LAMP-TR-154 July 2010
CS-TR-4964
UMIACS-TR-2010-09

COMPLEMENTARITY AND SIMILARITY: RELATIONSHIPS BETWEEN TEXT-MINED TERMS AND SOCIAL TAGS FOR IMAGE DESCRIPTION

Judith L. Klavans, Hyoungtae Cho, Rebecca LaPlante

Computational Linguistics and Information Processing
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742-3275
jklavans@umd.edu, hcho5@cs.umd.edu, laplante@umd.edu

Abstract

In this paper, we present our results on comparing the language of social tags with text-mined terms for images. We have developed a novel modification of the standard term frequency/inverse document frequency metric (tf*idf) (Salton & Buckley 1988) over tags and terms to identify and filter terms which discriminate images for searchers. Since tags serve as additional input, we refer to this modification as the T-tf*idf Measure, i.e. Tags-term frequency as an inverse of document frequency, where "document" in this case refers to the either the tag or term dataset. We present the results of several variations on this measure, and demonstrate the impact on output. We discuss evaluation of our results on the ability of the metric to reflect human judgments through experiments which illustrate the value of the approach.

Keywords: text-mining, social tag, tf*idf, image description

The support of this research by the Institute of Museum and Library Services under grant LG-30-08-0117-08 is gratefully acknowledged.

Complementarity and Similarity: Relationships Between Text-Mined Terms and Social Tags for Image Description

Judith L. Klavans, Hyoungtae Cho, Rebecca LaPlante

University of Maryland, College Park

Abstract. In this paper, we present our results on comparing the language of social tags with text-mined terms for images. We have developed a novel modification of the standard term frequency/inverse document frequency metric (tf*idf) (Salton & Buckley 1988) over tags and terms to identify and filter terms which discriminate images for searchers. Since tags serve as additional input, we refer to this modification as the T-tf*idf Measure, i.e. Tags-term frequency as an inverse of document frequency, where “document” in this case refers to the either the tag or term dataset. We present the results of several variations on this measure, and demonstrate the impact on output. We discuss evaluation of our results on the ability of the metric to reflect human judgments through experiments which illustrate the value of the approach.

Keywords: text-mining, social tag, tf*idf, image description

1 Project Context

“T^3: Text, Tags, and Trust to Improve Image Access for Museums and Libraries” is a collaborative, cross-disciplinary project comprised of academic researchers, digital librarians, and museum professionals. We explore the application of techniques from computational linguistics and social tagging to the creation of linkages between the formal academic language of museums and the vernacular language of social tagging. We use text mining algorithms, taxonomies, and lexical resources to identify suggested terms and thus to aid users in tagging and retrieving images based on tags assigned from many different perspectives. We use the trust a user places in particular metadata sources, e.g. other users or other sources, to infer a weighted set of results for their searches. Consideration of these weights in ranking algorithms--along with term relationships from lexical resources – has the potential to produce high-quality, focused and personalized retrieval of works from image collections.

2 Term Weighting for Tags and Text

The focus of this paper is on working with museums and libraries for high quality use of terms and tags to identify images relevant to a user’s need, to cluster images according to similarity of terms from text and user-assigned tags used to describe them, and thus to develop new computational linguistic ranking algorithms in the process of achieving this application outcome. Whereas approaches to term weighting have been in the information retrieval (IR) and computational linguistics (CL) fields since the 1950s (Luhn 1958) and are proven to be successful in contributing to document search, comparison, and retrieval (Salton and Buckley 1988), the application of these methods to the problem of text-based image access over a combined data set of social tags combined with terms from surrounding text has not, to our knowledge, been researched. Note that text-based image access contrasts with content-based image access, which relies on properties of the image such as color, shape, texture, etc. to determine similarity and difference, and to characterize semantic features, e.g. face recognition or scene identification.

