Yi, K. (2014). A Semantic Similarity Approach for Linking Tweet Messages to Library of Congress Subject Headings using Linked Resources: A Pilot Study. Advances In Classification Research Online, 24(1). doi:10.7152/acro.v24i1.14676

A Semantic Similarity Approach for Linking Tweet Messages to Library of Congress Subject Headings using Linked Resources: A Pilot Study

Kwan Yi
Library Science Program
College of Education
Eastern Kentucky University
215 Combs Building, 521 Lancaster Avenue
Richmond, KY, 40475-3102 USA
Kwan.yi@eku.edu

ABSTRACT
The objective of this study is to propose, implement, and test a framework of assigning relevant Library of Congress (LC) subject headings to tweet messages. In this study, the task of assigning LC headings is considered an automatic classification task that identifies relevant LC subject headings for given tweets. The classification task is conducted in two stages. In the first stage, tweets are clustered so that similar tweets are grouped together. In the second stage, the degree of similarity between a cluster of tweets and LC subject headings is measured by a popular similarity metric, Jaccard Coefficient (JC). In this pilot study, five selected tweet clusters and nine LC subject headings were carefully chosen and used. This pilot study demonstrates a positive result for the proposed approach of identifying subject headings for tweets. In three cluster cases out of the five, JC selected the most relevant headings as the largest degrees of similarity. For the other two cases, JC was not successful in ranking the most relevant within the top three headings. In the next step, a more sophisticated clustering method will be explored and applied. Also, all possible LC subject headings will be employed to identify LC subjects for tweets in the next steps of this study

Keywords
Twitter, Library of Congress subject headings, semantic similarity, Jaccard coefficient, classification.

INTRODUCTION

As a fast-growing social media tool and a microblogging service, Twitter is a source of producing and delivering a wealth of information, with about a billion registered users and 500 million tweets per day on average (Smith, 2013). Twitter is a social networking application in which people can connect to each other and share microblogging posts called tweets. People receive/send tweets mostly from/to those who they follow or by whom they are followed. The Library of Congress (LC) decided to archive tweets for both Congress and the public, and signed an agreement with Twitter in April 2010 to gain access to all public tweets (Gross, 2013). In a white paper released in January 2013, LC explained the rationale for the agreement:

> Archiving and preserving outlets such as Twitter will enable future researchers access to a fuller picture of today's cultural norms, dialogue, trends and events to inform scholarship, the legislative process, new works of authorship, education and other purposes. (Library of Congress, 2013, p. 1)

It was reported as of January 2013 that the Library of Congress has archived all public tweets posted since 2006 and has accumulated more than 170 billion tweets totaling 133 terabytes (Library of Congress, 2013, p. 2). However, it is not yet planned when and how the archive will be available to researchers or the general public.

Researchers' interest has inclined more toward all publicly available tweets than toward tweets restricted to any individuals (Java, Song, Finin, & Tseng, 2007; Kwak, Lee, Park, & Moon, 2010; Wu, Hofman, Mason, & Watts, 2011). With the growing masses of tweets, researchers have faced the challenge of helping people use microblogging posts for their information needs (Efron, 2011). An approach to tackle the problem is to organize tweets in various ways: classifying into news and non-news using a naïve Bayes classifier trained on a corpus of tweets marked as either news or junk (Sankaranarayanan, Samet, Teitler, Lieberman, & Sperling, 2009); classifying into news, events, opinions, deals, and private messages, based on a learning model trained with eight features such as author, time-event phrases, opinioned words, etc. (Sriram, Fuhry, Demir, Ferhatosmanoglu, & Demirbas, 2010); topical clustering using a supervised

Advances in Classification Research, 2013, November 2, 2013, Montreal, QC, CANADA.

ISSN: 2324-9773

clustering method known as Rocchio classifier (Rosa, Shah, Lin, Gershman, & Fredeking, 2011) and using a topic model (Karandikar, 2010); and classifying tweets into substance, style, social, and status using a supervised learning model (Ramage, Dumais, & Liebling, 2010).

