by others. Zhang & Moffat [14] found that there was no hour of the day at which the MSN search engine was completely unused from within the US. The US, however, is a somewhat larger geographical area then the University of Otago (and has a larger population).

Publicly available search engine logs tend to cover a very short period of time. The MSN log is one month in length, the Excite logs are one day, and the Alta Vista log is about six-weeks. From such short logs it is not possible to make any observations about seasonal user behaviour, analyses have been restricted to daily patterns.

Zhang & Moffat [14] present a day-by-day analysis of the MSN log, which covers May 2006. They show a clear drop in use over weekends and a pattern of peaking early in the week and dropping towards the end.

Shown in Figure 4 is the total number of Wikipedia requests per day seen in the proxy log. Use is clearly seasonal varying from fewer than 1,000 accesses per day in December to over 14,000 accesses per day in June and October. Unsurprisingly the peak is around the university's exam period.

It is reasonable to conclude from this seasonal access pattern that the Wikipedia forms an important part of the student study regime at the University of Otago. If this is the case then it is also reasonable to expect many of the most frequently requested pages to be related to academic study.

The 20 most frequently requested Wikipedia articles are shown in Table 2. The homepage (*Main Page*) is the most viewed Wikipedia page, being requested with more than 23 times the frequency as the next most popular page. This is as expected as many users will enter the Wikipedia via the homepage rather than typing an article's URL manually.

Column 3 shows a manual classification of the given pages into the categories *Work-Related* (*W*), *Informational* (*I*) and *Entertainment* (*E*). Of the top 20, half (10) can be considered work-related while the other half are entertainment (2) and informational (8). Most of the work-related pages are medical, reflecting the importance of the medical sciences to the University. This provides further evidence that the Wikipedia is, indeed, being used by students as an aid to their study during the exam period.

It should be noted that the classification is ad-hoc, and was arbitrarily chosen by the two authors. In particular, all medical pages in the table are classified as work-related on the assumption that these pages are mostly requested by the university's large number of medical students, rather than by people seeking medical advice. The classification of some pages is clearly ambiguous; the *Treaty of Waitangi* page could be considered informational due to the treaty's relevance to the location of the university (New Zealand), or work-related due to its potential relevance to History students.

Plotted in Figure 5 is the number of times each of the 340,477 requested articles was retrieved (ordered by frequency). There are a small number of pages requested a very large number of times. (Those articles appear to be informational pages about the Wikipedia, Wikis, New Zealand, the University of Otago, and death!) This distribution of request frequencies suggests that more useful results could come from clustering pages by subject area. We hypothesise that this would show other subject areas being looked up with comparable frequency to the medical sciences, but that those requests would be distributed among a greater number of pages, leading to their absence in Table 2.

It is not only reassuring that the Wikipedia is used for study purposes within the university, but also reassuring that it is not primarily used for smut. Spink et al. [13] provide a list of the 75 most frequently seen search terms in the Excite query log, the top 10 of which are: *and*, *of*, *sex*, *free*, *the*, *nude*, *pictures*, *in*, *university*, *pics*. It appears as though the Wikipedia is being used honourably by students.

3.2. Session Statistics

Identifying a user's session in a search engine query log has proven to be problematic because it is not clear what the user is doing between one log entry and the next. The same problem exists when looking at a proxy log such as the one used in this study, because only the user actions that result in an HTTP request are recorded.

The proxy log used in this study distinguishes users, and identifies the requested page, dates, time, etc., but not the referrer. Therefore, although it is known what was done, by whom, and when, it is not certain what a user was doing before making a particular request. Identifying a user's session under these circumstances is problematic because without the referrer it is difficult to identify the start (or end) of a session.

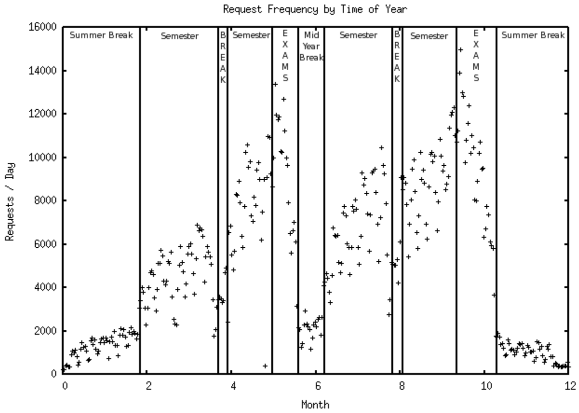

Figure 4: Access to the Wikipedia by date. Semester-times, breaks and examination periods are indicated.

