Providing access to textual and image resources : analyzing tagging practices

Špiranec, Sonja; Ivanjko, Tomislav

This is a preprint version of a published paper. For citing purposes please use:

Špiranec, Sonja; Ivanjko, Tomislav. Providing access to textual and image resources: analyzing tagging practices // Proceedings of the 21st international BOBCATSSS conference: from collections to connections, turning libraries inside out / Taskin, Z. ; Cakmak, T. ; Dogan, G. (eds). Ankara : Hacettepe University Department of Information Management, 2014. 32-35

Providing access to textual and image resources: analyzing tagging practices

Sonja Špiranec

Faculty of Humanities and Social Sciences, University of Zagreb/Department of Information and Communication Sciences, Ivana Lučića 3, 10000 Zagreb, Croatia. Email: sspiran@ffzg.hr

Tomislav Ivanjko

Faculty of Humanities and Social Sciences, University of Zagreb/Department of Information and Communication Sciences, Ivana Lučića 3, 10000 Zagreb, Croatia. Email: tivanjko@ffzg.hr

Abstract: *The process of describing and accessing different information objects in the online environment today mostly relies on different search engine algorithms and their automatic indexing methods derived from page content and structure. While this approach can produce relevant and satisfying results for different textual information objects, the difficulty in identifying the subject matter of non-textual, imaginative, or complex materials is confessed as being one of the most challenging aspects of organizing information. In order to provide access to such materials, user generated keywords have also emerged as an alternative method in describing online sources where users associate terms (tags) to information objects thus enabling keyword based classification.*

This study explores tagging approaches when different types of information sources are being tagged (textual resource vs. image). The research included a total of 80 students assigning tags on two sample resources - a scientific paper and a sample photograph. Differences between the two resources in tag assignment were analyzed with regard to tag number and distribution, both by using descriptive statistic and by introducing the Jaccard similarity coefficient – a metrics for comparing the tag sets created by each participant with the original keywords assigned by the author itself (scientific paper) or an expert indexer (image). The results showed differences that can be very important in designing tools for accessing and describing different information objects for various heritage institutions, such as museums with primarily visual objects, or libraries, housing primarily textual collections.

Keywords: *tagging, folksonomies, tag efficiency, photographs*

Introduction

Subject analysis is the primary activity through which access to different information objects is achieved. The conceptual process of subject analysis consists of the examination of an object (or an item), determining what the object is about, and expressing this aboutness in a concise manner (Chu, 2010). As such, determining the aboutness of an object is concerned with the process or act of assigning meaning to that object and translating this meaning to others through the use of keywords and terms. The process of describing and accessing different information objects in the online environment today mostly relies on different search engine algorithms and their automatic indexing methods derived from page content and structure. The access is then enabled through a web interface that allows keyword or phrase searching that connects the user need with the relevant results. While this approach can produce relevant and satisfying results for different textual information objects, the difficulty in identifying the subject matter of an information object is confessed as being one of the most challenging aspects of organizing information, where even with the most traditional information resources, determining and identifying what an item is about can be difficult and time consuming. With non-textual, imaginative, or complex materials, the process can be even more demanding (Taylor & Joudrey, 2009). Apart from search engines that rely on algorithmic solutions, or professional indexing based on expert understanding of knowledge organization schemas and terminology in a given subject area,

This is a preprint version of a published paper. For citing purposes please use:

Špiranec, Sonja; Ivanjko, Tomislav. Providing access to textual and image resources: analyzing tagging practices // Proceedings of the 21st international BOBCATSSS conference: from collections to connections, turning libraries inside out / Taskin, Z. ; Cakmak, T. ; Dogan, G. (eds). Ankara : Hacettepe University Department of Information Management, 2014. 32-35