A naïve Bayes classifier and the learning models that were used in previous studies above belong to the machine learning approach. Machine learning approach is a methodology that learns from experience automatically and predicts correct decision based on the learned experience. Despite its popularity, a drawback of the approach is that it requires preparing a good quality of labeled training and test datasets that are associated with costly efforts. Beyond this great extent of effort, it can be further challenging to create such labeled datasets for a classification task with numerous classes. To deal with many classes, this study proposes to use a semantic similarity metric for the classification, which does not need any costly datasets. The task of classifying tweets poses the innate challenge of data sparseness due to the short length of the messages. Moreover, tweet messages do not often conform to standard document structure and grammatical rules, and they lack label information (Chen, Amiri, Li, & Chua, 2013). To alleviate and compensate for the sparseness issue, this study attempts to cluster similar tweet messages to treat all messages in a cluster as a whole, and also to integrate external resources to extend the short tweet message.

The objective of this study is to propose, implement, and test a framework of assigning relevant Library of Congress (LC) subject headings to tweet messages. In this study, the task of assigning LC headings is considered an automatic classification task that identifies relevant LC subject headings for given tweets. The classification task is conducted in two stages. In the first stage, tweets are clustered so that similar tweets are grouped together. In the second stage, the degree of similarity between a cluster of tweets and LC subject heading is measured by a popular similarity metric, Jaccard Coefficient (JC) that will be discussed later. In this pilot study, five selected tweet clusters and nine LC subject headings were carefully chosen and used in the experiment.

TWITTER, TWEET, AND ITS LINKED STRUCTURE

Twitter is a popular social networking application, service, and website. A Twitter user is a person or an organization that is identified by a unique user name. In Twitter, users are allowed to post a short message, which is up to 140 characters long, known as *tweet*. Tweets are also called microblog posts, in the sense that each post in Twitter is restricted to 140 characters, compared to a post in a blog that comes without any restriction in length. The short length of each tweet may influence users to make use of shortened forms of words or phrases, such as abbreviations.

There are a number of abbreviations and symbols being used in Twitter that have special meanings. RT stands for "retweet." To retweet a tweet message is to rebroadcast a tweet posted by someone else to your followers. In a retweeted message, the "RT *@username*" is often prefixed to the original message, giving credit to the original poster, where *username* is the Twitter ID of the original author, and a retweeter's opinion or comment is optionally added at the end (i.e., "- very insightful!' at the example below). An example[1] of using RT is:

Example 1:

> Original Tweet: Interview with one of the creators of Twitter! http://bit.ly/DobCk
>
> Retweeted message: RT @thepodcast Interview with one of the creators of Twitter! http://bit.ly/DobCk - very insightful!

Although a tweet is usually sent by a Twitter user to those who follow the user, it can be directly sent to any specific user by appending a '@' symbol to the intended user name (e.g., "@thepodcast" in the example above). In this case, a pair of the '@' symbol and user name precedes a text of a tweet. An example of the direct message can be:

Example 2:

> @thepodcast what services from Twooting.com?

Additionally, Twitter supports the use of *hashtags*. Tweets contain words prepended by '#' (i.e., hash symbol), which are referred to as *hashtags* (e.g., "#FF" from the example below). Hashtags are introduced in Twitter to have people place a relevant keyword or phrase (without any space within) in order to categorize their own tweets. Twitter instructs people not to over-user hashtags but to use them to identify the relevant topic of the tweet. In this way, it helps people Twitter search. Hashtags can be placed anywhere within a tweet. A tweet example[2] containing a hashtag is:

Example 3:

> I don't do #FF, ever, but I'm willing to make one exception for @origiful and @briggles, Twitter's ambassadors of quan.

A linked structure related to tweets is drawn at Figure 1. A tweet consists of content, a hashtag, and a URL; the hashtag and URL are optional in a tweet. For example, the original tweet in Example 1 above contains content and a URL (i.e., http://bit.ly/DobCk), but Example 3 has content and a hashtag (i.e., "#FF"). As shown in the figure, tweets can be linked to Internet resources on the web through URLs or hashtags within tweets. The feel lucky space in Figure 1 refers to web pages that are returned as the first Google research result when hashtags without the '#' symbol are

[1] http://www.twooting.com/members/twooting/faq/VIEW/00000011/00000011.html

[2] https://support.twitter.com/articles/49309-using-hashtags-on-twitter

ISSN: 2324-9773

submitted to the Google search engine. Also, tweets can be linked to social tags used in social websites and their associated resources through the direct match between social tags and hashtags (without the '#' symbol) in social websites.