Table 2: Top 20 most retrieved pages, classified as Work-Related (W), Informational (I) or Entertainment (E)

Page	Requests	Class
Main Page	75583	I
Wiki	3256	I
New Zealand	1686	I
Deaths in 2008	1315	I
University of Otago	861	I
Dunedin	859	I
Standard deviation	857	W
Wikipedia	806	I
Dopamine	669	W
Blood pressure	561	W
The Dark Knight (film)	557	E
Aldosterone	556	W
Glycolysis	546	W
Tyrosinase	541	W
Gossip Girl (TV series)	541	E
Treaty of Waitangi	516	I
Tuberculosis	514	W
Meningitis	512	W
Multiple sclerosis	511	W
HIV	510	W

He & Göker [2] define a web search session as a set of consecutive requests by a user with no longer than some time limit from one request to the next. They conclude that for web search log analysis the optimal time is between 10 and 15 minutes. There was, however, very little difference observed between the sessions produced using a time limit of 15 minutes and those produced using a time limit of 60 minutes.

It is reasonable to assume that a user navigating the Wikipedia will spend longer reading documents than a user searching the web spends reading a results list. For this reason, and for this study, a session is defined as a set of consecutive requests by the same user with a gap of no more than 60 minutes between adjacent requests. Further investigation is needed to determine whether or not this is a suitable time limit for proxy logs.

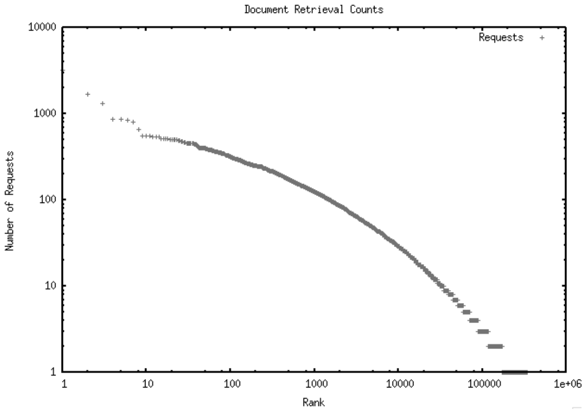

Figure 5: Number of times each document is retrieved ordered from most to least frequent.

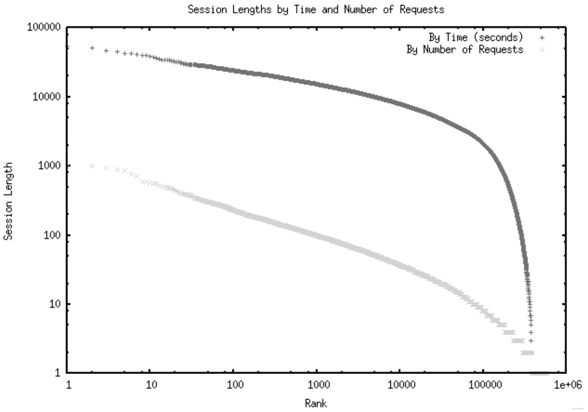

Figure 6: Session lengths ordered from longest to shortest. Sesson times in seconds and in number of clicks are both shown.

Table 3: Top 20 non-wikipedia session origins

Count	Source
2030	http://rds.yahoo.com/
1527	http://nz.wrs.yahoo.com/
1433	http://content.answers.com/
427	http://hk.wrs.yahoo.com/
203	http://s.scribd.com/
203	http://wrs.search.yahoo.co.jp/
154	http://au.wrs.yahoo.com/
149	http://mycroft.mozdev.org/
130	http://tw.wrs.yahoo.com/
129	http://sp.ask.com/
110	http://uk.wrs.yahoo.com/
82	http://www.scribd.com/
81	http://digg.com/
76	http://static.getfansub.com/
76	http://www.microsoft.com/
68	http://www.nationmaster.com/
60	http://www.apple.com/
57	http://wrs.yahoo.com/
52	http://i.ixnp.com/
52	http://pixel.quantserve.com/

Session length can be measured in several ways including the number of requests and the total time between the first and last request. In the case of a single-request session, however, the session time must be considered to be zero because it is impossible to tell how long the user spent looking at the single page that was requested.

