Content-Based Similarity Measures of Weblog Authors

Christopher Wienberg, Melissa Roemmele, and Andrew S. Gordon
Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive
Los Angeles, CA 90094
{cwienberg,roemmele,gordon}@ict.usc.edu

ABSTRACT

With recent research interest in the confounding roles of homophily and contagion in studies of social influence, there is a strong need for reliable content-based measures of the similarity between people. In this paper, we investigate the use of text similarity measures as a way of predicting the similarity of prolific weblog authors. We describe a novel method of collecting human judgments of overall similarity between two authors, as well as demographic, political, cultural, religious, values, hobbies/interests, personality, and writing style similarity. We then apply a range of automated textual similarity measures based on word frequency counts, and calculate their statistical correlation with human judgments. Our findings indicate that commonly used text similarity measures do not correlate well with human judgments of author similarity. However, various measures that pay special attention to personal pronouns and their context correlate significantly with different facets of similarity.

Author Keywords
Similarity measures; Weblogs; Personal pronouns

ACM Classification Keywords
I.2.7 Natural Language Processing: Text analysis

General Terms
Human Factors

INTRODUCTION

Much of current research in social media analysis is motivated by the desire to make predictions about people. For example, predictions of political sentiment can be instrumental in soliciting campaign donations, and predictions of consumer preferences can be monetized through directed advertising. In pursuing this goal, researchers have found much success in the application of social network analysis, treating populations as nodes in a connected graph upon which predictive models can be developed and evaluated. Information flow, for example, has been modeled as contagion [11], where the fact that nodes in graphs represent human beings is often immaterial to the analytic technique. These analyses have led to contentious claims about social contagion of health conditions [3] and behavioral preferences [4], which in turn has produced a rich research debate about the role that homophily plays in understanding the results of social network analyses [12].

Homophily, that “birds of a feather flock together,” is the observation that individuals tend to associate with similar others. The phenomenon of homophily makes it difficult to estimate the role of influence or contagion in social networks, e.g. does a person start smoking cigarettes because their friends who smoke have influenced their behavior, or does the propensity to smoke help explain why they are all friends in the first place? Recent analyses of Shalizi and Thomas [23] show that homophily and contagion are generically confounded in observational social network studies, necessitating caution in the interpretation of graph-based influence models.

This interplay between contagion and homophily encourages new research in social media analysis with a greater emphasis on the nodes in graphical models, i.e. the qualities, opinions, and behaviors of the actual people that these nodes represent. With a better understanding of how individual people are similar, it may be possible to untangle the confounding roles of contagion and homophily in studies of social influence. Accordingly, there is a strong need for reliable measures of the similarity of social media authors that are not based on their relationships to others (the social graph). For the practicalities of scale, these must be content-based measures that can be completely automated. Here, we focus our interest on measures that compute similarity based on the text authored by social media users.

The relationship between written text and the personal characteristics of its authors has been a rich area of research in the fields of psychology and sociology, and is increasingly of interest in the field of computational linguistics. Previous efforts have shown that many traits of people, such as their demographics and health status, can be inferred from automated analyses of their writings. We hypothesize analogous methods can be used to develop new automated measures of the similarity between people.

In this paper we describe our efforts to identify a reliable measure of the similarity between people based on their online

WebSci'13, May 2–4, 2013, Paris, France.

writings. We focus our attention on the personal writing of prolific bloggers, who provide us with ample data for analysis and evaluation. We present a novel approach for gathering gold-standard similarity data, where third-party raters are asked to read the posts of pairs of weblog authors and make judgments as to their similarity across several dimensions. In this work we explore multiple facets of the similarity between people, such as their demographic, cultural, and religious similarity. We explore several techniques for computing the textual similarity of different authors' posts, motivated by findings in previous work in network-based classification, psychology, and sociology. We investigate the degree to which these measures correlate with human judgments of similarity, and discuss the implications of our findings for future work in social media analysis.

PEOPLE AND THEIR TEXT

When people write, they reveal who they are. This disclosure of identity in written text occurs in myriad ways that are at times even outside the writer's awareness. Tasks in automatic text analysis take advantage of this idea. For instance, research in authorship attribution assumes that each writer leaves a "stylistic footprint" on her work. Computers can be used to analyze the style of a text and match it to that of a known writer [9].

