Word-Spotting for Automatic Tag Suggestion in the BYU Historic Journals Project

Douglas J. Kennard
Brigham Young University
kennard@cs.byu.edu

Bryan S. Morse
Brigham Young University
morse@cs.byu.edu

Abstract

The BYU Historic Journals project provides a prototype repository for users to upload scanned journals, letters, and other writings of their ancestors, or to add reference information about where to find those materials. Among other things, the system allows users to add tags that identify the people written about by their corresponding PersonIDs from the new FamilySearch system. This enables users to easily search for writings referring to any of their ancestors – not only the writings by them, but also the writings about them. In this paper, we describe research that is currently in its early stages and is intended to aid users in the tagging process by automatically suggesting tags. We use word-spotting (related to automatic handwriting recognition) to offer rank-ordered lists of tag suggestions to the user based on tags that the user has already specified. We also use word-spotting to help the user search for additional occurrences of tagged words, even if those occurrences have not been tagged.

1. Introduction

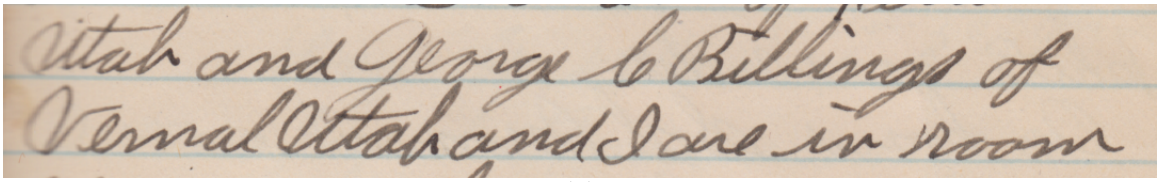
The personal writings of our ancestors (their journals, letters, etc.) help us to know and appreciate them at a much deeper level than would ever be possible by just learning genealogical facts about them such as names, dates, and places. But such writings are usually passed down (sometimes arbitrarily) from one person to another over multiple generations. Since an ancestor may have hundreds or even thousands of descendants within just a few generations, it is often very difficult to know where to find those writings, or even to know if they exist at all.

Most journals also contain writings about many *other* people, not just the person who wrote it. It is likely that many other people wrote about our ancestors, even if they didn't write about themselves. But it is usually even more difficult to find writings by other people about our ancestors than it is to find the writings of our ancestors, themselves.

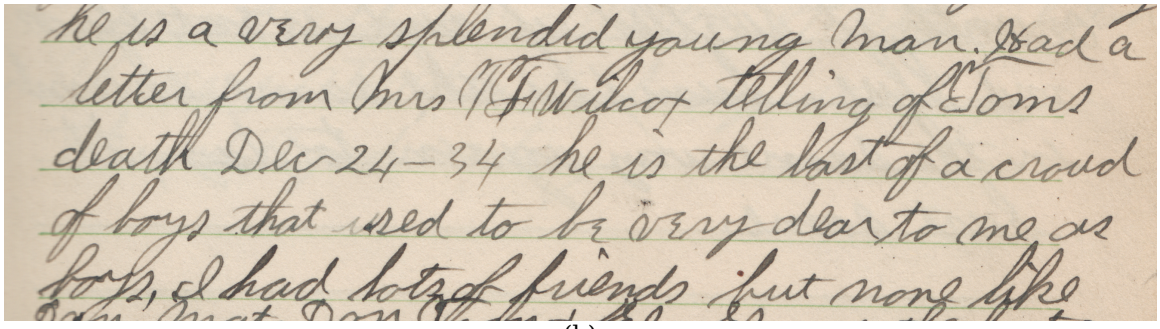
The BYU Historic Journals project seeks to provide a solution to these problems and allow users to easily contribute, find, and access writings both *by* and *about* their ancestors. The project provides a repository system to which users can upload materials or, alternatively, provide information about where the materials can be found. Users can tag materials with the PersonID identifiers from the new FamilySearch system that correspond to the person who wrote the materials. In addition, they can tag writings about other people within the materials with the PersonIDs corresponding to the people written about. Tagging writings with unambiguous identifiers in this manner allows very powerful search capability with very simple search techniques.

One of the reasons we choose to host the project at BYU instead of just developing the technology is to guarantee that policies are in place to ensure that information shared about people is appropriate, and would not be offensive or embarrassing to their descendants, nor to the subjects of the writings themselves, since we will be among them again after this life: “And that same sociality which exists among us here will exist among us there...” (D&C 130:2).

We describe the system used for the Historic Journals project – including other useful features such as direct connections, rosters, and implicit connections – in a paper recently submitted for consideration for publication in the Joint Conference on Digital Libraries (JCDL 2009) [1]. In this FHT paper, we do not focus on the system itself. Instead, we focus specifically on tools we are currently developing that are meant to aid users in the process of tagging.



