Searching for Language Resources on the Web: User Behaviour in the Open Language Archives Community

Baden Hughes

Department of Computer Science and Software Engineering
The University of Melbourne
Parkville VIC 3010, Australia
badenh@csse.unimelb.edu.au

Abstract

While much effort is expended in the curation of language resources, such investment is largely irrelevant if users cannot locate resources of interest. The Open Language Archives Community (OLAC) was established to define standards for the description of language resources and provide core infrastructure for a virtual digital library, thus addressing the resource discovery issue. In this paper we consider naturalistic user search behaviour in the Open Language Archives Community. Specifically, we have collected the query logs from the OLAC Search Engine over a 2 year period, collecting in excess of 1.3 million queries, in over 450K user search sessions. Subsequently we have mined these to discover user search patterns of various types, all pertaining to the discovery of language resources. A number of interesting observations can be made based on this analysis, in this paper we report on a range of properties and behaviours based on empirical evidence.

1. Introduction

While much effort is expended in the curation of language resources, such investment is largely irrelevant if users cannot locate resources of interest. The Open Language Archives Community (OLAC) (Simons and Bird, 2003) was established to define standards for the description of language resources and provide core infrastructure for a virtual digital library, thus assisting to address the resource discovery issue for this domain. The OLAC Search Engine is deployed as a web database application[1] integrated with related OLAC infrastructure for registration, validation, harvesting, and aggregation of OLAC data providers. (The detailed functionality of the search engine itself is described elsewhere (Hughes and Kamat, 2005).) Given that OLAC is the natural aggregation point for publishing catalogues of language resources (currently having over 30 data providers and over 30K language resources described), it is hoped that the findings in this paper will be of interest to major language resource publishers in understanding user search behaviour and language resource desiderata; and to service providers who are building resource discovery mechanisms specifically for language resources.

The structure of this paper is as follows: a brief introduction to OLAC is provided; followed by a description of the source data itself. From here, we consider a number of different dimensions of language resource search discovery, providing quantitative analysis for each point of interest, and subsequent discussion. Finally, we identify a number of items for future work, and draw conclusions about the nature of user search behaviour in the context of language resources.

2. Background

The Open Language Archives Community (OLAC) is a consortium of linguistic data archives, at the time of writing consisting of 32 archives and a corresponding catalogue of 29,000 language resources described by metadata. (For a more detailed description of OLAC, we refer interested readers to (Bird and Simons, 2003) and (Simons and Bird, 2003).) OLAC metadata is based on Dublin Core, with a number of extensions to the Dublin Core Metadata Set for relevant conceptual domains such as language, linguistic type, subject language, linguistic subject and linguistic role.

Derived from the model adopted within the OAI, the OLAC model has a two-tiered approach to implementation. *Data providers* are the institutional language archives which publish their XML-based metadata according to the OAI Static Repository standard. Individual archives use a variety of software to manage their catalogues internally. *Service providers* leverage the OAI Protocol for Metadata Harvesting to harvest the XML expressions of metadata catalogues. Within the OLAC community, typical practice is to aggregate these into an SQL database using the OLAC Harvestor and Aggregator. Service providers can then build services which utilise the union catalogue of OLAC metadata. The OLAC Search Engine is an example of this type of service implemented over the union OLAC catalogue.

3. Data Sources

Standard (Apache) web server access logs are available for the OLAC Search Engine. Since July 2004, these access logs have been collected on a continual basis by an automated harvesting script, resulting in the longitudinal aggregate data for the analysis presented in this paper. Using this raw data, a log parsing and statistical reporting script has been written in the Perl scripting language.

In total, the data consists of approximately 1.3 million queries in 450K user sessions collected over 500 days. This equates to approximately 2,600 queries per day, in around 900 user sessions per day. A user session on average includes 2.8 search instantiations.

