Routledge
Taylor & Francis Group

Creativity Research Journal

ISSN: 1040-0419 (Print) 1532-6934 (Online) Journal homepage: https://www.tandfonline.com/loi/hcrj20

The Semantic Network Model of Creativity: Analysis of Online Social Media Data

Feng Yu, Theodore Peng, Kaiping Peng, Sam Xianjun Zheng & Zhiyuan Liu

CREATIVITY RESEARCH JOURNAL, *28*(3), 268–274, 2016

ISSN: 1040-0419 print/1532-6934 online
DOI: 10.1080/10400419.2016.1195618

The Semantic Network Model of Creativity: Analysis of Online Social Media Data

Feng Yu
Xi'an Jiaotong University, China

Theodore Peng
Tsinghua University, China

Kaiping Peng
Tsinghua University, China & University of California, Berkeley

Sam Xianjun Zheng
Siemens Corporate Research

Zhiyuan Liu
Tsinghua University, China

The central hypothesis of Semantic Network Model of Creativity is that creative people, who are exposed to more information that are both novel and useful, will have more interconnections between event schemas in their associations. The networks of event schemas in creative people's minds were expected to be wider and denser than those in less creative people's minds. Based on this theory, data from Chinese online social media, also known as "Weibo microblogging," were analyzed. Each user's score consisted of the metric of coverage, which represented the spread of the network, as well as the metric of density, which represented the interconnections among nodes in the network. The results showed that occupations had a significant effect on people's creativity score. Academic scholars and writers in general had higher scores compared to other groups, such as entertainment celebrities and sport stars. The implications and limitations of this method of quantifying people's creativity were discussed.

Creativity is central to human imagination and problem solving, which was often considered to contain two dimensions: novelty and utility (Diedrich, Benedek, Jauk, & Neubauer, 2015; Paletz & Peng, 2008, 2009; Runco & Jaeger, 2012). The two dimensions of novelty and utility of creativity are consistent with *T* talent theory suggesting that a creative person not only has wide range of knowledge, but also has deep understanding in some domains (Lin, 2001). In this study, we proposed a Semantic Network Model to reconstruct the usefulness and novelty dimension of creativity and attempted to find a way to measure it in online social media.

According to constructionism learning theory, knowledge representations, which are stored in semantic memory, are responsible for the generation of new ideas (Sowa, 1991). Semantic memory is often described using a network metaphor (Collins & Loftus, 1975). The network consists of concepts, represented as nodes, and connections between concepts, represented as lines. Each line has a weight, representing how strong of an associative connection there is between the two concepts it bridges. When one node is activated, some residual activation spreads to other nodes in proportion to how strong of a connection they have.

Correspondence should be sent to Professor Feng Yu, Institute of Social Psychology, School of Humanities and Social Sciences, Xi'an Jiaotong University, 28# Xianning West Road, Xi'an 710049, P.R. China. E-mail: yufeng@sem.tsinghua.edu.cn

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hcrj.

Creativity can be viewed as activating the creator's schematic knowledge both for what is novel and for what is useful (Diedrich et al., 2015; Runco & Jaeger, 2012). When two knowledge structures are activated at the same time, they become more associated in memory, such that activating one increases the likelihood of activating the other. This means that a creative person can come up with ideas through many activations of the knowledge structure. It can be assumed that the creator will, on average, be exposed to more information than those who are less creative.

Novelty represents the coverage of network. Creativity contains novel ideas that are unrelated to existing knowledge. We generated fruitful ideas for solving a problem, but the possibility of occurrence of the novel idea will activate farther conception nodes. Evidence suggests that creators generate many different ideas when faced with a problem (e.g., in alternative uses task participants were asked to list as many possible uses for a common house hold item; Guilford, 1967). The central hypothesis of usefulness of novelty is that creative people will activate more nodes and that the activations will be more wide-spread compared to normal people.

Utility represents the density of the network. When facing a problem-solving situation, the creator not only generates more ideas and activates more remote conception nodes, but they also generate many ideas that are converged together. The convergence of these conception nodes makes the solution useful. High density activation of nodes, which makes people focus on a specific area and boosts creativity, is similar to rumination under depression (Cohen & Ferrari, 2010; Verhaeghen, Joormann, & Khan, 2005). The central hypothesis of the utility of creativity is that creative people will have more interconnections between event schemas in their memory than the regular population, and that these interconnections exist between more converging nodes.

Figure 1 depicts the nodes of knowledge activation of the semantic network. When facing a problem, a creative person will activate more conception nodes in the semantic network, which covered a very large part of the entire network space. For the utility dimension, nodes near each other activated together but demonstrated a lack of coverage, as they only covered a small part of the network space. For the novelty dimension, nodes far away from each other activated and covered a large part of the network space but were low density. For the normal population, both coverage and density were small.