(a)



(b)

Figure 1. Some references are easier to tag than others. (a) the PersonID for "George C. Billings of Vernal, Utah" is easy to look up on the new FamilySearch Website. (b) additional contextual information from later in the journal must be used to determine who Mrs. T.F. Wilcox is.

2. The Tagging Process

Although tagging journals with PersonIDs will open the door for very powerful searches to be performed, the process of tagging can take significant effort on the part of the user. While in some cases, journal authors refer to people in a straightforward manner with enough information to simply look them up on the new FamilySearch website (Figure 1a), in other cases the user must glean clues from context (Figure 1b) or use prior knowledge about the author, the author's family, and perhaps even consult other records such as church records or local newspapers to gather enough additional information to look up their PersonID to create the tag.

Since the tagging process takes significant effort anyway, we believe that tools that can in any way reduce mundane, monotonous parts of the tagging process will allow the user to be more efficient as they tag, or at least make the user's experience more pleasant.

We observe that even though many people are mentioned only once, there are also many people who are referenced many times over throughout a journal. Sometimes there are many pages between references to the same person. Due to the sheer number of people mentioned in a particular journal, a user may find it a nuisance to have to look back through a long list of tags for previous tags for that person so they can tag them again. It would be just as much of a nuisance to look the person up in FamilySearch again just to avoid looking back through a list.

We currently look at two ways of aiding the user in the mundane task of tagging the same people over and over in a journal. First, when a user is tagging a reference to a person, we automatically provide suggestions in a drop-down box. The suggestions are rank-ordered by how similar the words currently being tagged appear to be to words previously tagged.

Second, after a tag has been created for a person, we give the user the ability to search for other places in the journal that seem to have the same word(s), and might need to be tagged as the same person.

Since users can perform traditional text searches if a transcription has been provided for a particular journal, our word-spotting tools are specifically meant to be used when a complete transcription of the journal has not been provided. When transcriptions are available, a standard text search would provide the same basic functionality to the user as our second tool. The first tool could be augmented to make tag suggestions based on the text of the transcription.

3. Methods

Both of our ways of aiding users in the tagging process rely on word-spotting to look for words that seem to be similar to each other. Unlike automatic handwriting recognition for transcription purposes, word-spotting does not require algorithms that try to figure out which words are written. It is a simpler problem in which the algorithms must only try to find words that look similar to an example of the word being searched for. The user decides whether it is right.

Preliminary Processing

Our approach requires several steps to be performed after the images are uploaded, but before the tagging tools can be used:

1- Preprocessing (clean the image, find ink)
2- Segmentation (separate the lines of text, break textlines into words)
3- Compute features for each word (to use in word-spotting)
4- Save the information for use by the tagging tools

Preprocessing steps can include things like filtering out noise, removing borders and rule lines, and binarizing the images to separate ink from background.

For textline separation, we assume that lines are reasonably straight and that the spacing between lines is fairly consistent. While this is not always true, it is in the majority of the cases we see, especially for journals written in pre-printed books that are intended specifically for journal-writing.

We use a straightforward gap-metric approach to break the textlines into separate words. Such a simple method would not be accurate enough for transcription purposes, but since we are simply doing word-spotting as a tool to suggest words (instead of transcribe a document), mistakes are much more tolerable.

The features we use include whole-word features based on some that are used in word-spotting research by Rath and Manmatha [2]. They include: word profile, upper profile, lower profile, and black-to-white transition counts, each calculated after correcting for the writer's slant. The first few low-order Fourier coefficients from each of these features are placed into a feature vector. The resulting feature vector is the same length for any word, so the similarity (or difference) between any two given words can be easily and quickly computed using a simple distance metric for the two corresponding feature vectors, such as Euclidean distance.

The previous steps are performed offline as soon as images of the journal pages are uploaded. Later, when a user is actually tagging the journal, the tagging tools use the saved, precomputed features. Since the heavy computation is done offline, the tagging tools can be used at interactive rates.

Real-time Tag Suggestions

When a user selects a rectangular area to tag, an asynchronous (AJAX) request is sent to the Historic Journals server specifying the region being tagged. The server checks which words fall in that region, and returns a list of the tags that have already been created, rank-ordered by how similar their features are to the word being tagged. The browser provides the list as suggestions to the user (including both the names and PersonIDs), who can either use one of the suggestions or just add a PersonID that isn't on the suggestion list. An example of how the completed tag suggestion interface might look is shown in Figure 2.

Searching For More Occurrences of a Tag Word

For any tag that has already been created, the user can select an option to search for other occurrences of that word. When the user does this, an AJAX request tells the server which tag is being searched for. The server returns a list of where words are in the diary that have features similar to the features of the tagged word. The browser provides a result window with a link the user can click to quickly go to each location, along with an option to go ahead and tag the word with the same tag. A possible interface for the tool is depicted in Figure 3.