In Figure 6 the sessions from the proxy log are shown ranked from the longest to the shortest. In total there were 577,973 sessions. When measured by time, the longest had 26 requests over 86,441 seconds (1 day and 41 seconds), and the median had 2 requests over 93 seconds. It is reasonable to conclude that the longest session is not human generated (one click an hour for a day) and so there are, in all likelihood, robots running at the university that are downloading data from the Wikipedia each hour.

When measured by number of requests, the longest had 2,340 requests over 8,550 seconds (a mean of one click every 3.76 seconds for 2 hours 22 minutes and 30 seconds), and the median had 3 requests over 93 seconds. Again it is reasonable to conclude that the longest session is not a human, but a robot.

In some cases users chose to search the Wikipedia using a search engine. In these cases they might have either added the word *Wikipedia* to their query or site-restricted their search to a *wikipedia.org* site.

Table 3 shows the top 20 non-Wikipedia site origins appearing in the log. It is important to recall that the analysed log only includes requests that contain the substring *Wikipedia* – and so this table does not truly reflect the number of sessions originating outside the Wikipedia. It is surprising that Google does not appear, but this is possibly because of Google's use of asynchronous requests for result lists on supporting browsers.

Coupling this result with the number of requests for the Wikipedia homepage leads to the conclusion that the students tend to go directly to the Wikipedia and then search, rather than using an Internet search engines to find information in the Wikipedia.

3.3. Link Statistics

The primary motivation for this investigation is the understanding of how users navigate the Wikipedia so that this knowledge may be used to improve the performance of link recommender systems.

For the purpose of this investigation a user is deemed to have clicked a link in order to retrieve an article if, within a session, there was a page requested earlier in that session that contains a link to the retrieved page.

An alternative would be to consider only the user’s most recently requested page as a potential link source, which would reduce the number of false positives. This was rejected because of anecdotal evidence that users surfing the Wikipedia have multiple pages open at once, meaning that the user’s click sequence may resemble part of a breadth-first traversal of the link graph.

For brevity, the term *click* will hereafter be used without qualification to refer to a request that is believed to have been caused by a click on a particular link. It is important to note that this information may not be accurate, and a proxy log with referrers should ideally be used in future research.

Presented in Table 4 are the top 20 most clicked links. Of particular note is the link from the homepage to Deaths in 2008. This can be directly attributed to the link "Recent Deaths" at the bottom of the "in the news" section of the homepage. Of the top 20 links, 13 are clearly work-related while 5 are entertainment and 2 are informational.

Shown in Figure 7 is the distribution of link clicks ordered from most popular to least popular. By inspection it can be seen to roughly follow a power-law distribution. Most links are clicked only once but some links are very popular.

Figure 8 shows the distribution of clicked links on a per document basis. It can be seen that of the links in a document, very few were clicked even though there are many links in the documents. This cannot be explained by the presence of "boilerplate" links such as the *What links here* link because these links are not included in the collection from which the relevant data was extracted.

Table 4: Source and target articles of the 20 most clicked links.

Source	Target	Clicks	Class
Main Page	Deaths in 2008	3092	I
NAD	Nicotinamide adenine dinucleotide	282	W
Nicotinamide adenine dinucleotide	FAD	239	W
Tyrosinase	Melanin	233	W
Lactate	Lactic acid	219	W
ADH	Vasopressin	206	W
Tyrosine	Dopamine	202	W
Heroes	Heroes (TV series)	186	E
Main Page	Wikipedia	181	I
Melanin	Melanocyte	179	W
South Park	List of South Park episodes	176	E
Gossip Girl	Gossip Girl (TV series)	174	E
Adjuvant	Immunologic adjuvant	162	W
Thiamine pyrophosphate	Pyruvate dehydrogenase	161	W
Heroes (TV series)	List of Heroes episodes	151	E
House (TV series)	List of House episodes	141	E
Systole	Systole (medicine)	133	W
Vitamin E	Tocopherol	133	W
Melanin	Melanoma	124	W
Diaphragm	Thoracic diaphragm	124	W