Library of Congress (LC) subject headings and related sources are depicted in Figure 2. The three related sources are subject authority records, bibliographic records, and LibraryThing[3] records.

In this study, the following external resources related to tweets and LC subject headings are identified and used for the classification task:

- Titles, keywords, and descriptions of web pages used within tweet messages (refer to Web space in Figure 1) for Dataset A;
- Social tags from Delicious.com that are related to hashtags (i.e., terms prefixed by a hash symbol '#' in tweet messages) (refer to Social tag space in Figure 1) for Dataset A;
- Title, keyword, and description of webpages (refer to Feel lucky space in Figure 1) for Dataset A.
- Semantically related terms from LC authority file for Dataset B;
- Social tags from LibraryThing for Dataset B;
- Titles of bibliographic records for Dataset B.

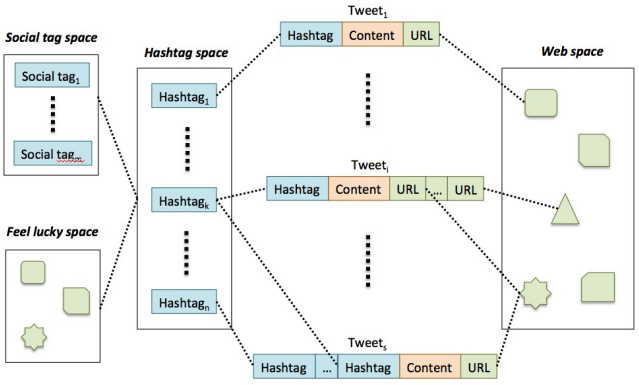

Figure 1. Linked structure related to tweet messages.

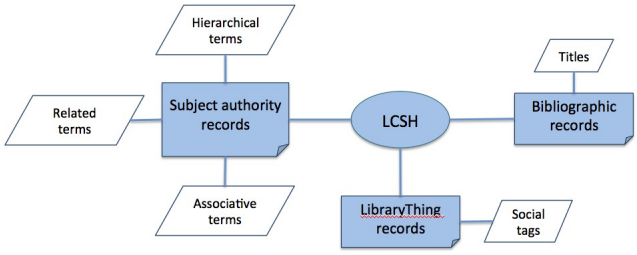

Figure 2. Resources semantically linked to LC subject headings.

SIMILARITY METRIC

The statistical similarity metric is a popular way of measuring the relative similarity of objects in semantic relatedness (Manning and Schütze, 2003). The similarity distance or degree between two objects has been applied to a wide range of text-based information organization, retrieval, process, and management applications such as text categorization (Sebastiani, 2002) and word sense disambiguation (Navigli, 2009). In this pilot study, we attempt to measure the degree of similarity between a group of tweet messages and an LC subject heading using a popular similarity metric, Jaccard coefficient. Jaccard coefficient (JC) was originally proposed as a way of measuring the level of association between documents and queries in information retrieval (IR) (van Rijsbergen, 1979). In the context of IR, documents and queries are regarded as two separate lists of terms. When entries of terms on the lists are either 0 (absence of term) or 1 (presence of term), the association between the two by JS is measured by the following expression:

$$\frac{|D \cap Q|}{|D \cup Q|}$$

where D and Q are terms in the lists for a document and a query, respectively. The numerator of the expression is equal to the number of terms common to both D and Q, and the denominator is equal to the number of terms in D plus in Q.

In our experiment, we use JC to measure the degree of the similarity between a LC subject heading and a cluster of tweet messages (see below the section METHOD for the reason of using a cluster tweet messages, instead of a single tweet). For our classification task, D and Q used in the expression above correspond to a dataset for a cluster of tweet messages and a dataset for a LC subject heading, respectively. The degree range of the similarity by JC is between 1 (the best similarity) and 0 (the least similarity).