However, it is not necessary to identify who exactly a writer is in order to know something about them using text analysis. A lot of work focuses on automatically determining certain traits or attributes of people from their writing. For instance, in the field of psychology, the "Big Five" personality traits (extraversion, neuroticism, agreeableness, conscientiousness, and openness [7]) are considered the most important personality dimensions. Several studies have identified features of written language that strongly correlate with these traits in their author [1, 13, 15, 17, 24]. The same is true for gender. By analyzing their written language, computers can classify the gender of individuals at a level much higher than chance [10, 14]. Additional work indicates it is possible to predict the age range of an individual based on their writing [19, 21]. Certain patterns in written language can even reveal other phenomena, such as depression [22], personal distress [18], and cultural upheavals [5].

The above findings largely depend upon how written text is analyzed. The choice of which language features to focus on is an important concern of this text analysis research. Many of the studies above rely upon the Linguistic Inquiry and Word Count (LIWC) [16] tool for selecting features. LIWC divides words into various psychologically and linguistically based categories and evaluates the frequency of each category across an individual's text. In addition to LIWC, n-gram features (consecutive sequences of n words) have also been used to automatically identify individual traits in written text [6, 10, 15]. Some research has found promising results for using personal pronouns as features. Individuals vary systematically in their written pronoun use according to many identifying factors, including age, gender, and emotional disposition [17, 19, 22].

In short, we can use text analysis to characterize people based on their writing. In the current work, we apply this technique to weblog data in order to identify similarity between people.

SIMILARITY BETWEEN PEOPLE

Since early in the history of the web, maintaining a weblog has been a popular activity [20]. Once reserved only for the technically inclined, advances in software have made this activity accessible to people around the globe, even those with few technical skills. The genres of weblogs are as varied as the people who maintain them, with some writing about national politics, local news, celebrity gossip, personal anecdotes, or some combination of these, among many others.

When people think of weblogs, personal story weblogs often come to mind. In this genre, bloggers write personally about themselves, recounting the occurrences of their daily lives, in a manner akin to a public diary. A typical post in this genre will be a narrative of the author's day, providing details of both the interesting and mundane. These stories will be fairly long and typically contain many personal pronouns and past tense verbs. In addition to text, these stories are often supplemented with photographs taken in the context of the events narrated, creating a multimedia online presentation of a person.

Previous work by Gordon and Swanson [8] has shown these stories can be automatically identified. Using machine learning techniques, they identified nearly one million personal stories in the ICWSM Spinn3r 2009 weblog dataset [2]. Leveraging their story classification technology, we have collected a corpus of over 24 million personal stories posted in weblogs between 2010 and 2012. In addition to containing the content of a post, our corpus contains metadata, such as the date of posting and the weblog it was posted to. This metadata reveals an interesting property of this corpus: while many of the blogs were infrequently active, with just one or a handful of posts in the corpus, tens of thousands of weblogs had a high frequency of personal story posts—ranging from weekly to daily. These prolific personal story bloggers provide an immense amount of information about themselves in their writing, and the volume of data they provide makes them well suited for analysis.

The genre of personal story weblogs is particularly useful for understanding people. Specifically, we leveraged a subset of our corpus—the high frequency bloggers—to examine the use of their writing to measure their similarity to others.

We distilled our corpus of millions of personal weblog stories into a smaller dataset of posts from these prolific bloggers. We began by assuming that most blogs with personal stories are maintained by a single person, a reasonable assumption in our experience. Then, we selected all weblogs with at least 300, but less than 1200, posts in our story corpus. Choosing weblogs in this manner selected for bloggers who have provided a great amount of text about themselves, while filtering out spammers and other failure cases of the personal story classifier. Additionally, we filtered this for weblogs that are currently active, only selecting weblogs with at least 8 posts in the previous two months—an average rate of one post per week. We further filtered this data, selecting 260 blogs at random and confirming their accessibility on the web. We then randomly paired them, without replacement, for a set of 130 pairs of people, with no people appearing twice.

