The use of additional evidence in mining user-created descriptions for content structural design

Yan Wu*

Department of Library and Information Studies, The University of the West Indies, Mona, Kingston, Jamaica

Abstract. The use of a text mining approach for full automatic taxonomy creation for content management has proven with serious limitations. The high level semantics indicating relevant association of entities among the documents are often not explored. This study introduces a feasible method that allows identifying high level semantics into text mining procedures while providing for appropriate levels of document descriptions to support access and discoverability. Due to the effectiveness of categorization and adequacy of the structure created can be better determined by humans who are familiar to the documents, qualitative inquiry rather than a purely experimental design was applied. The study collected the data and run the text mining analysis with text analysis, clustering and topic extraction. Two examples show how to develop a faceted classification structure to support digital collection access and navigation using the method. The study indicates that the text-mining method supports taxonomy creation with more efficiency and accuracy when human domain and application knowledge are captured during data collection and text mining processing. The proposed method of taxonomy creation would support the creation of new knowledge.

1 Introduction

Taxonomy or controlled word lists have been applied to online content structural design to support user browsing and search of information, leading to better user experience. With a shared terminology and structure, taxonomy can also greatly improve communication among different groups of users for knowledge creation and reuse. The resulting taxonomy or controlled word lists also become knowledge base for the domain represented by the digital collection or digital library content. They are part of valuable resource for the online collection designed.

User-generated semantics have been applied to text-mining procedures to create taxonomy [1, 2]. The user-generated semantics provide user perspectives which could be used in content structural design in support of this unique group of users' search and site navigation. This comes along with the principle of user-centred design. Many studies have

* Corresponding author: Yan.wu@uwimona.edu.jm

been carried out using data mining techniques to extract topics and subtopics from the text documents and to automatically generate taxonomy directly from the document texts [3, 4]. Using automatic generation can provide an advantage in the process of taxonomy creation, but the results from these machine learning processes often suffer low interpretability [5, 6], therefore their usefulness. In contrast, using user-created descriptions summarizing or categorizing the document content can greatly increases the quality of the dataset if the context of data collection is correctly identified and human cognition captured.

Text-mining offers several means to automatically explore semantics for knowledge discovery, by pointing out concepts, rules, and patterns from a dataset. One known approach of text-mining is based on frequent term vectors, commonly known as bag-of-words (BOW), by reducing a document to a list of unrelated terms [6]. The problem of this approach is that it is usually resulting in loss of context and implicit knowledge present in the text. Improved machine approaches include those allow word vectors mapped into a vector space at sentence-level, paragraph-level, and document-level for distributed representation [6]. One technique is clustering analysis, which represents terms based on similarities of their meanings represented by a vector of quantifiable features, with word contexts are represented in vectors [5, 6]. In clustering analysis, vectors can also be represented by co-occurrence, which has been proven effective in determining taxonomic relations [5] which is adopted in the machine learning part of the study. Analysis at a sentence-level is also deemed as more efficient in extracting useful concepts from the user-created descriptions of this study.

This paper uses two examples to show a faceted classification structure creation supported by text mining procedures. The method is through creating a dataset in a meaningful and representative context using user-created descriptions of documents and adopting it into the text mining procedures.

2 Methods

In the study, results were compared between those of topic extraction and association from data mining procedures based on user-generated content and those created manually by the same group of users based on the sample journal documents during the first case of experiment. The second example shows the comparison of data mining results between two scenarios of user document annotations for a pictorial collection. Since there are different types of taxonomy can be created for a document collection, finding the true user-document interaction context is critical to the creation of taxonomy of the chosen type. Proper user-document interaction contexts were first identified in both cases in order to collect user descriptions of the document collection to build the baseline datasets. The understanding is that the effectiveness of text mining can be improved by using additional types of evidence which apply to a more specific domain or application. Due to the effectiveness of categorization and adequacy of the structure created can be better determined by humans who are familiar to the documents, qualitative inquiry rather than a purely experimental design has been applied.

3 Results and Discussions

The first case is a taxonomy creation procedure using text-mining analysis methods based on user-created descriptions for an online newspaper collection. The users were asked to create taxonomy and summaries for the sample journal documents of the collection, and the user descriptions and taxonomy lists were the dataset. Text mining analyses were used to identify major topics and subtopics from the resulting 18000 user created words. The

following analyses were performed: topic extraction, word co-occurrence, clustering analysis and link analysis.

Topic extraction is a natural language procedure that allows stopwords removed from the text and high frequency words to be extracted. This procedure identifies a number of the most prominent terms in the dataset. The first 20 topics and subtopics retrieved were evaluated individually after run.