METHOD

The ultimate goal of the experiment in this study is to identify the most relevant LC subject headings for a cluster of tweet messages. For the experiment, a dataset of 1,561,750

[3] http://www.librarything.com

ISSN: 2324-9773

tweets was collected from February 1st, 2013 to April 30th, 2013 using the Twitter Application Programming Interface. As a tweet is a short message limited to 140 characters, determining a relevant LC subject heading based on such a short message is a very difficult task and may not be feasible in some cases, due to the lack of context or message explicated within a single tweet. To deal with the challenge of short messages, we sought a way of grouping topically similar tweets into a same group. We relied on hashtags for clustering tweets. That is, tweet messages were placed in a same cluster if they shared at least one hashtag. Thus, a tweet containing multiple hashtags must appear in multiple clusters. Note that any tweet that does not hold a hashtag was excluded in this experiment.

All 1,561,750 tweets were clustered based on the hashtags listed within tweets, and a total of 156,341 clusters were produced. Figure 3 shows the sizes of clusters after the clustering process was applied. The sizes of clusters varied from the largest cluster consisting of 6,989 tweet messages to the smallest cluster with only one tweet. There was only one cluster with the largest cluster size, whereas there were 109,812 clusters (equivalent to about 70% of all clusters) with only one tweet. A best-fit trend line for the relationship between the cluster size and its frequency was drawn to evaluate the plotted data in Figure 3. The equation of the trend line and R-squared value is displayed at the top right corner of the figure. The high R-squared value indicates that the cluster size and its frequency have a power-law relationship.

Considering the size of cluster, five out of 2,829 clusters were randomly selected, each of which contains exactly five tweet messages. Hashtags associated with the selected five clusters are: #FOMC, #allergies, #Antarctica, #Islamic, and #Kentucky.

The tweets for the five clusters are as follows:

Tweets associated with #FOMC

Bank of England and Fed minutes the focus today - http://t.co/zoCgCkLu #BOE #FOMC #unemployment #forex #gbp #euro

Mar 21-2013 Thr - Econ Data -8:30 Jobless Claims -10:00 Existing Home Sales / Philly Fed -10:30 Nat Gas -#stocks #ETF - #FOMC Fed Bernanke

Love this from @aggieskitchen and @HarryandDavid #FOMC grapefruits! http://t.co/j990sxuhq2

#Risk-off: #Market sentiment rattled by #liquidation talk, #Fed minutes http://t.co/pBAbZkCe #FOMC

No Surprises From #FOMC http://t.co/vojThOodjy #Fed #Bernake #Yawn

Tweets associated with #allergies

#VitaminD found to be important in preventing food #allergies in babies] #health http://t.co/4UASsdzX7t

RT @postgradmed: Emerging #overthecounter #therapies for the treatment of the common #cold , #nasal #allergies , & #sinus #infections !! http://t.co/3R534TIf

RT @cnnhealth: Where do #allergies come from? http://t.co/iwgMOknKW3

Yay for a no pollen day! Pollen Count for 04/05/2013: Tree: absent.Grass: absent. Weed: absent. #allergies #asheville #wnc

North Texas Allergy Sufferers Facing High Pollen Counts - http://t.co/OuX0Zj36 via @CBSDFW #northtexas #allergies

Tweets associated with #Antarctica

RT @YaleE360: John Kerry urges creation of world's largest marine reserve in #Antarctica http://t.co/0gwuMfuQMa via @BloombergNews

RT @StateDept: #SecKerry: #Antarctica is the highest, coldest, windiest, driest, most pristine & most remote place on Earth. http://t.co/9kwiJjXPxe

RT @AllEarsDeb: Brian Morrow announces opening date May 24 2013 #Antarctica @SeaWorld http://t.co/u0qF1xkQ

RT @PopePolar How does the World's Saltiest Pond get its salt (and water)? http://t.co/ZMHhF5X2 #Antarctica

Meet Yeti, the South Poleís crevasse-detecting robot http://t.co/TvFjONa6Rt #ScientificMethod #Antarctica