We collected human judgments of people similarity through a web survey. To build and conduct this survey, we used Qualtrics, a commercial online survey service. We utilized their panels service[1], in which Qualtrics handles issues of finding respondents. Qualtrics provided us with responses from US citizens of at least 18 years of age.

Each survey respondent was presented with just one pair of people from the set of 130. The respondents were asked to take ten to fifteen minutes to familiarize themselves with each author by consulting posts, photographs, descriptions, and any other posted content to construct a mental image of the author. After learning about each author, respondents were asked to write a brief description—a short sentence or two—of the author. Then respondents were asked to rate how similar they believe the two authors are. Our survey asked participants to rate authors along several dimensions of similarity, and described them as follows:

- **Demographic** - "How similar are the bloggers in age, ethnicity, gender, family structure, etc?"
- **Political** - "How similar are the bloggers' political opinions? Do the bloggers have similar levels of political interest, even if they have different opinions?"
- **Cultural** - "How similar are the bloggers culturally? Do they follow similar traditions? Do they have similar lifestyles?"
- **Religious** - "Do the bloggers have similar religious beliefs? Is religion important to them in similar ways?"
- **Values** - "Do the bloggers value similar things? Would they choose similar trade-offs and life decisions?"
- **Hobbies/Interests** - "Do the bloggers do similar things in their spare time?"
- **Personality** - "Do the bloggers have similar attitudes and dispositions? Are they both funny, serious, etc?"
- **Writing Style** - "Do the bloggers write in similar ways? Do they both write for audiences with similar reading levels?"
- **Overall** - "Considering the above characteristics and any others you can think of, how similar do you think these bloggers are overall?"

Respondents were asked to rate each of these facets of similarity on a seven point Likert scale, ranging from completely dissimilar to completely similar. Respondents were given the option to skip any facet they did not have enough evidence to assess.

Table 1. Correlation of Ratings between Respondents

Similarity Facet	Correlation
Hobbies/Interests	0.034
Political	0.045
Demographic	0.151
Personality	0.151
Writing Style	0.194
Religious	0.224
Cultural	0.255*
Values	0.345*
Overall	0.127

* average $p < 0.05$

Table 2. Correlation of Facets of Similarity to Overall Similarity

Similarity Facet	Correlation
Demographic	0.411*
Cultural	0.425*
Political	0.440*
Religious	0.491*
Personality	0.666*
Hobbies/Interests	0.685*
Values	0.700*
Writing Style	0.709*

* $p < 0.001$

In addition to ratings for all 130 blogger pairs, we collected a second set of judgments for 72 of these pairs, for use in assessing inter-rater agreement. As each rater assessed only a single pair of bloggers, traditional measures of inter-rater agreement, such as Cohen's kappa and Krippendorff's alpha, are not applicable. Instead, we used correlation analysis to assess the difficulty of the task. For each facet of similarity, we collected all similarity values where two respondents rated the same pair of bloggers. We compiled two judgment sets, where one half of these double responses were treated as if they belonged to one judgment set, and the other response to a second judgment set. We performed correlation analysis on these judgment sets using Pearson's product-moment correlation. Since correlation will change depending on how a judgment pair is divided over these two sets, we computed inter-rater correlation 10,000 times in this fashion, each time randomly selecting how to divide a judgment pair. We report these correlation values in table 1.

Most interesting among this data is the relatively high degree of response correlation with respect to values, religion, and culture. In particular, values and culture are among the more subjective facets of similarity examined in this work. Despite this, correlations between raters with respect to these facets, though weak, are statistically significant ($p < 0.05$), meaning that the null hypothesis of no correlation at all is rejected. Also interesting is how poorly correlated judgments of demographics are. As characteristics of demographics are much clearer than other categories (age, gender, and ethnicity are much more clearly defined and understood than differing cultural traditions, for instance) we expected the respondents to agree most of all in this category. Similarly, though not as clearly defined as demographics, we expected overall similarity to be one of the categories with more highly correlated judgments. Overall, the weak correlations between responses indicate that judging similarity is difficult. As human

[1] `http://www.qualtrics.com/panels/`

Table 3. Personal Pronouns

Category	Pronouns
Subjective	i, we, you, he, she, it, they
Objective	me, us, you, him, her, it, them
Possessive	my, mine, our, ours, your, yours, his, her, hers, its, their, theirs
Reflexive/Intensive	myself, ourselves, yourself, yourselves, himself, herself, itself, themselves

judges have trouble agreeing, automated methods for judging the similarity between people are likely to be difficult to develop and may be prone to errors.