The logs consist not only of search instances against the OLAC aggregator, but also click through data for specific search results. User sessions are uniquely identified, and

[1] http://www.language-archives.org/tools/search

thus non-sequential interactions by users can be assembled into composites resembling an entire search interaction. Although in the raw data information such as IP addresses, browser types, operating system preferences are included, we do not focus on such data in this paper per se. Furthermore, we have removed queries executed by search engines ('robots', which periodically interact with the search engine via a gateway service), since they do not exemplify the type of user we are interested in profiling in this paper.

4. Results

In the following section, a variety of statistics are discussed.

4.1. Archive Popularity

In 78.7% of all the queries, a specific archive is specified as the target domain of a search. Within this scope, the relative frequency of search for each archive can be seen in Figure 1.

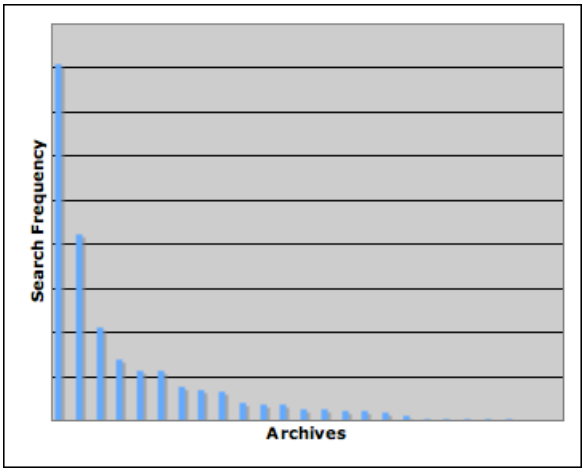

Figure 1: Archives ranked by frequency of directed queries

The SIL archive is a long established documentary archive, covering a very wide range of languages and largely holding documentary materials. As is shown later, the fact that a great variety of languages (many of which are minority languages) are searched for assists with this archive being ranked at the top of this list. What is interesting is the relative proportion of queries: clearly there are several different class of OLAC data providers based on query frequency. The first class consists solely of the SIL archive, which which is queried approximately twice as often as the next class. The second class consists solely of Paradisec; occurring twice as often as the third class which consists solely of Berkeley's SCOIL effort. It is not until much further down the list than more traditional language archives and data publishers appear.

4.2. Top Searches

Over the sample period, the top 5 searches were for: `arabic spoken language corpus`, `german morphology`, `dutch corpora`, `arabic news corpus` and `word lists`. Naturally some of these have valency given geopolitics during the collection period.

If we consider the top 100 searches, we find several interesting trends: spoken language corpora are more commonly searched for than written language corpora; multilingual corpora of both spoken and written varieties are increasingly in demand; and large minority languages are of considerable interest.

4.3. Search vs Click Through

Overall a statistically insignificantly low percentage of searches (0.04%) resulted in user click through behaviour. However, there are two possibly well founded reasons for such behaviour. The first is that many users view OLAC as a catalogue, (a list of resources), rather than as a digital library (a repository of materials). This perception is borne out by the fact that a relatively small number of the language resources found in OLAC are immediately retrievable for an end user; instead OLAC acts as the catalyst to an asynchronous process to obtain resources of interest.

4.4. Query Lengths

The overall length of queries is atypical of more generic web search environments: on average a user will supply 3.6 words per query compared with 2.x as described by other researchers in web query analysis. Unlike generic web search however, the expectation that these short queries will result in fulfilling their information need is reasonable: OLAC is a specific domain of enquiry and only has language resource information. Hence shorter queries do not necessarily imply a lower precision unlike on the web in general.

4.5. Search Operators

A small percentage of searches used various operators to constrain searches. The most dominant operators were '+' and '-', used to specify union and exclusion of query terms. The occurrence of these operators is in the order of 18% of queries, which correlates with generic web findings.