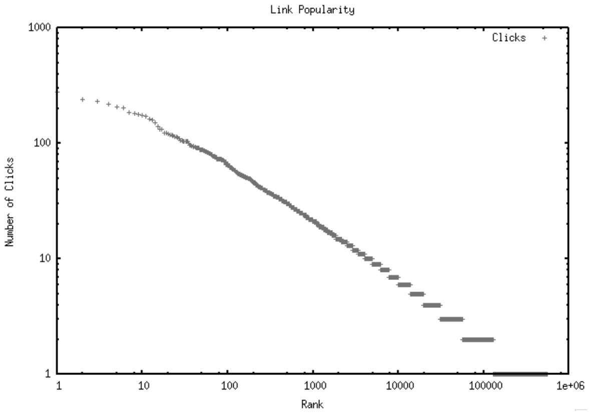

Figure 7: Frequency of use of clicked links

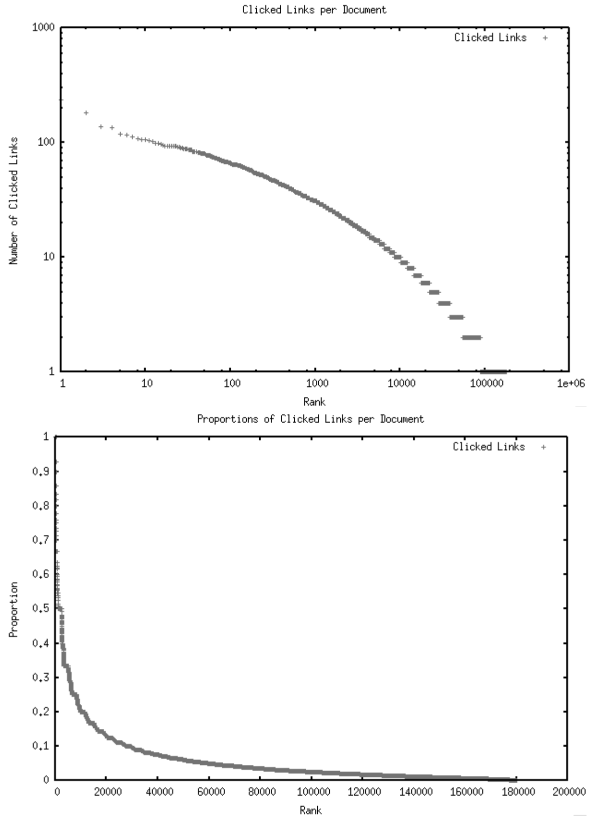

Figure 8: Number of clicked links per document by absolute count (above) and relative to the number of links in the document (bottom). In most documents only one link was clicked despite there being many links that might have been chosen.

Huang et al. [4] present the metric used in the INEX Link-the-Wiki track. It is a mean average precision (MAP) based metric which assumes that all relevant links are equally relevant. This assumption may not be valid; the users may show bias for certain links. In future work we will examine these potential biases by determining the prior probability of the click frequency distributions seen in each document. Given the already observed bias from the homepage to the recent deaths page it is reasonable to believe that some links are more popular than others. If this is the case then the appropriateness of the INEX Link-the-Wiki metrics should be examined.

6.4% of those links that are clicked are from the *See Also* section of the document whereas remaining 93.6% are from the running text. 6.4% is also the proportion of links in those documents that are *See Also* links. This suggests that there is no user preference to these links over the running text links. This is surprising because the *See Also* links are at the bottom of the page, although Fitts's Law may apply.

INEX offers two tasks in the Link-the-Wiki track: file-to-file linking, and anchor-to-BEP (best entry point) linking. In the former the task is to identify articles related to a new article to be added to the Wikipedia. This is equivalent to the task of adding *See Also* links to an article. In the latter task the link discovery system must identify anchor-texts in the running text of the new article and targets within the Wikipedia.

The discovery that running-text links appear to be as important as the *See Also* links suggests that the two INEX tasks are also equally important.

Potamias et al. [12] propose an algorithm for approximating the shortest path between two nodes in a large graph. Several hubs are chosen based on an estimate of their centrality in the graph, and a single-source shortest path calculation is performed from each hub to all nodes in the graph. The shortest path estimate for a pair of nodes is calculated by determining the length of the path between the nodes through each hub in turn, and taking the shortest of those paths.