We also examined the relationship of overall similarity to the various facets of similarity included in the data we collected. Using correlation analysis, we can see the relative influence of these facets on human assessments of similarity. The results of this analysis, presented in table 2, indicate writing style and personal values have the strongest impacts on third parties' judgments of people's similarity from text. Even the weakest correlations, however, are statistically significant, suggesting that people account for all of these facets, among others, when assessing the similarity between people.

COMPUTING PEOPLE SIMILARITY FROM TEXT

Similarity metrics have aided progress in many computational fields, such as Natural Language Processing and Computer Vision. For instance, in text retrieval, documents are ranked based on their similarity to a query. Text documents can be grouped using unsupervised machine learning techniques such as k-means clustering, which groups documents based on their similarity. Computer Vision researchers have developed similarity metrics to enable images to be used in analogous tasks. Two factors matter when computing similarity: the representation of the data and the metric used to compare. In this section, we present four representations of people, designed for use with the cosine vector similarity metric.

Bag-of-Words Text Similarity

The simplest strategy for working with arbitrary text is to represent it using the bag-of-words model, where each term is an individual feature in a long (typically sparse) vector. Word order, syntax, and sentence boundaries are all discarded. While simple, it is rather effective, and is commonly used in many natural language tasks, especially information retrieval.

While raw counts can be used, weighting schemes are a common addition to this model. Information retrieval researchers have observed that words that carry little information such as prepositions and articles appear frequently within and across most documents. Conversely, words that are very informative about the topic of a document, specialized terms like "senator" in documents about government policy and "insurance" in documents about health care, appear in only a few. Using raw term counts will result in common words like "the" having an inappropriate influence on measures such as document similarity.

The term frequency-inverse document frequency (tf-idf) weighting scheme was developed to balance between competing trade-offs. On the one hand, terms that appear frequently in a document tend to indicate the topic of that document; on the other hand, terms that appear in many documents, even if they appear frequently in a particular document, are unlikely to indicate the topic of that document. Tf-idf is a class of weighting schemes, rather than a specific method, and there are many ways to compute it. The general approach is to multiply a measure of term frequency (how frequently a term appears in a document) by a measure of the inverse of document frequency (how frequently a term appears across documents).

We represented authors in our weblog story corpus using bag-of-words term vectors, weighted with tf-idf scores. Raw term frequency in a document was used as the term frequency measure. Inverse document frequency was computed by taking the logarithm of the quotient of the size of the corpus divided by the number of documents a term appears in. These scores were computed over the sub-corpus of stories from bloggers included in our similarity survey. Each author was represented as the sum of all tf-idf weighted bag-of-words term vectors for all stories she has written in our story corpus.

Personal Pronoun Counts

An intuition underpinning this investigation is that personal pronouns are particularly important when understanding and examining characteristics of people. As personal pronouns directly reference people—in this context, the author or people the author knows and interacts with—their usage should be particularly indicative of what people are like. As work by Pennebaker and others have shown, pronoun usage in writing is indicative of many characteristics of the author, including gender [10, 14], age [19, 21], and even mental states, such as depression [22]. This is without regard to any additional context: simply what pronouns an author uses and in what frequencies are often sufficient information to make accurate predictions about the author.

Building upon this work, we examine the usefulness of personal pronoun frequency in assessing the similarity of two people. We began by assembling a list of 31 unique personal pronouns, presented in table 3. For this work, we considered four categories of personal pronouns: subjective, objective, possessive, and reflexive/intensive pronouns. We counted the occurrences of these pronouns in each author's posts, and used these counts to construct a vector of raw pronoun counts to represent each author.