user generated keywords have also emerged as an alternative method in describing online sources where users associate terms (tags) to information objects thus enabling keyword based classification (Sinclair & Cardew-Hall, 2008). The use of tags can generate a "folksonomy" ("folk taxonomy"), a bottom-up, socially-generated classification that differs from the traditional top-down method of organization (Vander Wal, 2005). Although tagging can be understood as a method of indexing, according to Peters (2009), folksonomies represent a "weak" method of knowledge representation because they don't have means to express semantic relations as traditional methods such as classifications, thesauri or ontologies have. Despite these drawbacks, user generated keywords are the cornerstone of image retrieval in the online environment because the process of indexing still requires significant human involvement (Stivilla et al., 2012). Automatic analysis of indexing is still in its infancy because computers may be able to decode colors or certain patterns, while meaning or aboutness is still out of the scope of automatic approaches. A number of studies were conducted in order to determine efficiency of user based image indexing, (Matusiak, 2006; Beaudoin, 2007; Jorgensen, 2007; Trant, 2007; Marlow & Naaman, 2006; Kellog Smith, 2006). Some research studies (Heckner, Neubauer & Wolff, 2008) indicate that type of service and nature of resources trigger different tagging behavior. The studies are consistent in the conclusion that more research data is needed to demonstrate the potential benefits of user tagging in the domain of images.

Research

Research aim, sample, materials and methodology

Previous researches have shown that tagging behavior might depend on the type of information object the user tags. This research will try to identify quantitative and qualitative differences in tagging different type of resources; images and textual objects. The research included a total of 80 students from the Department of Information and Communication Sciences assigning tags on two sample resources - a scientific paper and a photograph. The sample was selected from 40 undergraduate and 40 graduate students, to ensure that prior knowledge of the subject or the principles of controlled vocabularies don't influence the results. To represent the textual information, a scientific paper from the field of information science was selected. The paper had a total of 8 pages, and was stripped of author keywords to exclude simple copy-pasting and ensure original user tags were assigned ((http://www.sciencedirect.com/science/article/pii/S1096751608000183). As for the visual resource, a photograph depicting the main Zagreb square, Trg bana Jelačića, was selected. The photograph was in black and white, and was from the 1930s. The task of the participants was to analyze each resource and assign tags to it, with a minimum of five and a maximum of ten tags.

Results

First, the two folksonomies were analyzed using descriptive statistics. The total of 80 participants generated 1144 tags, from which 612 tags (53%) was assigned to textual, and 532 tags (47%) to the visual resource. The average number of tags describing the article was 7.65, while the corresponding number for the photograph was 6.65. It is important to mention that each participant had to assign at least 5 tags, with a maximum of 10 tags.

Further analysis showed that the folksonomy for the article generated 189 different tags (31%) while the folksonomy created in describing the photograph had 211 different tags (40%). The analysis of tag frequency distribution showed that the top ten tags (with the highest frequency) account for 44% (272) of total tags in the folksonomy describing the article, while that percentage was 36% (194) for the photograph. When those values are presented on a graph, we can see that both distributions follow the power law (Figure 1).

This is a preprint version of a published paper. For citing purposes please use:

Špiranec, Sonja; Ivanjko, Tomislav. Providing access to textual and image resources: analyzing tagging practices // Proceedings of the 21st international BOBCATSSS conference: from collections to connections, turning libraries inside out / Taskin, Z. ; Cakmak, T. ; Dogan, G. (eds). Ankara : Hacettepe University Department of Information Management, 2014. 32-35

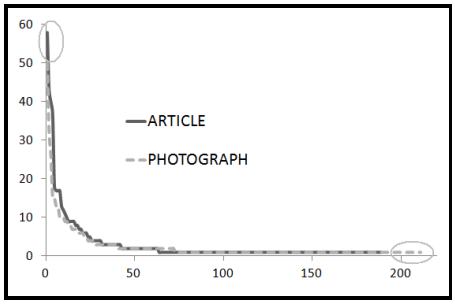

Figure 1. Tag frequency distributions compared

Although both folksonomies follow the power law curve, the comparison graph reflects differences where the article folksonomy produces a longer head and the photograph folksonomy distribution creates a longer tail. Following these results, we can conclude that participants created a more consistent folksonomy when describing the article, using fewer tags with higher frequencies. This can indicate that folksonomies describing textual resources could be more compact and produce power tags (i.e. tags with the highest frequency) that reflect the user warrant with a smaller vocabulary used.