Word co-occurrence was then used to derive categorical keywords. Co-occurrence analysis is the counting of paired concepts within the dataset, showing the relational strength between the identified keywords. Clustering analysis was also selected as an appropriate technique. The first 50 clusters of the analysis results were inspected one by one afterwards.

With the text-mining procedures, we were able to identify topics, subtopics as well as their relations from the current collection with confidence. We created a taxonomy using bottom-up induction by analyzing category names as well as their categorical structure. The accuracy of categorical keywords, instances, relations, and contexts were evaluated one by one through comparing dendrogram and results from the link analysis.

Two broad themes were identified first, website navigation design with user concern and student life with a subtopic on sports and so on (Figure 1). Major clusters were further explored to identify relations among concepts. We also evaluated specific keywords which are not self-explanatory but provided by text mining analyses (Figure 2).

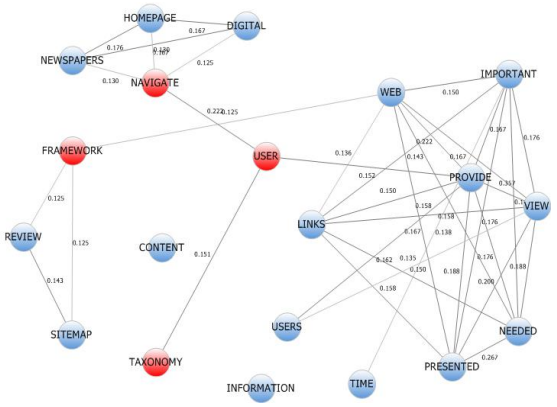

Fig. 1. Theme of user navagation showing in multidimensional scaling chart based on keywords co-occurrence mining.

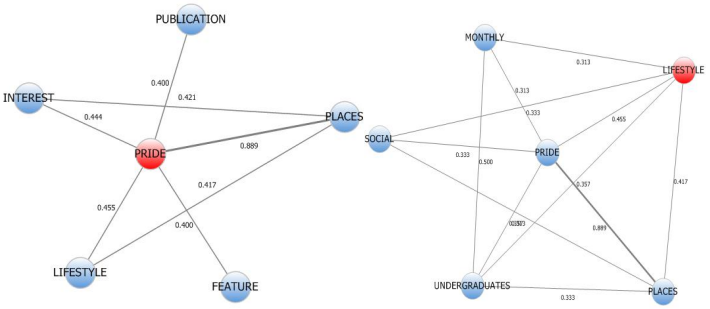

Fig. 2. ‘Pride’ and Figure 'Lifestyle’ as a major keyword.

Text-mining procedures applied successfully support a process to discover, characterize, and analyze user semantic inputs collected from a resource-based task. We have several findings. First, the results indicate that the taxonomy initially created manually by the same group users before text-mining analyses does not represent the actual users' way of viewing or accessing to the online journal collection. Using unsupervised learning by mining user-created semantics, a taxonomy is created for the sample collection. The topics include societies, lifestyle (places, pride, social, sporting, feature publication, community, and guild), dance, sports team, and pride. The topics and associations extracted from the user-generated descriptions through text-mining procedures allowed constructing a more user-friendly and meaningful taxonomy representing user interaction with the collection. With the support of text-mining analyses, a fuller list is drawn later.

Second, the results indicate that verbs are not good candidates for topic extraction in our case, and phrases are more precise presenting the patterns. For example, persons and important should be 'important persons'; digital and collections should be 'digital collections'; publication and monthly should be 'monthly publication'; and advertisement and columns should be 'advertisement columns'. Dance meeting, dramatic theatre, and so on should be preferred terms. With these findings, we manually labelled the content in order to populate the list. We used phases rather than single words when we deemed proper.

Also, the taxonomy shows that this group of users interacted with the document content as they are consumers rather than content creators, which the latters would follow a conventional classification scheme describing the content domain. This results in a more useful taxonomy reflecting online readers' perspectives. Their emotional responses are also captured.

In the second case, text-mining procedures were carried out using user descriptors of images to generate a taxonomy for an image collection. Text-mining results based on semantics from two scenarios of user document annotations are compared. In the two scenarios, the taxonomy induction that applies image annotations were involved in different levels of meaning making from the participants who were with the similar expertise level interacted with the images. This is in line with the 5-stage model of aesthetic development and Panofsky's three level of meaning-making of images [7, 8]. This method allows identifying high level semantic descriptors, so both basic level and higher level interpretations of images are captured.

