Identifying Genealogical Content on the Web

Dallan Quass
Foundation for On-Line Genealogy
dallan@folg.org

Introduction

It is well-known that there is a lot of genealogical content available on the Web, but not well-known exactly how much content there really is, or how effective various strategies are for accessing it. The goal of this presentation is to provide an introductory characterization of the amount of genealogical content available and an initial comparison of the effectiveness of three strategies in accessing that information – using search engines, genealogy webpage directories, and a machine-learning based webpage classifier.

The presentation will include estimates for the total amount of genealogical content available on the Web and the percent of it that is easily accessible using each of the three access strategies. The results will be based upon a sample of roughly 3 million genealogically-relevant pages that have been gathered along with information gathered from various search engines and webpage directories. So far the pages have been gathered but statistics have not been calculated. They will be calculated during the next several weeks before the workshop.

How Much Information Is Out There Really?

It is relatively easy to find the number of pages containing the word "genealogy" indexed by Google. However, this number is problematic for several reasons.

1. For unknown reasons, the number of pages containing the word genealogy as reported by Google varies by millions of pages from day to day.
2. It is unknown how many genealogically-relevant pages actually contain the word "genealogy"
3. It does not take into account the "deep web" – that is, genealogical content that is behind forms.

In this section we will estimate the total number of web pages with genealogical content. We will estimate the number of distinct hosts containing those pages, and the number of distinct persons and organizations who have posted genealogical content.

We will list the number of pages containing common genealogical words and phrases as reported by several search engines from day to day. We will also report on the percentage of pages containing the word genealogy that are not actually genealogical based upon a random sample, and we will report on the percent of pages that have been classified as genealogical by a machine-learning based classifier that do not actually contain the word genealogy. Considering all of this information together we will estimate the total number of search-engine accessible web pages that are relevant to genealogy.

Finally, we will estimate the amount of genealogical information available only on the "deep web" by reporting the number of genealogical pages containing forms based upon a random sample of genealogy pages, and the average amount of information behind those forms.

How Much of It Can You Find Using a Search Engine?

In this section we will report on the key words and phrases that commonly appear on genealogical web pages. The questions we will answer are:

1. For various key words and phrases, what percent of web pages (classified as genealogical by the machine-learning based classifier) contain that word or phrase?
2. What percent of genealogical web pages do not contain any of the common key words or phrases?

The answers to these questions will help us determine the effectiveness of using search engines to access genealogical content, since the most common way to access genealogical content using search engines is to append common genealogy key words or phrases to the query. That is, when searching for genealogical content for John Smith, one typically enters the query *"John Smith" genealogy* or the query *"John Smith" "family history"* into a search engine. This section addresses which key words and phrases are most effective (high recall, non-overlapping), and how much genealogical content is not returned by any of the top five key words and phrases.

How Much of It Can You Find Using a Website Directory?

As of Feb 8, 2006, Cyndislist.com reports 250,800+ links to genealogical web pages; Linkpendium.com reports 3,257,229 links. The question is how many of these links are to distinct web pages, and more importantly, distinct hosts. In this section we will report the number of distinct web pages and hosts linked to by Cyndislist.com and Linkpendium.com, and what percentage are still "alive." The estimate of the number of distinct hosts pointed to by these directories will be compared with the estimate of the total number of distinct hosts containing genealogical content from the previous section

in order to evaluate the effectiveness, specifically the comprehensiveness, of using directories to access genealogical content.

Building a Comprehensive Database of Genealogical Web Pages

In this section we will present our findings from building a machine-learning based classifier of genealogical content. Our classifier is operating at roughly 90% recall and 90% precision. Happily, it often identifies genealogical content in foreign languages even though it was trained solely on English-language data. We will discuss the steps we took to get it to this level, and compare the effectiveness of accessing genealogical content by searching a database of classified web pages against the strategies described earlier. Finally, we will describe our plans improve the classifier.