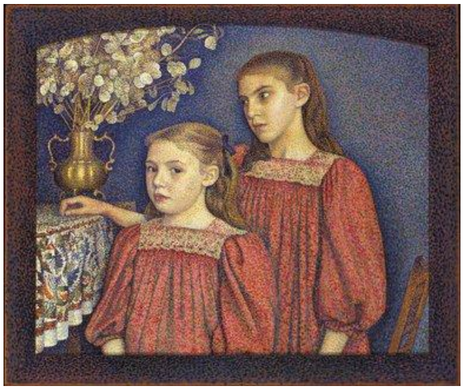
Text-based image access using text alone, without the use of social tags, has been explored, especially as a basic method before content-based image retrieval of visual image features became a more developed field (Smeulders et al. 2000). Thus, image search using proximal text is not novel (Sable, McKeown and Church 2002). However, it is well known that keywords and anchor text alone are not adequate to identify images. Traditional search over text is keyword-based and algorithms center mainly on the keyword: where and how often it (or they) might appear in the page title, on-page text, inbound anchor links, etc. Image search continues to be increasingly content-based (i.e. related to the visual features of an image, e.g. color, shape, people, objects) in general, but combined hybrid approaches using text and visual features are dominating the field.

The novel contribution of this paper is on using terms from text in combination with social tags as a new source of evidence for categorization and search. At the same time, our goal as humanities scholars is to understand the relationship between the language used to describe images in traditional ways (i.e. through descriptive text) in combination with newer ways (i.e. through social tagging and folksonomies (Bearman and Trant 2005). Thus, the underlying model of this paper relies on the abstract notions of complementarity and similarity. The principle of complementarity states that, for any reasonably complex system, the views of any two observers will be complementary, that is, it will be impossible to derive all the observations of one of the observers from the other (Easterbrook et al. 2001, Brunet et al. 2006). The principle applies whenever we have partial descriptions of the world from observers and may disappear if we ask observers to make increasingly detailed observations. Descriptions are partial by definition since filtering occurs as a result of perceptual limitations, cognitive abilities, personal values and experience, time limitations, etc. The notion of complementarity in the image description and tagging context refers to the fact that any two observers' descriptions could be redundant (if one observer's description can be reduced to the other), equivalent (if redundant in both directions), independent (if there is no overlap at all in their descriptions) or complementary, i.e. if none of the above hold. A corollary of complementarity is similarity, since words and phrases are often subtlely semantically related; determining a metric for similarity and complementarity is the larger goal of this research.

3 Experimental Paradigm

Our data set consists of social tag observations (12,600 by token, and 4,000 unique tags by type) from the steve.museum project combined with terms extracted from text (16,049 terms by token, and 4,788 unique terms by type) taken from the handbook descriptions of 165 images from the collection of the Indianapolis Museum of Art (IMA) (imamuseum.org). These tags were collected through a web interface as part of the larger "Steve Museum" project (www.steve.museum). As reported in Trant et al (2007), tags contributed directly by users might help bridge the gap between professional and public discourse by providing a source of terms not in museum documentation. The tag-term collection constitutes a novel research dataset for computational linguistics and one which we will provide to the larger community for analysis and use.

For each image, we analyzed the set of tags, and extracted terms from descriptive text. For example, consider the following image in Figure 1, where a selection of tags is given on the right and an excerpt from the IMA handbook is also provided.



vase	brass pitcher	tablecloth
flower	portrait	red dress
sister	moneyplant	boredom
face	match	double portrait
expression	dress	big sleeve
youth	chair	embroidery
pointillism	sibling rivalry	portrait

A member of the enthusiastic Belgian contingent who adopted Neo-Impressionism, Georges Lemmen wed intensity of mood with intensity of color to create a double portrait of commanding presence. The subjects, eight-year-old Jenny and twelve-year-old Berthe, were daughters of a family friend. Their penetrating gazes, typical of Lemmen's detailed, austere approach to portraiture, recall the precise likenesses of the northern Renaissance tradition. Nothing could be further from a conventionally sentimental image of childhood. (Excerpt from IMA handbook description, 2005)

Figure 1. Artist: Lemmen, Georges, Title: The Two Sisters or The Serruys Sisters, Nationality: Belgian

Our hypothesis was that tags would be more informal, whereas the text, even in a handbook, would be more formal. We anticipated little overlap, and that the overlap would be in words of certain categories, i.e. color, shape and representation (e.g. red, box, and woman). In order to test this hypothesis, we selected images of five types - Asian Sculpture, Abstract Painting, Representational Painting, Costumes, and Biblical Work - from the larger set of 165 images. We first ran the Morphy morphological analyzer (Lezius 1996) to compute raw overlap by type. Results of overlap on a random set of six images, one from each of the five types plus the image in Figure 1, is shown in Table 1 below.