The first corpus data are image descriptions created in blog format by 30 participants who gave descriptions to the individual images in the image collection. The second corpus data are image descriptions and annotations created by 12 participants who were with similar domain knowledge compared to the first group, and they carried out an image classification task to the selected images in the collection before annotations and descriptions were given to the grouped images.

Semantic clustering, link analysis, and topic extraction were carried out on the resulting 815 and 620 keywords collected from the two contexts. The extracted categories from the second dataset are much more precise and specific compared to categories extracted from the first dataset, and more adjectives were used. Basic semantic level keywords, such as "users" and "man' which are significant nouns in the first dataset are not highlighted in the second dataset. Concepts extracted in the second dataset are more concrete in their connotations. The centered categories in the second dataset are: "access", "storing", "photographed", "designed", and "decades", which are more comprehensive therefore satisfying in perspective in terms of describing the image collection.

Topics extracted from words and segmentation were set to be done by paragraph. Topic extraction on the first dataset generated 8 topics as best results after several runs. These topics are somewhat difficult to be associated to the keywords categorized therefore are less useful. The topics extracted are mainly names for persons, locations, and organizations, which are nouns describing the basic visual information depicted in images. Topic extraction

from the second dataset generated 30 topics, the first 8 are shown in table 1 compared to those from the first dataset. These topics are much easier to be interpreted, and most annotations are at higher semantic level rather than perceptual level.

Table 1. Compare the first 8 topics extracted on annotation vs. on high level meaning-making annotation of images.

N	NAME	KEYWORDS	EIGENVAL	% VAR	FREQ
1	ANN	SCHOOL; TWELVE; NICKNAMED; VANDALS; ST; RANGLIN; RAISED; PEN; PLAYED; LEAVING; ANN; INSTRUMENT; GUITAR; FATHER; ERNEST	19.35	2.42	15
2	BEATRIZ	LAMPADIA; BRAZIL; FRENCH; HASPO; NANCY; GOOGLE; YVONNE; INTERN; FELLOW; LIBRARIAN; BEATRIZ; VITAE; CLINE; DEMONSTRATES; HARVARD	18.05	2.25	15
3	BARBADOS	GOVERNOR; FOUNDING; GENERAL; BLACK; KARL; SWABY; HUGH; SULLIVAN; RALPH; HOLNESS; BARBADOS; LEO; VICE; SCOTT; CRAIG	16.46	2.06	15
4	TAGS; NATIONAL LIBRARY	TAGS; LIBRARY; JAMAICA; NATIONAL; RELATED; NB; SHOW; PROVIDE; GUIDE; CREATED; LUKE; WARM; RELATING; WATER; PRESERVATION	13.56	1.55	274
5	AGO; ATTACK	COUNTLESS; ATTACK; IDENTITY; CLEAR; YEARS; ROME; MILLION; AGO; DESTROYED; BURNED; GROUND; SARAJEVO; LIBRARIES; PEOPLE; CULTURAL	11.63	1.48	25
6	FRONT; YELLOWING	FRONT; YELLOWING; OWNED; POPULATION; ERA; HOUSES; GERMAN; BURNT; CORNER; LOWER; BRITTLE; JEWISH; VAD; YASHEEM; PAGE	10.96	1.50	40
7	AEROSOL; CONDENSED	EXTINGUISH; AEROSOL; SYSTEM; INTERVENTION; SPRINKLER; GASEOUS; CONDENSED; SYSTEMS; HUMAN; FIRES; EXAMPLES; INCLUDE; SUPPRESSION; AUTOMATIC; CONTROL	10.25	1.32	30
8	INDIVIDUAL; RESTORED	INDIVIDUAL; SHEETS; RESTORED; WASHED; CONTENT; COLOUR; PRESENTED; CHANGED; REMOVE; RACK; DRYING; ACIDIC; DRY; ITEM; PAPER	9.35	1.21	54
N	NAME	KEYWORDS	EIGENVAL	% VAR	FREQ
1	ACCESS	ACCESS; TIMELINE; DVDS; ADDED; SOUNDS; REPLAYING; EDITING; APPARENT; FORMED; REALIZED; PROCESS; IMG; INVENTIONS; MINICASSETTES; MINICASSETTE	22.77	4.65	15
2	CAPTURED	RALSTON; SOCIAL; OBJECTS; VICE; CAPTURED; CHOREOGRAPHER; JAMAICAN; CRITIC; EMERITUS; LATE; PIECES; DEDUCED; DEDICATED; SCHOLAR; MILTON	19.00	3.92	15
3	BLACK; CIVIL	PROTECTED; CIVIL; MOVEMENT; BLACK; LEADER; JUNIOR; LEARN; PHOTOGRAPH; RIGHTS; REASONS; USA; JEWELS; PEOPLE; DR; MARTIN	14.99	2.85	22
4	ANTICIPATED; CYLINDERS	OCCUR; ANTICIPATED; SUPPRESSANT; FIRES; STATE; LONG; CYLINDERS; UNALTERED; PREPARATIONS; THREATS; DISASTERS; READY; TIME; SAFETY; SHOWN	13.44	2.76	21
5	AIR; CONSERVATION	AIR; CONTROL; EFFECTIVE; MAINTENANCE; CONTRIBUTE; DOCUMENTATION; CONSERVATION; GENERAL; SCENES; QUALITY; SHELVING; RUNNING; ACTIVITIES; BUILDING; GROUPED	13.20	2.77	23
6	BAG; DEEMED	EXHIBIT; FLAG; PAN; EMBLEM; NATIONAL; STEEL; TOBAGO; BAG; TRINIDAD; TRINIDADIAN; SECTIONS; DEEMED; HOST; DISPLAYS	12.71	2.64	20
7	ACTIONS; ANYMORE	DEEPER; EASILY; LABELING; MEET; PRESENT; ANYMORE; TOOLS; ACTIONS; SIGNIFICANCE; MEANINGS; SORTED; PRODUCTS; PAST	11.96	2.47	16
8	APPOINTING; ATTENDED	ATTENDED; CAMPUSES; LEADERSHIP; ESTABLISHMENT; POSITIONS; PROCURING; PERTAIN; DOCUMENTING; TEACHERS; APPOINTING; LAND; ASPECT; CAMPUS; PERSONS; GROUPING	11.66	2.41	30

