Gruzd, A. (November, 2006). Folksonomies vs. Bag-of-Words: The Evaluation & Comparison of Different Types of Document Representations. Poster. The 17th Annual SIG/CR Classification Research Workshop, ASIS&T 2006.

Folksonomies vs. Bag-of-Words:
The Evaluation & Comparison of Different Types of Document Representations

Anatoliy Gruzd
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Champaign, Illinois 61820 USA
agruzd2@uiuc.edu

Purpose

Among the factors that influence the *effectiveness* of retrieval systems, the most influential is the quality of *document representation* (*docrep*) (Lancaster, 1998). Most Internet search engines rely on *docreps* automatically extracted from web pages (commonly called *Bag-of-Words*). Unfortunately, this automatic approach often introduces *noise* (items unrelated to the page's core topic) to *docreps*. One way to reduce noise is to utilize user-created *docreps* which are less susceptible to it. Until recently, it was impractical to rely on user-created *docreps* on Internet-size collections. This all changed when online bookmarking web-services such as *citeulike.org* and *del.icio.us* started to appear. These bookmarking web-services made it easier for the vast Internet communities to collaborate and produce community-generated descriptors (known as *folksonomies*). Due to their multi-representational nature (from various community members), *folksonomies* provide retrieval systems with *docreps* that tend to be more user-oriented. With this observation in mind, I am investigating whether *folksonomies*-based retrieval systems would yield more relevant results than conventional systems.

Approach

To formally answer this question, I followed White & Griffith's (1987) methodology to determine how well descriptors from *docreps* discriminate among related (*exhaustivity level*) and unrelated (*specificity level*) documents. First, I identified clusters of documents that are related to each other by their content. Second, I counted how many descriptors on average span more than one document in the cluster (referred to here as **Count1**). Finally, I counted how many documents on average outside the cluster are spanned by the descriptors (referred to here as **Count2**). Using the average values of **Count1** and **Count2,** I compared the two different *docrep* types. To visualize and interpret results, I used Paijmans' (1993) approach to plot **Count1** and **Count2** in a manner similar to that of a precision/recall graph.

Data collection

For my pilot study, I randomly selected a relatively small collection consisting of 190 web pages out of more than 42,000 web pages tagged as '*article*' in *del.icio.us*. These '*articles*' were selected because they contain substantial amount of textual information and usually focus on a single topic. All 190 web pages were then group into seven topical clusters (academics, economy, science, etc…). Due to the small size of this pilot sample, I was able to group them manually. However, for larger collections, we will need other criteria (independent from both *foksonomies* and *Bag-of-Words*) that can be used to automatically group related web pages. Some possible candidates may include metadata generated by web pages' creators, manually created Internet subject directories (e.g. *Yahoo! Directory*), or hyperlinks found on web pages.

Findings

Interestingly, my results demonstrated that *folksonomies*-based and *Bag-of-Words*-based *docreps* yielded a similar level of *exhaustivity*. On average, the number of descriptors that span three or more documents in each cluster are higher by only 1% for *folksonomies* vs. *Bag-of-Words*. However, *folksonomies*-based *docreps* have a higher *specificity* level than *Bag-of-Words*-based *docreps*. On average, for *folksonomies*-based *docreps*, the number of documents outside the cluster that are spanned by descriptors are about 10.43% less than for *Bag-of-Words*-based *docreps*. The preliminary results from this limited study demonstrated the potential advantages of *folksonomies* vs. *Bag-of-Words*. The difference probably comes from the fact that *Bag-of-Words* tends to include more common words; however, a lager scale study is needed to make more conclusive decisions.

In sum, the tools and techniques developed in this study, the implementation of White and Griffith's methodology and Paijmans' visualization proved to be an effective toolkit to evaluate and compare *foksonomies* vs. *Bag-of-Words.*

References

Lancaster, F. W. (1998). *Indexing and Abstracting in Theory and Practice* (2nd ed.). Champaign, IL: GSLIS, University of Illinois at Urbana-Champaign.

Paijmans, H. (1993). Comparing the document representations of two IR-systems: CLARIT and TOPIC. *Journal of the American Society for Information Science, 44*(7), 383-392.

White, H. D., & Griffith, B. C. (1987). Quality of indexing in online data bases. *Information Processing & Management, 23*(3), 211-224.