Left-Right Context

Expanding on the notion that pronoun usage in text is indicative of an author's personal characteristics, we examined an author representation that utilizes the context a pronoun is used in. Authors do not write about themselves in a vacuum. Rather, they describe their actions, emotions, and thoughts in any given situation.

Table 4. Example Left-Right Contexts

Sentence: “Yesterday, we had a fantastic day with you at the beach.”

Context Size	Contexts
1	“yesterday we had”, “with you at”
2	“yesterday we had a”, “day with you at the”
3	“yesterday we had a fantastic”, “fantastic day with you at the beach”
4	“yesterday we had a fantastic day”, “a fantastic day with you at the beach”

Words that an author uses in relation to people should be particularly indicative of properties of that person. While determining which words in a sentence describe a person is a challenging natural language processing task, a good heuristic is to look at the words immediately to the left and right of a personal pronoun, called the left-right context. We show the left-right contexts for an example sentence in table 4. Though these words will not all be related to the person mentioned, typically most will be.

For each author, we constructed a vector of every personal pronoun used and its accompanying left-right context. We considered varying sizes of left-right contexts. We looked at contexts of 0 (the raw personal pronoun counts described in the previous subsection) through 4 (the pronoun, with four words on the left and four words on the right). Larger context sizes did not also include smaller contexts (i.e. a left-right context of size 4 does not also include left-right contexts of size 3). Contexts did not cross sentence boundaries; a left or right context that would traverse a sentence boundary was truncated at the boundary. The vector representing an author includes the raw counts of all left-right personal pronoun contexts in that author’s posts.

LIWC Features

As described previously, the Linguistic Inquiry and Word Count (LIWC) application is software for analyzing text [16]. LIWC utilizes dictionaries to categorize words and counts the words in a document with respect to their dictionary category. This dictionary system allows LIWC to score documents against a vast array of categories, ranging from abstract categories regarding emotional state to more linguistic categories such as whether a word is a verb or a noun. The LIWC dictionary contains nearly 4,500 words, each one curated by hand and categorized by its linguistic and psychological content.

The LIWC application analyzes documents and returns a vector of decimal values, one value for each category of analysis. For this work, we analyzed the text of each weblog author in our similarity survey. We used the 2007 LIWC dictionaries, which provide 82 distinct categories. For each author, we concatenated all posts in our story corpus from that author and analyzed it with the LIWC application. We used the resulting vector, unaltered, as the representation for that author.

The Cosine Similarity Metric

We have described four vector-based representations of people based on text they write but have not described how to compare them. The cosine similarity metric is a widely used strategy for comparing the similarity of two vectors. While typically used with textual feature vectors, the cosine similarity metric can be used with any data represented as a vector of numeric values.

Intuitively, the cosine similarity of two vectors is high when they have similar feature distributions, and low when their feature distributions vary greatly. It ranges from -1 to 1, with strongly negative values assigned to vectors that are of opposite polarity, and strongly positive values assigned to vectors that are very similar. Values of 0 indicate no similarity, i.e. the vectors are orthogonal.

Cosine similarity is computed by taking the cosine of the angle between two vectors. Due to this it is invariant to the scale of the two vectors. This is a nice property for our study, enabling two authors to be considered similar even if one writes much more frequently than another. For vectors X and Y, cosine similarity is computed as:

$$CoSim(X,Y) = \frac{X \cdot Y}{||X||\,||Y||} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n}(X_i)^2}\sqrt{\sum_{i=1}^{n}(Y_i)^2}}$$

The cosine similarity of two vectors can be computed quickly, enabling us to compare the similarity of two authors effectively.

RESULTS

We examined the use of automatic methods for computing the similarity between people using correlation analysis. We began by calculating cosine similarity between all pairs of bloggers featured in the blogger similarity survey using each of the representations featured in the previous section. We then calculated the correlation between these automatic similarity metrics and human judgments of similarity, using Pearson’s product-moment correlation. We computed correlation with all dimensions of similarity included in the survey. The results of this analysis are presented in table 5.