4.6. Search Syntax

Approximately 28% of searches use some form of inline syntax, a feature which is supported by the OLAC Search Engine to allow users to further constrain their queries. The most common elements expressed in a constraint are linguistic subject (allowing the selection of language resources by specific linguistic types, eg morphology) and format (allowing the selection of a particular data format).

4.7. Searching By Language

OLAC is a community of language archives, and as such it is to be expected that language will be a primary constituent in various types of search behaviour. This is held up by empirical evidence: 72What we find however is quite interesting: the top 10 languages targeted by search are a mixture of both minority and majority languages. In rank order, the most popular languages targeted by search are: English, Dutch, French, German, Arabic, Quechua, Italian, Pidgin, Greek, Toba.

4.8. Searching by Location

The OLAC Search Engine supports search by specification of a country name related to the language resource in

focus, and uses the Ethnologue as a basis for this determination. Searching by country rather than by language name accounts for just over 15% of queries in the first instance.

4.9. Related Item Search

The OLAC Search Engine has a facility to find related language resources to a given search result. These results are presented by relation types in' the form of shared metadata elements: subject, type, format etc. A small but significant (8%) of user sessions used this function to explore resources of the same type after finding an initial result through searching.

4.10. Search by Modality

Are users searching for spoken or written corpora ? The answer is not immediately distinct: about 23% of user queries specified spoken language corpora; while about 32% specified written language corpora. What is revealing however is that in considering search revision behaviour and click through behaviour, searches targeted at spoken language resources provoked a higher number of derivative queries than written language resources.

4.11. Search by Creator

OLAC's constituency has a large number of users interested in documentary linguistic work. As such, it is not suprising that a considerable number of searches include a constraining element such as `creator` (12%), which identifies the linguist who collected field data of interest.

4.12. Top Non-Existent Resources

We were able to analyse the search log to determine the most commonly requested language resources for which no OLAC record is found. We propose that the following are languages would have a considerable resources market should they become available: Greek, Quechua, Hungarian. Furthermore, there is considerable interest in highly parallel multilingual corpora, not just in written but multimodal forms.

There are many other metrics derivable from this data collection, and for reasons of space we discuss only a few of them here. Interested readers can find more detailed analysis online[2].

5. General Findings

A number of interesting observations can be made based on this analysis, vis:

- users searching for language resources typically use longer and more specific queries than those exhibited in general web search environments.
- users searching for language resources typically ignore advanced operators to constrain searches.
- users searching for language resources are far more persistent (more queries per session) than in general web search contexts.
- users searching for language resources typically constrain their searches by one or more OLAC extensions, especially data type.
- users searching for language resources are often interested in viewing results of related resources particularly by language and type.
- users searching for language resources are interested in resources for languages which could be considered minority languages much more so than the major languages typically included in published corpora.
- users searching for language resources rarely specify the format of the language resource in question (eg data type, encoding type), but typically refine their queries to include specific format requirements.
- users searching for language resources consistently use more complex metadata than that published by data providers.

6. Related Work

Empirically grounded analysis of user search behaviour in online environments has been explored in a number of contexts, although language archives and searching for language resources have never been covered specifically. In the context of general web search (Silverstein et al., 1999) and (Spink et al., 2001) have considered large scale web search logs and analysed them across various dimensions. Other researchers have considered the linguistic properties of web queries themselves (Jansen et al., 2000). In digital library environments, there have been a number of studies which consider how query logs relate to higher-order user behaviour (Jones et al., 2000).

7. Future Work

A number of items of future work can be identified given that we now have a dynamic analysis framework. We focus divide these work items into three categories: those which are embedded within the OLAC search engine user interface and oriented at end-users directly; those which are provided external to the search engine and targeted towards language resource publishers; and those which allow comparison between user search behaviour against the OLAC Search Engine with other similarly positioned services.

7.1. Search Engine Enhancements

In the OLAC Search Engine we already provide a set of prepopulated search links for queries which have zero results (eg search by language or country name variants, search by more general terms). This function should be extended to also include popular searches (eg the top 10 searches of the last month), so that users are provided with potentially useful results from a single click post-initial-query.