In this study, social media data were used to test the semantic network theory based on natural language on Sina Weibo (Chinese version of Twitter), and this kind of Big Data method was started to be used in social and cultural psychology (Park et al., 2015). Because microblogs automatically record users' language, they were suitable for exploring the semantic network model of creativity. Although creativity could be divided into linguistic, mathematical, spatial, kinesthetic, rhythmic, social, introspective, and naturalist areas based on the Multiple Intelligences Theory (Gardner, 1985), this study focuses on linguistic creativity. Previous research found that academic scholars and writers will be the most creative people in the language area (Kogan, 2002). Entertainment celebrities are good at performing arts but not particularly creative at the language level (Glück, Ernst, & Unger, 2002; Kogan, 2002). Sport stars are good at dealing with sports but, on average, will be less creative than academic scholars and writers. Although many businessmen believe that creativity is about thinking outside the box, their creativity is usually evaluated by the

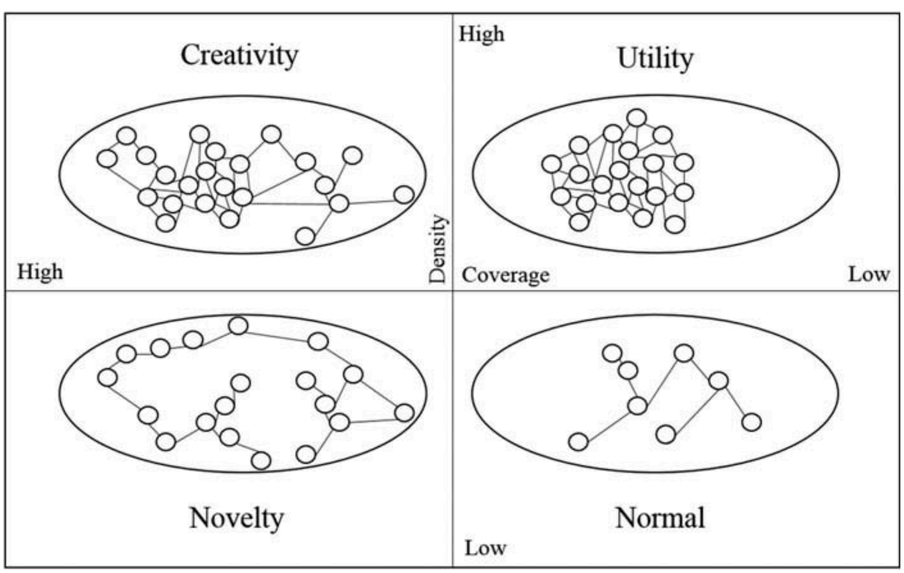

FIGURE 1 Semantic network model of creativity.

innovation in their products and unique corporate strategies (Petocz, Reid, & Taylor, 2009). Thus, to establish a criterion by which to analyze linguistic abilities, Weibo users were categorized by occupation. We hypothesized that academic scholars and writers were more likely to express their creativity on microblogs than non-scholars and writers. Using the two-dimension model, the coverage and density of academic scholars' and writers' microblog words will be higher than those of non-scholars and writers.

METHODS

Dataset

Microblog data were collected from Sina Weibo (Weibo.com) from 2010 to 2011. The dataset had more than 7 million users and more than 4 billion microblogs. The ten thousand most active users were selected according to the ranking of follower numbers in the dataset. By filtering out stop words and punctuations, Chinese word segmentation was performed on all their microblogs, which resulted in approximately 17 million messages. After filtering out words that occurred fewer than 10 times, a vocabulary of 180 thousand words remained. Using the vocabulary list and the users, we computed the coverage and density of each user to measure creativity.

Coverage

We collected messages of all microblog users, denoted as M. For each microblog user u in U, we collected the messages of u as M_u. For messages in M, we collected the global vocabulary as V, and for messages in M_u, we collected the user vocabulary as V_u. We defined the coverage of u as,

$$Coverage = \frac{V_u}{V}$$

The previous measure did not consider the differences between words. In reality, some words were more meaningful than others. Hence, we used TFIDF (term-frequency inverse-document-frequency) (Manning, Raghavan, & Schütze, 2008) to indicate the importance of a word. Formally, TFIDF of a word in a document set was defined as follows:

$$TFIDF_{w,M} = TF_w \times \log\frac{|M|}{|M_w|}$$

where TF_w was the frequency of w in M, $|M|$was the total number of messages in M, and $|M_w|$was the total number of messages in M that contain the word w. TFIDF reduced the importance of those highest-frequency words, which were usually function words and did not have specific meanings. Based on TFIDF statistics, we defined weighted coverage as,

$$Weighted\_Coverage = \frac{\sum_{w\in V_u} TFIDF_{w,M_u}}{\sum_{w\in V} TFIDF_{w,M}}$$

Density

For computing density, we used pointwise mutual information (PMI) to measure the relatedness between two words. Given a set of messages M, PMI of two words w and v was defined as

$$PMI(w,v) = \log\frac{p(w,v)}{p(w)p(v)}$$

where $p(w,v)$ was the frequency of messages having both w and v, $p(w)$ and $p(v)$ were the frequencies of messages having w and v separately. PMI was a popular method to measure the relatedness between two objects according to their co-occurrences (Lin, 1998). Similarly, computed global PMI and use PMI between two words, and then we defined the divergence of a user u as

$$Density = \frac{1}{|M_u|}\sum_{w,v\in V_u} \max(PMI_{M_u}(w,v) - PMI_M(w,v), 0)$$

to represent the density of a given user.