Tweets associated with #Islamic

#Hijab #Moeslem #Girls #Islamic Check out this site: http://t.co/j42TJIL7EU

We'll c again #Islamic leaders saying 1thing but doing another. However archaic #MuslimBrotherhood r at least honest. http://t.co/alxeF2i7Zu

An Early #Islamic Estate: a Digital Initiative on #Umayyad #Syria http://t.co/QDnAEJ03 via @sharethis

Why #Islamic #Leaders Should Speak Out Against #Terror - Mohamed Hemish: http://t.co/VcnL32pqNs

#Muslim hatred of #Jews smacks of envy. My #Dawn column against rampant #anti-Semitism in the #Islamic world at http://t.co/ueBxfS8j

Tweets associated with #Kentucky

RT @nolaprep: #Kentucky Coach John Calipari visits St. Augustine's Craig Victor http://t.co/Sb0Xf5C7Wz #UK #Bigblue #nolaprep

ISSN: 2324-9773

RT @McConnellPress: McConnell Congratulates @Real_JenLaw on OSCAR Win http://t.co/6jk7jwlYlC #Kentucky #Oscars

#PurpleKnights RT @nolaprep #Kentucky Coach John Calipari visits St. Augustine's Craig Victor http://t.co/bQjKR2fZld #UK #Bigblue #nolaprep

Remember our nation's heritage. "A Visit to #CampNelson National #Cemetery" http://t.co/xiqdCNRRPa #CivilWar #Bluegrass #Kentucky

Karma...fighting to put the "student" back in "student athlete" since 2013. Have a nice off-season Mr. Calipari. #Kentucky #NITFail #NIT

In this pilot study, not all LC subject headings were employed in calculating the semantic similarity of the five clusters, due to the time factor. Instead, after the manual review of the tweets in the selected five clusters, a total of nine LC subject headings were carefully selected to fully cover the topics of the five tweet clusters: Climatology, Earth sciences, Health education, Immune system, International economic relations, Public health, Religions, Sports medicine, and Sports stories, American.

The experiment of measuring the similarities between tweet clusters and LC subject headings was conducted based on the Datasets A and B that are described in the section **TWITTER, TWEET, AND ITS LINKED STRUCTURE**. To measure the effect of the data from 'Feel Lucky space,' the similarity was measured in two different cases: one with Dataset A and the other with Dataset A that excludes the data from the 'Feel Lucky space.' The performance measured without the 'Feel Lucky space' data was used as a baseline performance called Tweet baseline, and the one with the 'Feel Lucky space' data in Dataset A is called Tweet feellucky.

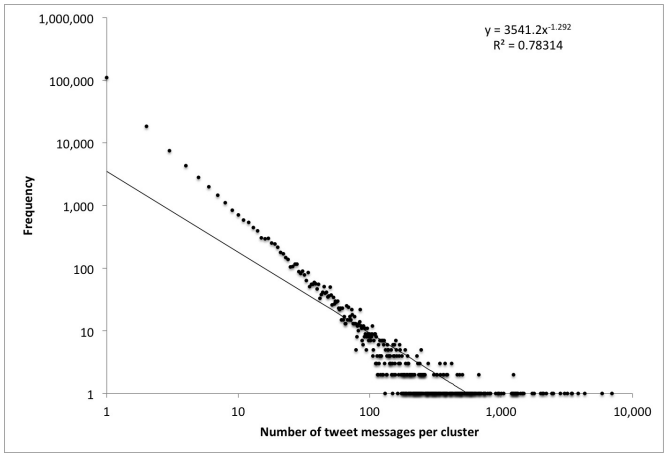

Figure 3. Result of the hashtag-based clustering.