Table 1. Overlap of Tags and Terms over Six Sample Images

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8
		Overlap	Tags + Terms	Tags	Terms	O/T+T	O/Tag	O/Term
Row 1	Image 1	9	91	36	64	9.9%	25.0%	14.1%
Row 2	Image 2	14	111	42	83	12.6%	33.3%	16.9%
Row 3	Image 3	13	140	82	71	9.3%	15.9%	18.3%
Row 4	Image 4	1	111	27	85	0.9%	3.7%	1.2%
Row 5	Image 5	19	121	68	72	15.7%	27.9%	26.4%
Row 6	Image 6	11	138	71	78	8.0%	15.5%	14.1%
	Average					9.4%	20.2%	15.2%

Table 1 shows that there is wide variety in the number of overlapping tags/terms, as shown in Column 2. However, the average appears to be between 9-10% over the combined tag-term set. Note that the overlap for tags alone averages 20.2% whereas the average for terms is 15.2% since the term set tends to be larger. Given this variance and given the difference in tag and term sets, in the next section of this paper, we show how we have normalized to prevent bias for images with larger tag sets or for images with longer textual descriptions. In future work, we intend to measure the variance, to examine the types of lexical items which tend to overlap, and the relationship between these items and their frequency in a large corpus. The goal will be to gain further insights on the nature of the overlapping terms and tags and the images they describe.

4 Complementarity, Similarity, and Image Description Metrics

After our initial observations, our goal was to identify which of these terms might be useful to characterize the individual distinguishing features of images, and thus perhaps be helpful to users in differentiating images. We compared this task to the information retrieval task of determining which documents might be relevant to a given query by using basic metrics such as tf*idf or, in future work, Latent Dirichlet Allocation (Blei et al. 2003) or Latent Semantic Indexing (Deerwester et al. 1990). The notion of term frequency (tf) has a long history, starting with early research on summarization (Luhn 1958) to the exploration of inverse document frequency (idf) (Spärck Jones 1972), to relevance weighting (Salton & Buckley 1988), leading to many variations to characterize topic and thus determine relevance to a query.

The original contribution of this paper is in exploring a new variation of tf*idf which is able to incorporate the notion of tags added to term frequency to measure the ranking of a term/tag pair vis a vis the overall vocabulary of the document set. For this paper, "document set" refers to two types of documents, where one type refers to the terms extracted from a piece of text related to an image, and the other type refers to the set of tags for that same image. Thus, in Column 2 of Table 2, each of these definitions of "document set" is clarified to show which lexical items are included. For example, in Row 1, the "document set" refers to the tag set alone. In Row 3, the "document set" refers to the text-mined terms alone. In Rows 2, 4-6, the "document set" includes both the tags and terms. Six scores were computed in order to compare the similarities and differences between tags and terms, and to gain an understanding of the impact of each on characterizing an image. These scores are outlined in Table 2.

Table 2. Computational Variations of T-tf*idf

	T-tf*idf Type	Data used for term frequency(tf), D_{tf}	Data set used for inverse document frequency (idf), D_{idf}	Metric
Row 1	T1	Tag	tag set	
Row 2	T2	Tag	tag set + text-mined terms	
Row 3	T3	text-mined term	text-mined terms	
Row 4	T4	text-mined term	tag set + text-mined terms	
Row 5	T5	Tag + text-mined term	tag set + text-mined terms	Arithmetic mean
Row 6	T6	Tag + text-mined term	tag set + text-mined terms	Harmonic mean

We experimented with six different variations to explore which would correlate with human judgments. We have labeled these six variations as T1-T6 in order to refer to them each independently. Term frequency (tf) can be computed as:

$$\mathrm{tf}(t_i)=\frac{N(t_i)}{\sum_{i}^{n} N(t_i)}\ , \tag{1}$$

where t_i is a term, $N(t_i)$ is the frequency of the term, $\sum_{i}^{n} N(t_i)$ is the normalization factor to indicate the frequency of all the terms in a particular document. The count is normalized to prevent bias for images with larger tag sets or for images with longer textual descriptions. We varied the formula to obtain various term frequencies. The data set used for inverse document frequency is also used for normalization of term frequency as follows:

$$\mathrm{tf}(t_i)=\frac{N(t_i)}{\sum_{i}^{n} N(t_i)}=\frac{N(D_{tf}(t_i))}{\sum_{i}^{n} N(D_{idf}(t_i))}\ , \tag{2}$$

where $D_{tf}(t_i)$ refers to the criteria in column 2 of Table 2 and where $D_{idf}(t_i)$ is the data set used for inverse document frequency depending on the criteria in column 3 of Table 2. The normalized tag frequency by tag denominator is defined as:

$$\text{tag-frequency}(t_i)=\frac{\text{Frequency of tag for one image}}{\text{Sum of \# all the tags for the same image}}\ . \tag{3}$$