In order to investigate how accurate participants were in expressing the subject matter of materials with tags, an investigation of semantic similarity was undertaken. Since the chosen scientific paper on which the tagging was done had author keywords assigned to it, and the photograph was also described by an expert indexer, the basis of the measure was comparing the tag sets created by each user with the original keywords assigned by the author or the expert indexer. The comparison was carried out using the Jaccard similarity coefficient equation, a statistic used for comparing the semantic similarity of sample sets as described and used in statistical NLP (Manning & Schutze, 2003) or the field of information retrieval (Peters, 2009), and applied to analyze folksonomies (Tsai, Hwang, & Tang, 2011). The basic formula for calculating the Jaccard coefficient (J*c*) is given as "J*c*=N*i*/(N*k*+N*t*-N*i*)" where Jaccard coefficient (J*c*) is defined by N*k* (original set of author keywords), N*t* (user tag set) and N*i* (number of elements in intersecting set). For example, the original author keywords {*academic integrity, authorship, citation, referencing, scholarly communication, social software, style guides, web2.0*} when compared to the chosen user tag set {*web2.0, citation, referencing, academic integrity, social bookmarking, authorship, web2.0 authoring, wiki*} had 5 tags in common {*academic integrity, authorship, citation, referencing, web2.0*}, so the Jaccard coefficient of this user tag set was 5/(8+8-5) =0.45. The measure has a range between 0 (no intersecting elements) and 1 (data sets are equal). This procedure was done for each user and generated a total of 80 Jaccard coefficients for each resource and the averaged data for each group are shown in Table 1. By using the same method, a set of power tags (top ten tags with the highest frequency) for each resource was also compared to the original author or indexer keywords.

Table 1. Jaccard similarity coefficients compared

	All tags	**average # of intersecting elements**
Article	0.21	2.5
Photograph	0.14	1.4
	Power tags	**# of intersecting elements**
Article	0.38	5
Photograph	0.23	3

When both folksonomies were compared with original author or indexer keywords, it was shown that the article folksonomy shares more common elements than that of photograph. Since many authors advocate the use of folksonomies as a complementary method of knowledge organization by using power tags extracted from folksonomies along with controlled vocabularies (Yi & Mai Chan, 2009; Mendes, Quiñonez-Skinner & Skaggs, 2009), research identified the power tags for each of the folksonomy and compared them to the expert keywords in order to see their potential in adding new value to the resources description. When comparing the number of shared

This is a preprint version of a published paper. For citing purposes please use:

Špiranec, Sonja; Ivanjko, Tomislav. Providing access to textual and image resources: analyzing tagging practices // Proceedings of the 21st international BOBCATSSS conference: from collections to connections, turning libraries inside out / Taskin, Z. ; Cakmak, T. ; Dogan, G. (eds). Ankara : Hacettepe University Department of Information Management, 2014. 32-35

elements between power tags and the folksonomy as a whole, we can see an increase in precision when only power tags are taken into consideration. The number of elements the participants share with the author or expert indexer has risen from 2.5 to 5 for the article, and from 1.4 to 3 for the photograph. Among the top ten tags with the highest frequency in the article folksonomy, five of them were the same as one of the eight author keywords, so the term overlap was 5/8 (63%). As for the photograph, three out of six terms assigned by the expert indexer were also present as a power tag, so the overlap was 3/6 (50%).

Discussion and conclusion

This research tried to explore the possible differences in tagging textual and image resources by including 80 participants in describing two different resources – a scientific article and a photograph. The results have shown that participants created a more consistent folksonomy when describing the article, using fewer tags with higher frequencies. This shows that describing textual resources was perhaps easier for the participants, because they already had an initial vocabulary given to them (the words in the article itself) so they could employ the copy-paste method even if their knowledge of the subject matter was limited. On the other hand, there were no such vocabulary indicators when dealing with the photograph, where the cognitive skills required for translating the visual signals into descriptive tags required additional processing specific for each participant, which generated a folksonomy with less shared terms and created a wider vocabulary base for including possible power tags. In order to investigate how accurate participants were in expressing the subject matter of materials with tags, an investigation of semantic similarity was undertaken, where the tag sets created by each participant were compared with the original keywords assigned by the author or the expert indexer by using the Jaccard similarity coefficient. When both folksonomies were compared with original author or indexer keywords, it was shown that the article folksonomy shared more common elements than that of photograph. Also, when comparing the number of shared elements between power tags and the folksonomy as a whole, there was an increase in precision when only power tags are taken into consideration. This again showed greater precision of the article folksonomy, where participants with little knowledge of the subject field, managed to describe the resource covering five out of eight terms the article author assigned himself.