The actual path taken in each session was computed and the lengths of the paths are shown in Figure 6. The shortest path they could have taken (from the start to the end of their session) can be estimated using the algorithm of Potamias et al. The difference is the *slack* in the session. That is, assuming the user has one information need per session and upon fulfilling it they stop using the Wikipedia, the number of wasted clicks (and consequently the amount of wasted time) can be estimated.

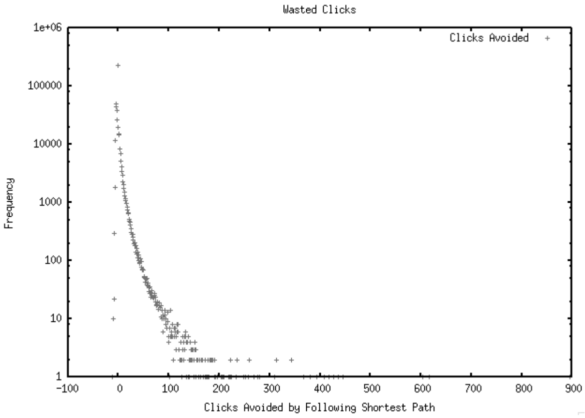

Figure 9: Number of clicks that could be saved if the user navigated the Wikipedia using the shortest path from the start of their session to the end of their session.

Figure 9 shows the difference between the actual length and the estimated shortest path for each session. Positive numbers indicate that clicks would be saved if the user had chosen the shortest path; negative numbers are due to users arriving at their destination by methods other than clicking links.

The shortest path estimation algorithm was used because of the number of sessions and the magnitude of the link-graph. It should be noted that the result is always pessimistic. It computes a number that is no smaller than the shortest path. Despite this, 83,761 (14%) user sessions would be reduced in length if the user had followed the shortest path. In 192,375 (33%) sessions the user found a path shorter than the estimated shortest path (perhaps by searching). 231,317 sessions are optimal. For the remaining 70,520 sessions no path could be found (the link graph is not strongly connected).

Assuming users are doing their utmost to find the information they seek, it is pertinent to ask why they waste so many clicks in their information seeking. Further investigation is needed; however it could be due to information overload. Given the extensive interlinking between Wikipedia articles, it may simply be too difficult to spot which links to click. If this is the case then a reduction in the size of the link graph (that is, the removal of links) may result in a better user experience. This result is in line with the manual assessment experiments of Huang et al. [5], which suggest that many of the links in the Wikipedia are not relevant. Further, since 33% of the sessions are shorter than the shortest path, it is reasonable to conclude that users' current response to viewing over-linked documents is to resort to searching.

The mean number of clicks that could be avoided if a user followed the shortest path is 0.018 clicks per session.

However, it is also possible that many of the wasted clicks seen are a result of users browsing Wikipedia for trivia, merely because they find it interesting. (For example, clicking links that go from the name of a day, month or year to a list of events that happened in that time period.) It is therefore important not to take the link-graph reduction goal to its logical conclusion by removing all trivial links, as this would diminish users' enjoyment of Wikipedia, which might in turn cause the non-trivial information content in Wikipedia to stagnate. Therefore, it is important to balance the removal of links that hinder navigation with the retention of links that, while not strictly relevant, are sometimes used and do not hinder navigation.

It is pertinent to ask whether the first document the user viewed *should have been* linked to the last document they viewed. Computing this is equivalent to solving the link discovery problem, but an estimate might be made using one of the previously published link discovery algorithms. The Itakura & Clarke [6] algorithm as implemented by Jenkinson et al. [7] is fast and might make a good candidate algorithm, as might Geva's title matching algorithm [1]. Computing the optimal link graph for the Wikipedia is left for future work.

4. Discussion and Conclusions

The University of Otago student proxy server logged all accesses to the Internet for the 2008 calendar year. From this log all accesses to the Wikipedia were extracted and analysed. In total 16,665,418 requests were made by 17,635 users.

The analysis suggests that students use the Wikipedia primarily as an encyclopaedia for study-related purposes. They typically use it for a very short period of time (a few minutes) and search from the Wikipedia rather than via an Internet search engine. They prefer to use it close to exams, and they use it at all times of the day and night.