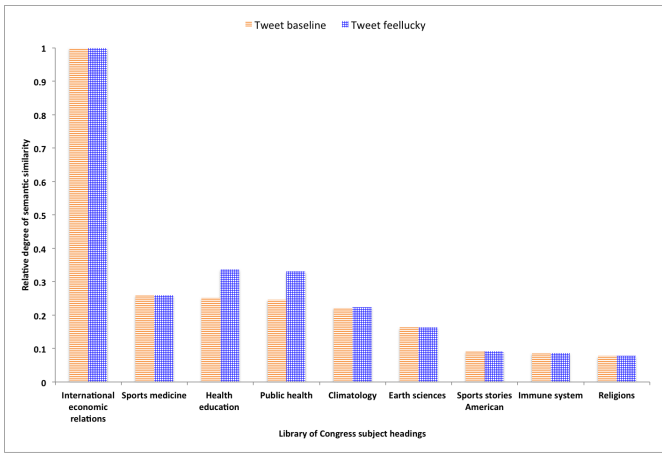

Figure 4. Semantic similarity result for the test cluster of #FOMC.

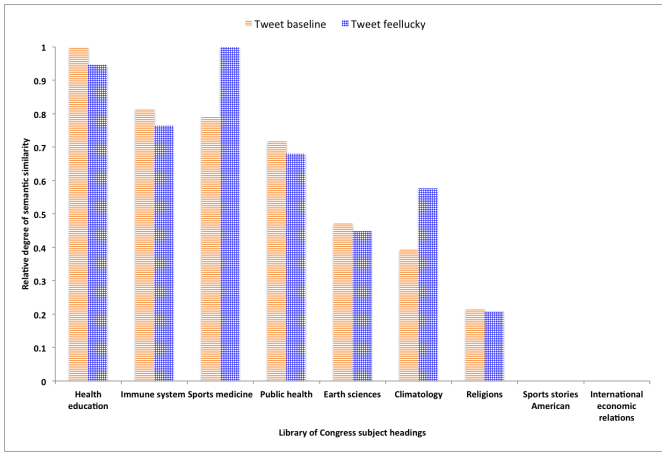

Figure 5. Semantic similarity result for the test cluster of #allergies.

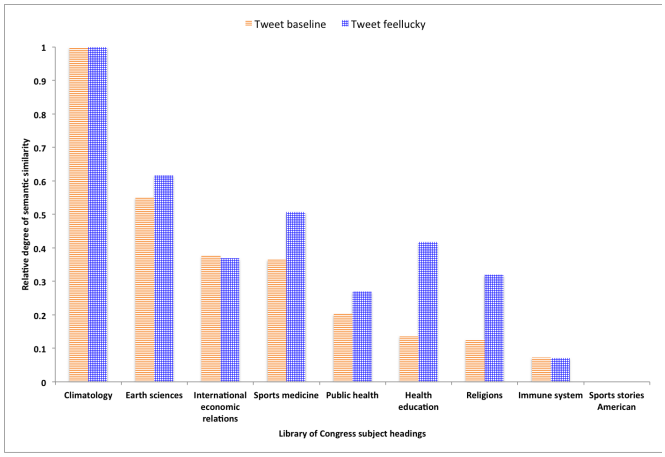

Figure 6. Semantic similarity result for the test cluster of #Antarctica.

ISSN: 2324-9773

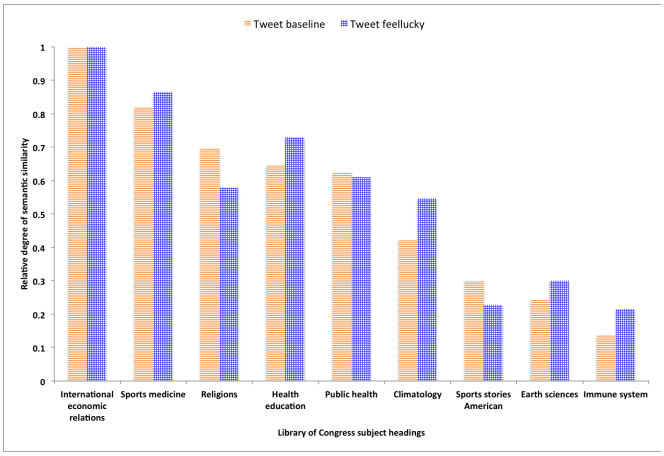

Figure 7. Semantic similarity result for the test cluster of #Islamic.