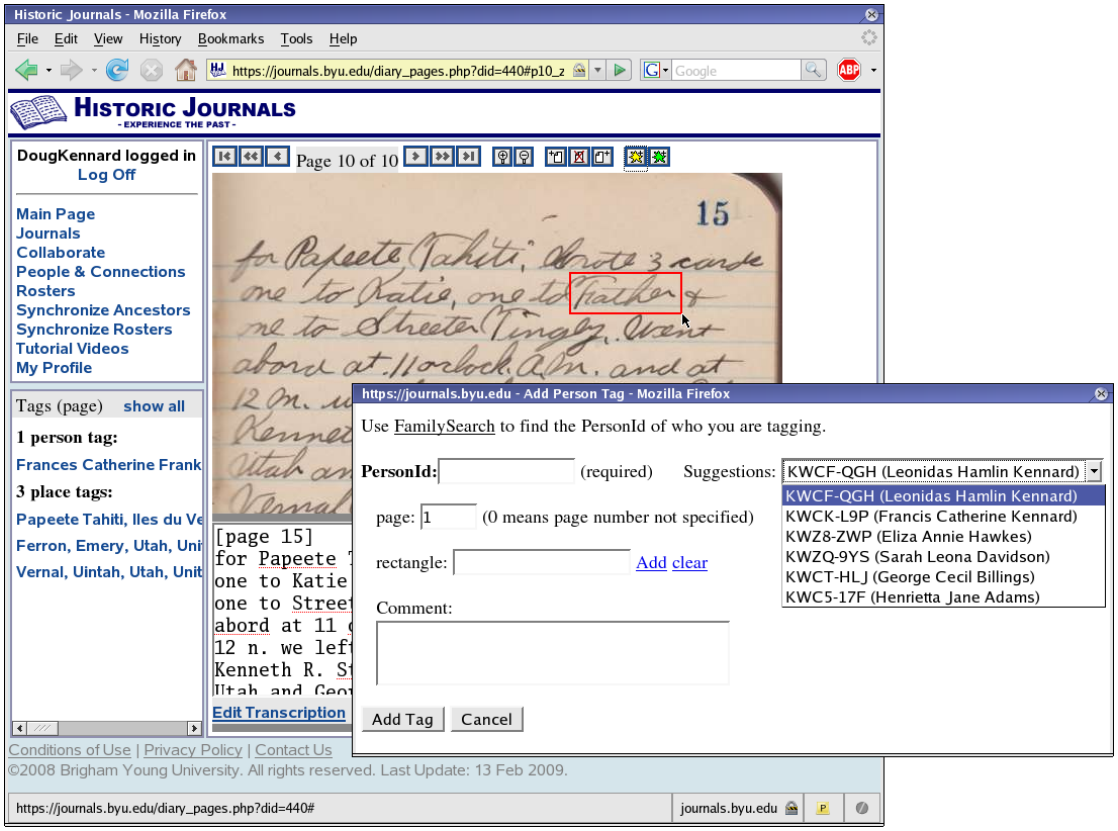


Figure 2. Mock-up of tag suggestion interface. When a rectangular region is highlighted for tagging, suggestions are made of previously tagged words, ranked in order of similarity to the current word.

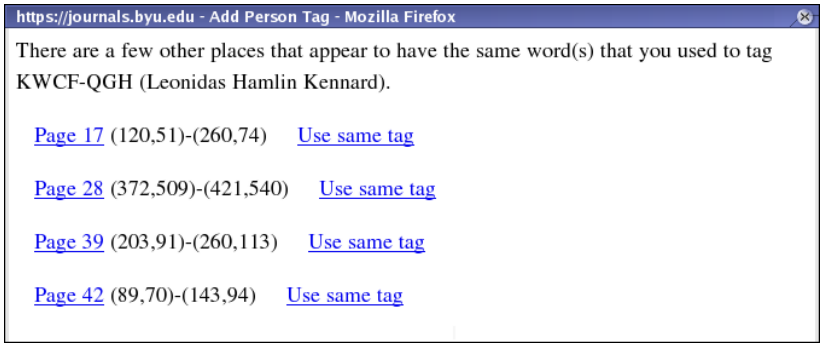


Figure 3. Possible interface for searching for more occurrences of a Tag Word.

Conclusion

We have described research currently in progress for aiding users of the BYU Historic Journals project with journal tagging. The tagging tools described provide real-time tag suggestions to users and allow them to search for more occurrences of words that have already been tagged. The tools are implemented using word-spotting algorithms found elsewhere in the literature.

References

[1] Douglas J. Kennard, William B. Lund, and Bryan S. Morse. "Improving Historical Research By Linking Digital Library Information to a Global Genealogical Database." (to appear) Joint Conference on Digital Libraries (JCDL 2009), Jun 15-19, 2009, Austin, Texas.

[2] T. M. Rath, R. Manmatha, V. Lavrenko. "A Search Engine for Historical Manuscript Images." SIGIR 04, Jul. 25-29, 2004, Sheffield, South Yorkshire, UK.