The normalized tag frequency by tag/term denominator (represented with the –c suffix for "complemented") is defined as:

$$\text{tag-frequency-c}(t_i)=\frac{\text{Frequency of tag for one image}}{\text{Sum of \# all the tags+terms for the same image}}\ . \tag{4}$$

In Table 2, Row 2 shows that the denominator consists of tags plus terms. Thus, compared to Row 1, which has tag-frequency(t_i) alone in the denominator, Row 2 adds term frequencies to tag-frequency in the denominator. The normalized term frequency by term denominator is defined as:

$$\text{term-frequency}(t_i)=\frac{\text{Frequency of term for one image}}{\text{Sum of \# all the terms for the same image}}\ . \tag{5}$$

The normalized term frequency by the tag/term denominator is defined as.

$$\text{term-frequency-c}(t_i)=\frac{\text{Frequency of term for one image}}{\text{Sum of \# all the tags+terms for the same image}}\ . \tag{6}$$

In (6) as in (4), term-frequency-c(t_i) means that tag-frequency is complemented with tag frequencies in the denominator.

In Row 5 of Table 2, we present a normalized tag/term frequency by tag/term denominator using an arithmetic mean:

$$\text{tag/term-frequency}(t_i)=\frac{\text{tag-freqency}(t_i)+\text{term-frequency}(t_i)}{2} \quad . \tag{7}$$

In contrast to (7) for the arithmetic mean, the normalized tag/term frequency by tag/term denominator using a harmonic mean is shown in (8):

$$\text{tag/term-frequency-h}(t_i)=\frac{2\times\text{tag-freqency}(t_i)\times\text{term-frequency}(t_i)}{\text{tag-freqency}(t_i)+\text{term-frequency}(t_i)} \quad . \tag{8}$$

The tag/term-frequency-h(t_i) means that the average of tag and term frequency is computed by harmonic mean which is the same formula as F-score, while the tag/term-frequency(t_i) uses an arithmetic mean. This is shown in (9). The interesting point in (9) is that if either of tag-frequency(t_i) or term-frequency(t_i) is zero, tag/term-frequency-h(t_i) becomes zero. In other words, this frequency measure provides positive term frequencies for overlapped types in tags and text-minded terms; zeros for others. For example, in Figure 1, Lemmen's 'The Two Sisters or The Serruys Sisters', the type 'girl' appears 10 times in tags and there are 124 tags for that image by token. Thus, as shown in (3) above, the tag-frequency('girl') = 10/124 = .08. As shown in (5), the term frequency is zero since there is no term 'girl' for this image. The numerator in T5 (9) is thus .08 + 0. The arithmetic mean is thus .04. For the same image, the numerator in T6 is 2 * .08 * 0 (as shown in (8)), and the denominator is .08 + 0, producing a 0 result for T6 (9).

$$\begin{aligned}&\text{T5 (Table 2): term/tag-frequency}(\ t_i\ )=\frac{0.08+0.00}{2}=0.04\\&\text{T6 (Table 2): term/tag-frequency-h}(\ t_i\ )=\frac{2\times0.08\times0.00}{(0.08+0.00)}=0\end{aligned} \tag{9}$$

In the same manner, we have an inverse document frequency formula as following:

$$\text{idf}(t_i)=\log_2\left(\frac{|Doc|+1}{|\{doc: t_i\in doc \text{ of } D_{idf}\}|+1}\right) \tag{10}$$

The number of documents, $|Doc|$, remains the same over various formulas, since we consider 165 image sets as documents for both tags and text-mined terms. The denominator of the formula indicates the number of documents consisting of D_{idf}, which is defined in Table 2, where t_i appears. We add 1 both for nominator and for denominator to avoid zero-division.