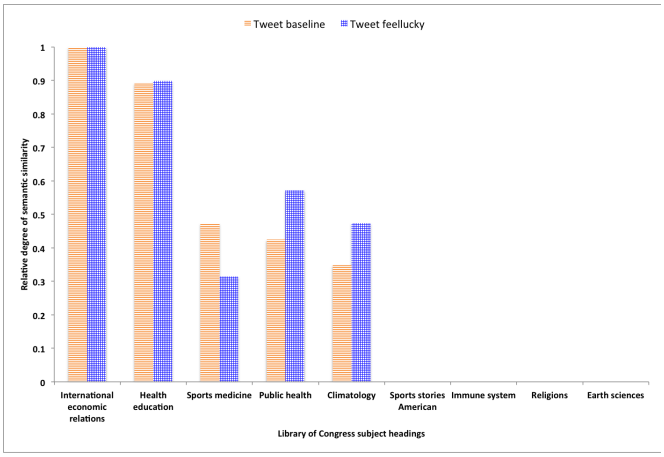

Figure 8. Semantic similarity result for the test cluster of #Kentucky.

RESULT

The similarity results with five different clusters are shown in Figures 4 through 8.

Results for the cluster associated with the hashtag #FOMC: FOMC stands for Federal Open Market Committee. In either baseline or feellucky, JC judged the LC heading 'International economic relations' with the largest degree of similarity, which is the most relevant for the cluster.

Results for the cluster associated with the hashtag #allergies: In either baseline or feellucky, all LC headings about health and climate that are associated with allergies received higher degree of similarity than headings not as closely associated with allergies. Note that the top ranked heading with the 'tweet baseline' dataset is 'health education,' but the top one with the 'tweet feellucky' is 'sports medicine.'

Results for the cluster associated with the hashtag #Antarctica: In either baseline or feellucky, the two top-ranked headings are the most relevant ones: 'climatology' and 'earth sciences.' Nevertheless, some non-relevant headings also received a considerable degree of similarity.

Results for the cluster associated with the hashtag #Islamic: In either baseline or feellucky, the most relevant heading 'Religions' was not selected as the top or second ranked heading in similarity. It may have occurred because the scope of the 'Religions' heading is excessively broad, and there is no strong evidence found in the dataset being used that connects religion and Islamic.

Results for the cluster associated with the hashtag #Kentucky: In either baseline or feellucky, the target relevant heading 'Sport stories, American' is not highly ranked in similarity. Out of the five tweet messages in the cluster, only one message is about the subject Kentucky state and the other three are about the University of Kentucky men's basketball team or the head coach.

Overall, any notable difference between the 'tweet baseline' and 'tweet feellucky' is not found. In four out of the five cluster cases, the same headings are top ranked by both 'tweet baseline' and 'tweet feellucky.'

Here are the findings of this study based on the observation of the experimental results. First, webpages of the URLs expressed within tweets appear to represent the target subject of the tweets well. Second, the subject scope of a hashtag used in a tweet tends to be broader than that of the tweet message. The inconsistency in subject between hashtag and tweet message leads to the lower quality of the hashtag-based clustering of tweet messages. Third, the Delicious 'related social tags' that were collected for target hashtags are likely to cover more diverse and broader subjects than what are covered by the hashtags. The coverage inconsistency remains problematic.

SUMMARY AND CONCLUSION

This study analyzed a big dataset (i.e., tweet messages) with the help of linked data (i.e., LCSH) and attempted to integrate the big data into the linked data. This is a pilot study of assigning relevant LC subject headings into a group of tweet messages using a popular semantic similarity metric, Jaccard Coefficient. A challenge with this task was to extract the key subject of tweets, each of which is short in length. Two approaches taken in this study for this task were clustering tweets and utilizing various external resources such as webpages, social tags, and authority/bibliographic records that are semantically related to tweet content and LC headings.

This pilot study that selected five cases demonstrated a positive result for the proposed approach. In the three cluster cases out of the five, JC selected the most relevant headings as the largest degrees of similarity. For the other two cases, JC was not successful in ranking the most relevant headings

ISSN: 2324-9773

within the top three subject headings. A plausible explanation for the unsuccessful results may be that the gap in semantics between hashtags and the subject headings considered is still too great; the key terms representing the subjects of hashtags may not be present in the datasets being used.