Cosine similarity on bag-of-words term vectors poorly approximates the similarity between people. Cosine similarity on bag-of-words term vectors correlates weakly with any facet of similarity, and none of these correlations are statistically significant. This result suggests that utilizing bag-of-words cosine similarity as a stand-in for the similarity of users, such as in some work analyzing homophily in social networks, is misguided.

A better indicator of the similarity between people is their usage of personal pronouns. As stated earlier, pronoun usage

Table 5. Correlation of Automatic Similarity Metrics to Human Judgments

Similarity Facet	Bag-of-Words	Pronoun	Pronoun Left-Right Context				LIWC
		Count	1	2	3	4	
Demographic	-0.108	-0.176	-0.038	0.080	0.080	0.033	-0.016
Political	-0.027	-0.118	-0.032	0.008	-0.036	0.008	0.003
Cultural	0.136	0.133	0.136	0.170	0.189*	0.200*	-0.107
Religious	0.173	0.222*	0.186	0.146	0.111	0.176	-0.015
Values	0.025	0.113	0.221*	0.262*	0.208*	0.133	0.002
Hobbies/Interests	0.034	0.111	0.213*	0.154	0.082	0.110	-0.025
Personality	-0.009	0.115	0.253*	0.253*	0.197*	0.172	-0.027
Writing Style	0.119	0.076	0.182*	0.123	0.078	0.105	-0.055
Overall	0.025	0.161	0.204*	0.171	0.056	0.048	0.002

* $p < 0.05$

has been shown to be indicative of many personal features of people. In the case of weblog authors this is borne out, with a weak but statistically significant correlation between the cosine similarity of two weblog authors' personal pronoun usage counts and human judgments of their religious similarity. To set this result in context, review of many personal stories in our corpus indicate that people who frequently mention their religious beliefs in their personal weblogs often refer to their personal higher power with pronouns such as "he" and "him." This trend in personal pronoun usage may cause people with similar religious beliefs to be highlighted by similarity over personal pronoun counts.

Cosine similarity over personal pronouns and their left-right contexts is a particularly strong approach to assessing the similarity of two people automatically. A context of just one word to the left and right of a personal pronoun reveals a great amount of information about a person. Using this information, cosine similarity on this representation has statistically significant correlations with human judgments of a pair's similarity with respect to values, hobbies, personality, and writing style, as well as overall similarity. Additional context shows mixed results, with improvements for a few individual facets of similarity and marked performance declines for others. Using two and three words of left-right context with the cosine similarity metric is still correlated—to a statistically significant degree—with human judgments of people's similarity with respect to values and personality. However, using three and four words of left-right context when computing similarity also shows a statistically significant correlation between the automatic method of assessing similarity and human judgments with respect to culture, a correlation not seen with fewer words of left-right context.

Using the results of processing text with LIWC as a feature vector for cosine similarity shows no correlation with human judgments of any facet of similarity. While surprising, it is unclear how much can be inferred from this finding. This is only a single, limited, non-traditional application of LIWC analysis. As it is non-traditional, poor performance using vectors from LIWC analysis may be explained by a handful of unhelpful dimensions dominating similarity calculations, for instance. These results are discouraging, but different strategies for using LIWC—such as careful selection of features for this task—may be more successful. Our results indicate, however, that off-the-shelf applications of LIWC are ineffective for the task of comparing the similarity between people.

Except for demographics and politics, all facets of similarity have at least one automatic method of measuring similarity with a statistically significant correlation to human judgments of similarity. This indicates that similarity metrics can be developed to specialize in determining the similarity of any single facet of similarity. For instance, a specialized similarity metric for religious similarity could be developed, likely relying heavily on personal pronoun count features.

However, the failure to find any statistically significant correlation with either demographics or politics is notable. Both of these facets of people have been targeted in social media research, with research on political views, social networks, and the outcomes of elections having seen significant work in recent years, and demographics often factoring into discussions of network homophily. While we cannot make any conclusions about different genres of social media data, that similarity with respect to these characteristics within personal story data was not correlated with any automatic similarity metric should cast some doubt on the applicability of these methods.