We apply the formula above into each case by defining tag-idf(t_i), term-idf(t_i), and tag/term-idf(t_i) as follows:

$$\text{tag-idf}(t_i)=\log_2\left(\frac{165+1}{|\{doc: t_i\in doc \text{ of tagset}\}|+1}\right) \text{ for T-tf*idf} - \text{T1} \tag{11}$$

$$\text{term-idf}(t_i)=\log_2\left(\frac{165+1}{|\{doc: t_i\in doc \text{ of terms}\}|+1}\right) \text{ for T-tf*idf} - \text{T3} \tag{12}$$

$$\text{tag/term-idf}(t_i)=\log_2\left(\frac{165+1}{|\{doc: t_i\in doc \text{ of tagset+terms}\}|+1}\right) \text{ for T-tf*idf} - \text{T2, T4, T5, and T6.} \tag{13}$$

5 Results

Note that T1 and T2 reflect a greater weighting for tags over terms since T1 uses only tags without the larger denominator from adding text-mined terms, whereas T2 increments the denominator by using tags and terms together. As shown in Table 3, the results in Columns 1 and 2 are only slightly different; the reason for this is that text-mined terms tend to occur in the same number of documents as do tags in tag clouds. Through exploration of each of the six permutations, it has become evident that the differences between T1 and T2 are minimal, as are the differences between T3 and T4. The reason for this is that T1 and T2 add weight to words that appear in tag sets, whereas T3 and T4 both emphasize term sets. Thus, through the full computational analysis, we have been able to simplify comparisons by eliminating two of the metrics which are redundant. In future work, we will determine a principled reason for selecting between T1 and T2, and between T3 and T4. As shown in Table 2, the tf*idf scores of T2 place more weight on the tags that appear mostly in the tag set by considering terms as an element of document for idf, as compared with T1. To take a concrete example, observe the rankings of the tf*idf scores of tags in Lemmen's 'The Two Sisters or The Serruys Sisters' from Figure 1. The ranks by tf*idf score of the tags in the T1 and T2 methods shown in Table 2, above, and their inverse document frequency in terms of text-mined terms are shown in Table 3. Underlined items are those occurring both as tags and terms.

Table 3: Rank of tag by tf*idf scores using the T1 and T2 computations

Tag	Rank in T1	Rank in T2	# of documents the tag appears in tag clouds	# of documents the tag appears in text-mined terms	Increase/decrease
girl	1	1	8	3	-163%
sister	2	2	2	1	-150%
pointillism	3	3	4	0	-200%
dot	4	4	6	0	-200%
dress	5	5	17	0	-200%
portrait	6	7	18	4	-178%
tablecloth	7	6	5	0	-200%
plant	8	8	5	3	-140%
brass	9	9	7	1	-186%
chair	10	10	9	1	-189%
same	11	17	2	8	200%
money	12	13	2	3	-50%
vase	13	12	11	3	-173%
two	14	11	13	0	-200%

As shown in Table 3, the rank of the tag, 'same' that appears frequently in other image descriptions for text-mined terms is lower, since the T2 method also takes account of how meaningful a tag is among text-mined terms with regard to term specificity. This example demonstrates the goal of this research in formulating a metric to eliminate less "meaningful" terms both across a general vocabulary (e.g. the word 'same' has a higher frequency in a corpus) than a term such as 'pointillism'), as computed across a specialized vocabulary and then compared to a general vocabulary.

Table 4 shows how, for the same image, we observe the difference between T3 and T4 methods using the same observation of results from their formulas. Table 4 shows the impact of frequency of occurrence for terms in tag clouds compared with frequency in text. For example, compare the rank of terms such as, 'portrait', 'old' and 'tablecloth' that often appear in tag clouds compared with their frequencies in text. Not surprisingly, the top ranked item is the creator, the artist, Lemmon. The most general term, 'portrait' is ranked third (3) for T3 whereas it is ranked sixth (6) by T4. Interestingly, 'tablecloth' is ranked thirteenth (13) by T3 in comparison with a ranking of (32) by T4. This type of divergence is not common, as noted in Table 4 by the underlining in the Table. As we evaluate with user input, we will utilize human judgments to determine how to account for these differences, and possibly how to combine the intuitions from T1/T2 and T3/T4 into a single metric which reflects the impact and significance of each term or tag. Through these methods, described in the future work section, we will gain insights into the categorization of tags and terms, as seen through thesaural and other classifications (e.g. Panofsky 1972, discussed below).