The analysis of the link statistics suggests that there is some bias in the users' click pattern, as very few of the available links are clicked, but further work is needed to determine the nature of this bias. Users appear to click on a very small proportion of the links in a document, but there is no bias towards *See Also* or running-text links. If indeed there is bias, then it may be appropriate to re-examine the metrics used to measure the performance of link discovery systems.

On the assumption that a user is trying to fulfil one information need in each session, the amount of slack in a user session was computed. In 14% of sessions the user did not choose the shortest path from the start of their session to the end. In 33% of cases the user found a path shorter than the shortest path which suggests that the link-graph of the Wikipedia is not helping those users and they are resorting to methods other than browsing in order to find their information.

This study was conducted with the goal of improving link discovery systems such as those seen in the INEX Link-the-Wiki track. The results suggest that by removing non-useful links from the Wikipedia (simplifying the graph) the user will find it easier to browse in order to fulfil their information need, but it is important not to take this too extreme, and to remove *harmless* links merely because they are not relevant, as this would decrease the utility of the Wikipedia.

Further work might be conducted on the proxy log. Previous studies have suggested that 4-digit year links are not considered relevant by INEX assessors. The nature of the links the user clicked remains unknown, as does the nature of relevant links in the INEX assessments.

The INEX Link-the-Wiki track has two tasks. In the file-to-file task a set of randomly selected documents are chosen from the Wikipedia. The links between those documents and the Wikipedia are removed and the system must predict the links that were present. As a consequence of the Wikipedia log entries having been extracted from the full proxy log, there now exists a complete year-long log of which articles were chosen and which links were clicked. This log might be used as the source of articles for the INEX track. If the articles were chosen from those accessed in the log then performance could be measured relative to those links that were clicked.

The log might also be used in the Link-the-Wiki anchor-to-BEP task in which the link discovery system must choose anchors and target document / best entry point pairs. Although best entry points are not typically linked to in the Wikipedia, the anchor text and target document pairs can be deduced from the Proxy log using the method outlined above.

Much of this study was devoted to understanding how university students use the Wikipedia. It is heartening to see the use is generally related to their study, but disheartening to see that use is driven by the examination schedule.

5. References

[1] Geva, S., *GPX: Ad-Hoc Queries and Automated Link Discovery in the Wikipedia*. INEX 2007 pp. 404-416.

[2] Göker, A. and D. He, *Analysing Web Search Logs to Determine Session Boundaries for User-Oriented Learning*, In *Adaptive Hypermedia and Adaptive Web-Based Systems*. 2000. pp. 319-322.

[3] Huang, D.W., et al., *Overview of INEX 2007 Link the Wiki Trac.* INEX 2007 pp. 373-387.

[4] Huang, W.C., S. Geva, and A. Trotman, *Overview of INEX 2008 Link the Wiki Track*, INEX. 2008p. 314-325.

[5] Huang, W.C., A. Trotman, and S. Geva, *The Importance of Manual Assessment in Link Discovery*, SIGIR 2009

[6] Itakura, K.Y. and C.L. Clarke, *University of Waterloo at INEX2007: Adhoc and Link-the-Wiki Tracks*, INEX 2007. pp. 417-425.

[7] Jenkinson, D., K.-C. Leung, and A. Trotman, *Wikisearching and Wikilinking*, in *pre-proceedings of INEX 2008*. 2008.

[8] Kamps, J., M. Koolen, and A. Trotman, *Comparative Analysis of Clicks and Judgments for IR Evaluation*,.WSCD 2009.

[9] Metzger, M.J., A.J. Flanagin, and L. Zwarun, *College student web use, perceptions of information credibility, and verification behavior.* Comput. Educ., 2003. **41**(3):271-290.

[10] Mihalcea, R. and A. Csomai, *Wikify!: linking documents to encyclopedic knowledge*. CIKM 2007. pp. 233-242.

[11] Milne, D. and I.H. Witten, *Learning to link with wikipedia*, CIMK 2008 pp. 509-518.

[12] Potamias, M., et al., *Fast shortest path distance estimation in large networks*, CIKM 2009.

[13] Spink, A., et al., *Searching the Web: The public and their queries.* JASIST 2001. **53**(2):226-234.

[14] Zhang, Y. and A. Moffat. *Some Observations on User Search Behavior*. ADCS 2006. pp. 1-8