DISCUSSION

We have presented a novel approach to assessing the similarity between people. This work has shown that the similarity between people can be assessed automatically from text, an encouraging finding for social scientists. Previously, there was no clear strategy for automatically comparing people, let alone evaluating those comparisons. Now, our technique of collecting a gold-standard set of third-party human judgments opens an avenue for more quantitative research into the similarity between people.

Our findings support the intuition that personal pronouns are highly salient when studying people from text. However, our results indicate that personal pronouns alone are generally insufficient to measure the similarity between people. Instead, the context in which personal pronouns—references to people—are placed is vital to this. Our results using left-right context to compute the similarity between people are encouraging. Just a single word of context appearing to either side of a personal pronoun reveals a significant amount of information about a person. Our finding of statistically significant correlations between cosine similarity over a vector of personal pronoun left-right contexts of one word and human similarity judgments of values, hobbies/interests, personality, writing style, and overall similarity is a strong indicator that analyzing references to people and the words that are used to describe them is the right approach when studying and comparing people through text. That different context window sizes provide correlations with human judgments of different facets of similarity—such as cultural or religious similarity—is a sign that similarity metrics can be developed to specialize in particular facets of similarity that are of interest to researchers, and that focusing on references to people is fundamental to analyzing people themselves.

There remains work to be done in this avenue of research. Foremost, we must address the issue of low inter-rater agreement in human judgments of blogger similarity. We observe statistically significant agreement for the similarity facets of values and culture, and therefore are most confident in our conclusions with respect to these facets. Although the other facets appear to be trending in the right direction, it will be necessary to obtain stronger inter-rater agreement in order to use the judgments as gold-standard test data. There are several approaches that could be pursued. For example, additional time, instruction, and training of our raters should yield more consistent judgments of what makes two people similar or dissimilar. Additionally, using a few trained raters rather than hundreds of untrained raters would permit more traditional measures of inter-rater agreement, such as Cohen's kappa or Krippendorff's alpha. Alternatively, we could simply collect much more data, including more overlapping data. With more data, random variation between raters will be smoothed out, yielding stronger and more confident correlations between raters. Also, with more overlapping judgments (i.e. many raters judging the same pair of people), techniques like averaging judgments, or selecting judgments with agreement from multiple raters, become available. By addressing concerns about inter-rater agreement, we can be more confident in our conclusions across all facets of similarity.

A side benefit of collecting more data is that it enables the use of supervised machine learning techniques, where what makes people similar can be learned directly from the data. Regression analysis could be performed on each facet of similarity, allowing researchers to analyze the similarity between people in a manner specialized to the task of interest (i.e. researchers studying political issues in social networks can compare users on their political similarity). This prospect of learning what makes people similar directly from data is a tantalizing avenue for future work.

We have shown that the similarity between people, as they portray themselves on the web, can be computed automatically. However, a concern for this work and most work in social media analysis is the underlying assumption that people present themselves honestly on the web. The web provides people an opportunity to change how others perceive them. Users can emphasize different parts of their personality, explore contrasting opinions, or even completely fabricate properties about themselves, all in the comfort of anonymity. It is important to know how true-to-life an online persona is. Crucial future work to social media analysis is to assess how accurate an online persona is to the person at the keyboard.

Our work has repercussions for social media and other researchers interested in quantitatively analyzing and comparing people. According to our results, traditional techniques for assessing the similarity between people, like a bag-of-words model or LIWC analysis, do not measure what researchers expect. Fortunately, other simple techniques leveraging personal pronouns and the context they appear in are available to researchers when looking to compare the similarity between people from text. Researchers who are currently using techniques like bag-of-words similarity should consider what it is they wish to measure. If similarity of content is all that is needed, this technique suffices. However, we encourage researchers to look deeper, at the similarity of the people generating that content. When doing so, researchers should take care to use a technique that works well for the facet of similarity they are interested in measuring.