Table 4: Rank of terms by tf*idf scores using the T3 and T4 computations

Text-mined Terms	Rank in T3	Rank in T4	# of documents the term appears in text-mined terms	# of documents the term appears in tag clouds	increase/decrease
Lemmen	1	1	1	0	-100%
Neo	2	2	4	0	-100%
portrait	3	6	11	11	0%
likeness	4	3	3	0	-100%
intensity	5	5	3	1	-67%
Belgian	6	4	3	0	-100%
impressionism	7	8	7	4	-43%
hue	8	7	9	1	-89%
old	9	38	13	21	62%
tendril	11	9	1	0	-100%
tendency	12	10	1	0	-100%
tablecloth	13	32	1	4	300%
rigorous	14	12	1	0	-100%

In order to show how lexical items are ranked according to each of the six variations on T-tf*idf, Table 5 provides the output. First, starting from T6, we observe that the overlapping items consist of 11 lexical items, where lexical item could be a tag or term; in T6, of course, each lexical item occurs as a tag and a term. Note that several morphological operations will need to be modified to optimize these results. For example, the compound noun 'moneyplant' which appears in the image in Figure 1, and which is commonly used to symbolize sincerity and honesty, has been tokenized into 'money' and 'plant', and then counted as two separate and independent tokens, which they are not. This is a case which has an obvious correct answer, unlike the issue of morphological analysis applied to the terms 'impressionism' and 'impressionist', which could be incorrectly normalized to 'impression' thus creating an inaccurate count for the specialized technique terms regarding the impressionist style, which is morphology related but semantically unrelated to 'impression'.

Table 5: Top 15 ranks over six T-tf*idf methods

Rank	T1	T2	T3	T4	T5	T6
1	girl	girl	Lemmen	Lemmen	girl	portrait
2	sister	sister	neo	neo	Lemmen	Lemmen
3	pointillism	pointillism	portrait	likeness	sister	tablecloth
4	dot	dot	likeness	Belgian	pointillism	plant
5	dress	dress	intensity	intensity	neo	brass
6	portrait	tablecloth	Belgian	portrait	portrait	money
7	tablecloth	portrait	impressionism	hue	dot	frame
8	plant	plant	hue	impressionism	tablecloth	double
9	brass	brass	old	impressionist	plant	vase
10	chair	chair	impressionist	tendril	brass	complementary
11	same	two	tendril	tendency	dress	colors
12	money	vase	tendency	rigorous	money	N/A
13	vase	money	tablecloth	portraiture	vase	N/A
14	two	youth	rigorous	Jenny	frame	N/A
15	youth	frame	portraiture	dialogue	likeness	N/A

Similarly, the prefix ‘Neo’ in Figures 4 and 5, which was extracted from ‘Neo-Impressionism’ shown in the text describing the image in Figure 1, is tokenized independently of ‘Impressionism’ which results in a somewhat skewed count both for ‘Neo-’ and for ‘Impressionism’. This type of morphological analysis and tokenization, although a basic operation, is an ongoing difficult challenge across applications (Klavans and Tzoukermann 1992).

6 Evaluation and Validation

While these tf*idf scores show the frequency of a tag or term within a given document set, can they accurately predict the real usefulness or importance of a term for a user? To answer this question, we have run a formative evaluation with three subjects to compare the value a person places on the perceived usefulness of a given tag or term to distinguish an image to the value of that same tag or term as derived from the T-tf*idf calculation to test our methods. Now that we have generated the results from the six T-tf*idf variations, we are positioned to run these independent validation studies. A sample of the survey tool used is provided in Figure 2 and selected results are provided in Table 6.

1 useful ← → 5 not useful

1 useful ← → 5 not useful

1	2	3	4	5	Word or Phrase	1	2	3	4	5	Word or Phrase
□	□	□	□	□	friend	□	□	□	□	□	girl
□	□	□	□	□	money	□	□	□	□	□	approach
□	□	□	□	□	renaissance	□	□	□	□	□	Lemmen
□	□	□	□	□	vase	□	□	□	□	□	double
□	□	□	□	□	family	□	□	□	□	□	division
□	□	□	□	□	big	□	□	□	□	□	tree
□	□	□	□	□	Lemmen	□	□	□	□	□	drape
□	□	□	□	□	artist	□	□	□	□	□	likeness
□	□	□	□	□	impressionism	□	□	□	□	□	year
□	□	□	□	□	match	□	□	□	□	□	plant
□	□	□	□	□	childhood	□	□	□	□	□	tablecloth