Similarly, we would like to enhance the graphical result display interface to include reference to prepopulated queries relevant to the initial query and the result content. Since we already have support for the display of 'similar by type' links in the search results of the OLAC Search

[2] `http://www.language-archives.org/tools/search/statistics`

Engine, this should be straightforward to implement. The result would be that an end user can see similar queries related to their initial query based on the aggregate of user search behaviour to date.

Perhaps os most interest would be the development of a series of reports similar to Google Zeitgeist[3], allowing insight to the most popular search types. Given the popularity of this feature amongst web search engines generally, we believe that it would have the desired effect of both promoting the OLAC Search Engine as well as interest in language resources more generally.

7.2. Data Publisher Enhancements

Perhaps the most obvious extenstion for data publishers is to implement a standardised search report per OLAC data provider, similar to those provided for metadata quality evaluation (Hughes, 2004). This would allow individual data providers to continually analyse search activity for language resources they publish, and compare overall end user search behaviour within OLAC against specific archive search activity.

Another area for further work is better support for the OLAC Search Engine's API. While the basic search API has been deployed and documented, there is little evidence from user log data that data publishers are promoting this mechanism for searching their own content eg by copying the form based user interface or precomposing a query string for their own archive.

7.3. Comparison and Cross-validation

Perhaps of community wide interest would be the ability to compare and cross-validate the data from the OLAC Search Engine against other similar services both within the Open Language Archives Community (eg against the search activity on the LinguistList OLAC Search[4]) and external to OLAC (eg against the search activity on MPI's IMDI digital library framework). In addition, we would like to be able to compare the referrer log data from the OLAC Search Engine against the inbound activity to data publishers catalogues - however this last function would clearly require considerable cooperation from language resource publishers.

Under the auspices of a new academic research project, We are currently obtainining access to a large web search log from Microsoft's MSN search engine, and intend to extract from the 15 million queries provided relevant language and language resource search data for further cross-comparison.

8. Conclusion

We have reported a number of metrics regarding user search behaviour in the Open Language Archives Community, based on search logs from the OLAC Search Engine. Despite this previous research in related areas, we believe this is the first paper to specifically consider user search behaviour in language archives. It is hoped that the findings in this paper will be of interest to major language resource publishers in understanding user search behaviour and language resource desiderata; and to service providers who are building resource discovery mechanisms specifically for language resources.

[3] http://www.google.com/zeitgeist

[4] http://www.linguistlist.org/olac/olac-search-advanced.html

9. References

Bird, S. and G. Simons, 2003. Extending Dublin Core Metadata to support the description and discovery of language resources. *Computing and the Humanities*, 37:375–388.

Hughes, B., 2004. Metadata quality evaluation: Experience from the Open Language Archives Community. In *Proceedings of the 7th International Conference on Asian Digital Libraries*. Springer Verlag: Berlin.

Hughes, B. and A. Kamat, 2005. A metadata search engine for digital language archives. *DLib Magazine*, 11(2).

Jansen, B.J., A. Spink, and A. Pfaff, 2000. Linguistic aspects of web queries. In *Proceedings of the American Society of Information Science*. American Society of Information Science.

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

Silverstein, C., M. Henzinger, H. Marais, and M. Moricz, 1999. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(3):6–12.

Simons, G. and S. Bird, 2003. The open language archives community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, 18:117–128.

Spink, A., D. Wolfram, B.J. Jansen, and T. Saracevic, 2001. Searching of the web: the public and their queries. *Journal of the American Society of Information Science and Technology*, 52(3):226–234.

Acknowledgements

The research in this paper has been supported by the Australian Research Council through Special Initiative (E-Research), grant number SR0567353 "An Intelligent Search Infrastructure for Language Resources on the Web".