These results can give additional indications that folksonomies created around textual resources could be more effective in covering the subject matter of the resource, while those created around visual resources could generate folksonomies that have fewer terms shared with those chosen by an expert in the field. The participants were able to express the aboutness of the article with fewer tags, and those tags were also more similar to those assigned by an expert in the field. Following this results we can conclude that participants were more efficient in tagging the textual resource and those original subject expert keywords were more efficient in ensuring the user warrant. On the other hand, the photograph folksonomy had more heterogeneous terms assigned, showing less consistent vocabulary and a smaller overlap with expert indexer terms. These results showed differences in user approach to textual and image resources that should be taken into consideration when designing tools for accessing and describing different information objects for various heritage institutions, such as museums with primarily visual objects, or libraries, housing primarily textual collections.

References

Chu, H. (2001). Research in image indexing and retrieval as reflected in the literature. *Journal of the American Society for Information Science and Technology, 52*, 1011–1018.

Jorgensen, C. (2007). Image access, the semantic gap, and social tagging as a paradigm shift. *18th Annual ASIS SIG/CR Classification Research*. Retrieved December 7, 2012 from http://journals.lib.washington.edu/index.php/acro/article/download/12868/11366

Kellog Smith, M. (2006). Viewer tagging in art museums: comparisons to concepts and vocabularies of art museum visitors. 17th ASIS SIG/CR Classification Research Workshop. Retrieved December 7, 2012 from http://journals.lib.washington.edu/index.php/acro/article/view/12492/10991.

Manning, C.D., Schütze, H., (1999). *Foundation of statistical natural language processing.* Cambridge, Ma.; London : The MIT Press

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. *Proceedings of the 17th Conference on Hypertext and Hypermedia*. Retrieved December 7, 2012 from: http://www.danah.org/papers/Hypertext2006.pdf.

This is a preprint version of a published paper. For citing purposes please use:

Špiranec, Sonja; Ivanjko, Tomislav. Providing access to textual and image resources: analyzing tagging practices // Proceedings of the 21st international BOBCATSSS conference: from collections to connections, turning libraries inside out / Taskin, Z. ; Cakmak, T. ; Dogan, G. (eds). Ankara : Hacettepe University Department of Information Management, 2014. 32-35

Matusiak, K. (2006). Towards user-centered indexing in digital image collections. *OCLC Systems & Services: International digital library perspectives, 22*, 283-298.

Mendes, L. H., Quiñonez-Skinner, J., & Skaggs, D (2009). Subjecting the catalog to tagging. *Library Hi Tech*, *27*, 30-41.

Peters, I. (2009). *Folksonomies: indexing and retrieval in Web 2.0*. Berlin: De Gruyter.

Sinclair, J., & Cardew-Hall, M. (2008). The folksonomy tag cloud: when is it useful*? Journal of Information Science*, *34*, 15-29

Stvilia, B., Jörgensen, C., & Wu, S. (2012). Establishing the value of socially created metadata to image indexing. *Library & Information Science Research*, *34*, 99-109.

Taylor, A. G., & Joudrey, D. N. (2009). *The organization of information* (3d ed.). Englewood, CO: Libraries Unlimited.

Trant, J. (2007). Exploring the potential for social tagging and folksonomy in art museums: proof of concept. *New Review of Hypermedia and Multimedia*, *12*, 83–105.

Tsai, L. C., Hwang, S. L., & Tang, K. H. (2011). Analysis of keyword-based tagging behaviors of experts and novices. *Online Information Review, 35*, 272-290.

Vander Wal, T. (2005). *Explaining and showing broad and narrow folksonomies*. Retrieved December 7, 2012 from http://www.vanderwal.net/random/entrysel.php?blog=1635.

Yi, K., & Mai Chan, L.(2009). Linking folksonomy to Library of Congress subject headings : an exploratory study. *Journal of Documentation, 65*, 872-900.