Perhaps the most unexpected finding in our work is with respect to people's values. Across this work, similarity with respect to values stands out for strong, statistically significant correlations. Similarity with respect to values has the strongest inter-rater correlation (one of only two facets with statistically significant inter-rater correlation) indicating humans can easily understand and compare the values of others. Human judgments of similarity with respect to values are strongly correlated with human judgments of overall similarity, indicating that a person's values are important to understanding a person and how he compares to others. Techniques to automatically compute the similarity between people have some of the strongest correlations with human judgments of similarity of values, indicating that this facet of similarity is particularly amenable to text analysis. This finding is particularly surprising because the values of a person are rather abstract. Compared to a more concrete concept like demographics, the values of a person are difficult to define and compare and we expected it to have among the poorest inter-rater correlation. Additionally, compared to facets of similarity like hobbies or writing style, which should have a strong impact on vocabulary choices, values has a poorly specified vocabulary associated to it, and we expected text analysis to struggle to measure it.

These results should be encouraging to researchers interested in the values of people. They indicate that values are an important facet of people and that they can be analyzed using simple techniques applied to the text people write.

ACKNOWLEDGMENTS

The projects or efforts depicted were or are sponsored by the U. S. Army. The content or information presented does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

REFERENCES

1. Argamon, S., Dhawle, S., Koppel, M., and Pennebaker, J. Lexical predictors of personality type. In *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America* (2005).
2. Burton, K., Java, A., and Soboroff, I. The icwsm 2009 spinn3r dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media*, ICWSM 2009 (San Jose, CA, 2009).
3. Christakis, N., and Fowler, J. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine 357*, 4 (2007), 370–379.
4. Christakis, N., and Fowler, J. The collective dynamics of smoking in a large social network. *New England Journal of Medicine 358*, 21 (2008), 2249–2258.
5. Cohn, M., Mehl, M., and Pennebaker, J. Linguistic markers of psychological change surrounding september 11, 2001. *Psychological Science 15*, 10 (2004), 687–693.
6. Fast, L., and Funder, D. Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology 94*, 2 (2008), 334.
7. Goldberg, L. An alternative "description of personality": the big-five factor structure. *Journal of Personality and Social Psychology; Journal of Personality and Social Psychology 59*, 6 (1990), 1216–1229.
8. Gordon, A., and Swanson, R. Identifying personal stories in millions of weblog entries. In *Proceedings of the Third International Conference on Weblogs and Social Media, Data Challenge Workshop*, ICWSM 2009 (San Jose, CA, 2009).
9. Holmes, D. Authorship attribution. *Computers and the Humanities 28*, 2 (1994), 87–106.
10. Koppel, M., Argamon, S., and Shimoni, A. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing 17*, 4 (2002), 401–412.
11. Lerman, K., and Ghosh, R. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*, ICWSM 2010 (Washington, DC, 2010).
12. Lyons, R. The spread of evidence-poor medicine via flawed social-network analysis. *Statistics, Politics, and Policy 2*, 1 (2011), Article 2.
13. Mairesse, F., Walker, M., Mehl, M., and Moore, R. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research 30*, 1 (2007), 457–500.
14. Nowson, S., and Oberlander, J. The identity of bloggers: Openness and gender in personal weblogs. In *Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs* (2006).
15. Oberlander, J., and Nowson, S. Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics*, COLING-ACL '06 (2006), 627–634.
16. Pennebaker, J., Francis, M., and Booth, R. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* (2001).
17. Pennebaker, J., and King, L. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology 77*, 6 (1999), 1296–1312.
18. Pennebaker, J., and Lay, T. Language use and personality during crises: Analyses of mayor rudolph giuliani's press conferences. *Journal of Research in Personality 36*, 3 (2002), 271–282.
19. Pennebaker, J., and Stone, L. Words of wisdom: Language use over the life span. *Journal of personality and social psychology 85*, 2 (2003), 291–301.
20. Rosenberg, S. *Say Everything: How blogging began, what it's becoming, and why it matters*. Crown Publishers, New York, 2009.
21. Rosenthal, S., and McKeown, K. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (2011), 763–772.
22. Rude, S., Gortner, E., and Pennebaker, J. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion 18*, 8 (2004), 1121–1133.
23. Shalizi, C., and Thomas, A. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research 40*, 2 (2011), 211–239.
24. Yarkoni, T. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality 44*, 3 (2010), 363–373.