With the news that the Library of Congress (2013) decided to archive all public tweet messages, the need to manage and organize tweet messages has arisen, which this study contributes to. In addition, this study has a direct impact on the potential of accessing tweet messages in library catalogues (Chang and Lyer, 2012) through the LC Subject Headings, which is the most popular controlled vocabulary in academic libraries.

There are three areas to be improved by future studies. First, a simple and brute force approach of clustering tweets is used in this pilot study. As shown in the #Kentucky cluster, the five tweets in the cluster cover three different topics: University of Kentucky basketball coach John Calipari in three tweets, Kentuckian Jennifer Lawrence winning an Oscar for Best Actress as one tweet, and Kentucky itself at another one. In the next steps, a better clustering method needs to be adapted that effectively utilizes subjects of the webpages for the URLs listed in tweets since the subjects turn out to represent well the subjects of tweet messages. Second, other data sources matching the subjects of the webpages for the URLs listed within the tweets need to be identified and integrated into the automated classification process. Third, all possible LC subject headings will be employed for the future application of the proposed approach in the next steps of this study to explore issues about LC subject headings.

ACKNOWLEDGMENTS

I would like to thank Judith Yi and Marcia Rapchak for their work of reading and editing this manuscript.

REFERENCES

Chang, H.C., & Iyer, H. (2012). Trends in Twitter hashtag applications: Design features for value-added dimensions to future library catalogues. *Library Trends, 61*(1), 248-258.

Chen, Y., Amiri, H., Li, Z., & Chua, T. S. (2013, July). Emerging topic detection for organizations from microblogs. In L. Goeuriot & L. Kelly (Eds.), *Proceedings of the 36th International Conference on Research and development in Information Retrieval (ACM SIGIR 2013)* (pp. 43-52). New York, NY: ACM.

Efron, M. (2011). Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*, *62*(6), 996-1008.

Smith, C. (2013, October 17). By the numbers: 53 amazing Twitter stats. Digital Marketing Ramblings… Retrieved October 18, 2013 from http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/

Gross, D. (2013, January 7). Library of Congress digs into 170 billion tweets. Retrieved August 20, 2013 from http://www.cnn.com/2013/01/07/tech/social-media/library-congress-twitter/

Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (pp. 56-65). ACM.

Karandikar, A. (2010). *Clustering short status messages: A topic model based approach* (Master thesis, University of Maryland)

Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web* (pp. 591-600). ACM.

Library of Congress. (2013, January). Update on the Twitter archive at the Library of Congress. Retrieved October 15, 2013 from http://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf

Manning, C. D., & Schütze, H. (2003). *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys, 41*(2), 1-69.

Ramage, D., Dumais, S. T., & Liebling, D. J. (2010). Characterizing Microblogs with Topic Models. Paper presented at the Fourth International AAAI Conference on Weblogs and Social Media, Washington, DC. Paper retrieved from http://nlp.stanford.edu/~dramage/papers/twitter-icwsm10.pdf

Rosa, K. D., Shah, R., Lin, B., Gershman, A., & Frederking, R. (2011). Topical clustering of tweets. *Proceedings of the 34th International Conference on Research and Development in Information Retrieval Workshop on Social Web Search and Mining* (pp. 841-842). New York, NY: ACM.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., & Sperling, J. (2009, November). Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 42-51). ACM

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1) 1-47.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010, July). Short text classification in twitter to improve information filtering. In H.-H. Chen, E.N. Efthimiadis, J. Savoy, F. Crestani, & S. Marchand-

ISSN: 2324-9773

Maillet (Eds.), *Proceedings of the 33rd International Conference on Research and Development in Information Retrieval (ACM SIGIR 2010)* (pp. 841-842). New York, NY: ACM.

van Rijsbergen, C.J. (1979). *Information retrieval*. London: Butterworths.

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011, March). Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 705-714). ACM.

ISSN: 2324-9773