Describing images requires individuals to have certain domain knowledge, and this also affects the nature of the semantics they would create. According to Jaimes and Chang [8], user image annotations can be categorized as GenericOf, SpecificOf and About[ness], which general, specific, and abstract world knowledge is required accordingly to formulate descriptions at the conceptual levels. From our study, different tasks involved while image annotation was carried out, affected the results and quality of semantics created. The results indicate that involving individuals in a more complex task in the image annotation process, more specificof and high level abstract semantics were produced.

This paper reports the studies of the text-mining method in taxonomy creation when human domain and application knowledge are captured during data collection and processing. Our goal is to make the text-mining results based on user-created descriptions useful in taxonomy creation, in our cases, by showing us the user preferred point of view interacting with the content and with richer concept associations beyond lower-level semantics. Text mining method is based on clustering of entities supported by an underlying structure of a co-occurrence network. Text-mining analyses performed on the sample datasets show the efficiency of the suggested approach. Next, we will do iterative

analyses to tune the taxonomy using a fuller document collection in each case. We also plan to validate the taxonomy after finding a proper level of test and then test the taxonomy online.

References

1. L. Zheng, Z. Wu, J. Yin, J. Li, Y. Deng, S.WTCluster, "Utilizing tags for web services clustering," International Conference on Service Oriented Computing, pp. 204–218, (2011)
2. S. Dasgupta, S. Bhat, Y. Lee, "Taxonomic clustering of web service for efficient discovery," *Proceedings of International conference on Information and knowledge management*, pp. 1617–1620, (2010)
3. N. O. Andrews and E. A. Fox, "Recent developments in document clustering," Dept. CS, Virginia Tech, Blacksburg, VA, Tech. Rep. TR-07-35, (2007)
4. T. H. Cheng and C. P. Wei, "A clustering-based approach for integrating document-category hierarchies," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, **38**, no. 2, pp. 410–424, Mar. (2008)
5. H. Yang and J. Callan, "A metric-based framework for automatic taxonomy induction," In: *Proceedings of the Joint Conference of the ACL and the AFNLP.ACL;* 271–279, (2009)
6. Y. Bengio, H. Schwenk, J. S. Senecal, F. Morin, and J. L. Gauvain, "Neural probabilistic language models," In *Innovations in Machine Learning*, pp. 137–186, Springer, (2006)
7. K. DeSantis, A. Housen, *A Brief Guide to Developmental Theory and Aesthetic Development*, New York: Visual Understanding in Education, (2005)
8. E. Panofsky, Iconography and Iconology: An Introduction to the Study of Renaissance Art, In E. Panofsky (Ed.) *Meaning in the Visual Arts: Papers in and on Art History*, pp. 26-54, (1975)