Figure 2. Ranking experiment of tags and terms by human subjects for ‘The Two Sisters or The Serruys Sisters’

We first needed to address two questions for the user evaluation design: (1) we needed to determine how the tags and terms associated with each image would be represented, and (2) we needed to identify a manageable set of information for a human evaluator to review. The data set which we have processed to use in the next phase of this project includes the original set of tags and terms extracted from text, with single words and multi-word phrases. This set of tags and terms had been modified for use in this experiment to tokenize multiword phrases into their component parts, resulting in the phrase 'eight-year-old' being represented as three separate tokens, 'eight', 'year', and 'old'. Thus, the counts for terms and tags often result in phrases being considered as independent words, as discussed above for 'money plant'. In order to maintain consistency between human judgments and their application to evaluation of the weighting metrics, we did not retro-correct and recombine phrases such as 'eight-year-old'; rather, we left them split to reflect the performance of the automatic tokenization. In future work, we will explore the impact of this decision on overall results.

In addition, the full set of images and related tags and text is too large to reasonably expect a human subject to evaluate and rate the 'usefulness' of all terms and tags for each image. As shown above in Table 1, there is an average of approximately 120 tags/terms per image. To eliminate overloading our subjects and thus potentially invalidating our results, we have selected one image from each of the five image groups discussed in Section 3 and one additional image (that in Figure 1) for a total of six images to use in the evaluation. However, even with this reduction, the volume of the tags and terms associated with the six images was still too large for human subjects. Thus, we took a random sampling of the tags and terms for each image to end with a test set of 15 tags and 15 terms for a total of 30 tags and terms for each user to evaluate for each image. The random sampling was created by isolating the set of normalized, decoupled tags and sorting the set by frequency, with duplicates removed. Proper names which had been incorrectly changed to lower case were manually corrected. Lastly, we did not change the spelling of any tag or term that we felt may have been misspelled because of the difficulty in determining the correct word intended by the individual who originally assigned the tag or wrote the text that included the term.

Table 6 shows our initial results for the formative evaluation. Note that Subject 1 (S1) consistently ranked all items lower than Subject 2 (S2); similarly for S3, who was consistently lower than S2.

Table 6. Rank of Tags and Terms over One Image

Terms				Tags			
	S1	S2	S3		S1	S2	S3
childhood	3	2	5	money	5	4	1
likeness	5	4	3	eyes	2	3	5
year	5	5	1	match	4	3	4
double	5	3	4	red	2	3	5
Lemmen	5	1	5	dress	2	3	5
family	4	2	4	double	3	3	5
friend	4	4	2	big	5	5	1
approach	5	4	1	tablecloth	4	4	2
presence	5	4	4	dry	5	5	1
impressionist	3	4	4	tree	5	5	1
division	5	5	1	vase	4	3	2
renaissance	3	4	3	girl	3	1	4
impressionism	3	4	4	drape	4	4	1
artist	2	3	3	sleeve	3	3	4
sentimental	3	3	4	plant	5	5	1

We illustrate in Table 6 the kind of data we have collected to explore ways to match human judgments to those of our T-Tf*idf metrics. Note that the highest ranked items for tags are 'girl', 'vase', and 'drape', whereas for terms, the highest ranked is 'artist'. In this small sample set, no single item received two top rankings of "1", although many received more than one "5", indicating that it may be more difficult to achieve popularity in ranking as opposed to skepticism. Of course, a much larger sample across many images with more subjects will enable us to draw more solid conclusions, and will permit us to accurately link human judgments to the output of our metrics. It seems, at first glance that the results in T1/T2 appear to match human judgments although more data is needed. This is our next step for future work on this data.

7 Conclusions

In this paper, we present a novel metric for examining the role of social tags and text-mined terms for images, as part of a research project involving text-based image access in museums and libraries. Using computational linguistic techniques to extract and normalize terms from text, we present our first results on overlap, complementarity, and similarity. We have developed the T-tf*idf Measure, i.e. Tags-term frequency as an inverse of document frequency. Our results from evaluation of variations on this metric have been compared with human judgments to determine agreement or differences with our findings.

