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

Douglas J. Kennard and Dr. Bryan S. Morse
(BYU Computer Science Department)
Journals, Letters, other writings

- Open the door for interest in Family History

- Bring ancestors to life

- Help us understand / appreciate them
Problem: Journal Access

Did he have a journal?

Does it still exist?

Who has it?
   (1 of 900 living descendants)

How can I read it?

Has anyone else written ABOUT him?

Did he write about others?
   (their descendants would want to know)

Do my other ancestors have journals?
BYU Historic Journals Project

Search for writings by or about ancestors

Submit Journals
Scanned Images
Transcriptions
Reference Information

Collaboration
Transcribe (if desired)
Tag names (link to PersonIDs)
Tag places (place authority)

PersonIDs (FamilySearch)
Policies to protect privacy, avoid embarrassing:
  - living descendants
  - ancestors

“And that same sociality which exists among us here will exist among us there…”

(D&C 130:2)
System Details

Joint Conference on Digital Libraries (JCDL 2009)

Demo / Q&A: today in demo session

This Presentation

Word-spotting tools to aid with tagging
The Tagging Process

Person identifier: KWCF-QGH
The Tagging Process

“George C Billings of Vernal Utah” - easy

“Mrs. TF Wilcox” - additional context
Observation

We often tag the same people many times

- look them up the PersonID again
- look back through our list of tags

(neither is difficult, but both are inconvenient)
Proposed Tools

1- Suggest previous tags (order by similarity)
Proposed Tools

2- For a tagged word, spot other occurrences
Proposed Tools

Word-spotting:

- We don’t need high accuracy  (unlike transcription)
- Just find words that look similar   (simpler problem)
- Mistakes are tolerable   (user selects)
Current Status

- Just getting started (for this application)

- Leverage code from our previous HR efforts:

Methods

Preliminary Processing (offline):

1- Preprocessing  (clean image, find ink)
2- Segmentation  (lines of text, words)
3- Feature Extraction
4- Save features for later use
1 - Preprocessing

- Clean image (filter noise)
- Remove borders, background
- Binarize image (find ink)
2 - Segmentation

- separate lines of text (profiles)
- word separation (gap metrics)

slant removal (shear)
3 - Feature Extraction

(Rath, Manmatha, Lavrenko, SIGIR 2004.)

Projection Profile
Upper Profile
Lower Profile
Transition Counts

- Treat each as a 1-D signal
- Compute Fourier Transform
- Store Low-order Fourier Coefficients
4 - Save Feature Vectors

- Save feature vector of each word

For two words, their “difference” is calculated as Euclidean Distance between their feature vectors
[page 15]

For Papeete Tahiti. Wrote 3 cards
one to Katie, one to Father &
one to Streeter Tinge. Went
aboard at 11 o'clock a.m. and at
12 n. we left the pier. We, Elders
Kenneth R. Stevans of Ferron
Utah and George C Billings of
Vernal, Uintah, Utah, Utah.

Use FamilySearch to find the PersonId of who you are tagging.

PersonId: 
page: 1  (0 means page number not specified)
rectangle:  Add clear
Comment:

Add Tag  Cancel
Conclusion

- We proposed tools to help users tag journals:
  
  Real-time Tag Suggestion
  
  Search for other occurrences of Tag Words

- The tools are based on word-spotting

- We are currently in early stages of the research

- We hope tools will increase convenience for users
Thank You