In future research we will pursue three directions: first, we intend to utilize thesauri to classify the types of terms. Specifically, the Art and Architecture Thesaurus (AAT) provides the opportunity to focus on the domain specific terms of image description, and explore overlapping categories such as color, shape, or material. We anticipate that Panofsky's pre-iconographic and iconographic distinctions will also contribute to an understanding of the subject description analysis. Second, we will investigate the nature of the overlapping tag/term set in terms of the category, type and nature of lexical item. Finally, we intend to explore in much greater depth different types of metrics that cluster, categorize and assign tags and terms to topics, such as Latent Dirichlet Allocation (Blei et al. 2003) or Latent Semantic Indexing (Deerwester et al. 1990) utilizing coocurrence as well as occurrence with probabilistic modeling. In all cases, we will evaluate with human subjects to ensure that our intuitions are supported by objective experimental data.

Acknowledgements. The steve.museum project was initiated and led by Jennifer Trant, Archives & Museum Informatics, Susan Chun, Independent Museum Consultant and Rob Stein of the Indianapolis Museum of Art. Steve.museum are partners in the T3 project (umiacs.umd.edu/research/T3). All tagging data analyzed in this article was generously made available to us. We have benefitted from discussions with Jennifer Golbeck and Dagobert Soergel, both of whom are members of the T3 group.

References

1. Art & Architecture Thesaurus (AAT). Getty Vocabulary Program. Los Angeles: J. Paul Getty Trust,Vocabulary Program. (1988-)
2. Bearman, D. and Trant, J. Social Terminology Enhancement through Vernacular Engagement: Exploring Collaborative Annotation to Encourage Interaction with Museum Collections, D-Lib Magazine, 11(9) (2005)
3. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I. Latent Dirichlet Allocation. In: Journal of Machine Learning Research 3: pp. 993–1022. (2003).
4. Brunet, G., Chechik, M., Easterbrook, S., Nejati, S., Niu, N., and Sabetzadeh, M. A manifesto for model merging. In Proceedings of the 2006 international Workshop on Global integrated Model Management (2006)
5. Chai, J.Y., Zhang, C., Jin, R. An Empirical Investigation of User Term Feedback in Text-Based Targeted Image Search. ACM Transactions on Information Systems, vol. 25 no. 1 (2007)
6. Choi, Y., Rasmussen, E.M. Searching for Images: The Analysis of Users' Queries for Image Retrieval in American History. Journal of the American Society for Information Science and Technology, vol. 54 no. 6, 498--511 (2003)
7. Deerwester S., S. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman. Indexing by Latent Semantic Analysis. In: Journal of the American Society for Information Science 41 (6): 391–407. (1990).
8. Easterbrook, S., Chechik, M. A framework for multi-valued reasoning over inconsistent viewpoints, Proceedings of the 23rd International Conference on Software Engineering, 411-420 (2001)
9. Klavans J. L. and E. Tzoukermann. Morphology. In: S. Shapiro, editor, *The Encyclopedia of Artificial Intelligence*. John Wiley and Sons, New York (NY). second edition, (1992)
10. Lam, T., Singh, R. Semantically Relevant Image Retrieval by Combining Image and Linguistic Analysis. In: Advances in Visual Computing, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, pp. 770--779 (2006)
11. Lezius, W. Morphologiesystem Morphy. In: R. Hausser (ed.): Linguistische Verifikation. Dokumentation zur Ersten Morpholympics Niemeyer: Tübingen. 25--35. (1996)
12. Luhn, H.P. The Automatic Creation of Literature Abstracts. In: IBM Journal of Research and Development, Vol. 2, No. 2, 159—165 (1958)
13. Panofsky E., Studies in Iconology: Humanistic Themes in the Art of the Renaissance. New York: Harper & Row. (1972)
14. Sable, C., McKeown, K., Church, K. W. NLP Found Helpful (at least for one text categorization task). In: Proceedings of the Acl-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 172--179 (2002)
15. Salton, G., Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management (1988)
16. Smeulders, A.W. M., Worring, M., Santini, S., Gupta, A., Jain, R. Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence (2000)
17. Spärck Jones, K. A statistical interpretation of term specificity and its application in retrieval. In: Journal of Documentation 28 (1): 11–21. (1972)
18. Trant, J., Bearman, D., Chun, S. The eye of the beholder: steve.museum and social tagging of museum collections. In: Proceedings of the International Cultural Heritage Informatics Meeting - ICHIM07, Toronto, Canada. (2007)