Creativity

As the semantic network model of creativity suggested, creativity is the combination of coverage and density, so we used the product of coverage Z score and density Z score to compute creativity. Which is:

$$Creativity = Z(Weight\ Coverage) * Z(Density)$$

RESULTS

The most active 10,000 users were chosen for this study, and only microblogs of certified users were examined in this study. The total number of users was 2308 (Mean of the amount of microblogs = 848.10, SD = 1232.324; Mean of followers = 1,593,454.40, SD = 2,836,558.010). Based on their reasons for certification (VIP users in Sina Weibo have yellow "V's" after their IDs as official certification. They have all been real-name authenticated and have provided certification reasons), we coded the certified users into five categories: academic scholars and writers, business leaders, entertainment celebrities, sport stars, and others. Two psychologists coded separately. We used a conservative principle, which allowed a user into a category only when both coders were in agreement, supporting high inter-rater reliability. In case of disagreements, we always put the user in the category of "others".

The coding results contained 284 academic scholars and writers, 505 business leaders, 1,258 entertainment celebrities, 168 sport stars and 93 others. Weight coverage and density were calculated using formulas and methods described in the methods section. Scores of weight coverage and divergence were standardized to make them comparable. Figure 2 depicted the Z score of weight coverage and density of each group. Using Z score of weighted coverage as dependent variable and category of users as independent variable, linear regression found occupation categories significantly predicted the coverage, $F(4, 2303) = 34.495$, $p < .001$, $R^2 = 0.057$. Table 1 depicted the results of this regression, and the coverage of academic scholars and writers was significantly smaller than those of business leaders (effect size, $r = .138$), entertainment celebrities ($r = .326$), and sport stars ($r = .179$). Using Z score of density as dependent variable and category of users as independent variable, linear regression found occupation categories significantly predicted the coverage, $F(4, 2303) = 67.175$, $p < .001$, $R^2 = 0.057$. Also, Table 1 depicted the results of this regression, and the density of academic scholars and writers was significantly bigger than those of business leaders (r = .–100), entertainment celebrities ($r = –.402$), and sport stars ($r = –.200$).

Figure 3 depicted the creativity of five groups. Creativity was calculated by Zcoverage × Zdensity. Using creativity as dependent variable and category of users as independent variable, linear regression found occupation categories significantly predicted the coverage, $F(4, 2303) = 3.711$, $p < .01$, $R^2 = 0.006$. Also, Table 1 depicted the results of this regression, and the density of academic scholars and writers was significantly bigger than those of business leaders (r = .–064), entertainment celebrities ($r = –.120$), and sport stars ($r = –.031$). The result suggested that academic scholars and writers are more creative than business leaders, entertainment celebrities, and sport stars at a linguistic level, which supported our hypothesis.

Figure 4 depicted the word cloud of typical highly creative and uncreative persons. Further content analysis was conducted to show the most frequent words each category

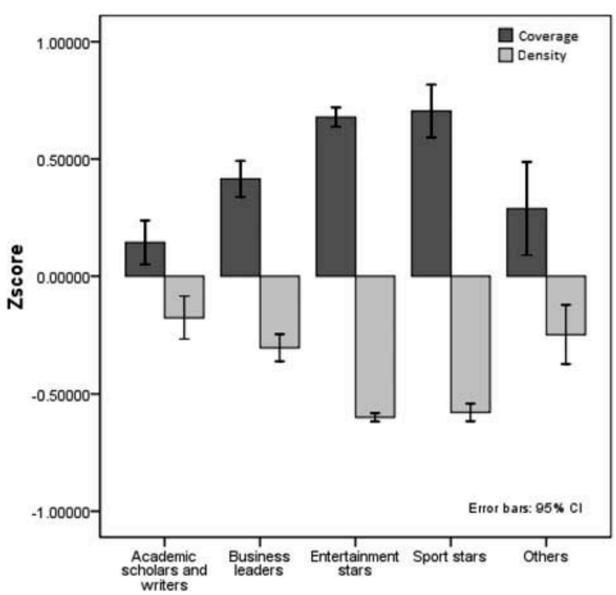

FIGURE 2 Coverage and density of different groups.

TABLE 1
Regression Results of Category of Users on Predicting Coverage, Density, and Creativity

(N = 2308)	M	SD	Standard Beta	t	p
Coverage as dependent variable					
D1	.2188	.41353	.138	4.612	.000
D2	.5451	.49807	.326	10.251	.000
D3	.0728	.25985	.179	7.260	.000
D4	.0403	.19669	.035	1.528	.127
Density as dependent variable					
D1	.2188	.41353	−.100	−3.454	.001
D2	.5451	.49807	−.402	−12.982	.000
D3	.0728	.25985	−.200	−8.329	.000
D4	.0403	.19669	−.027	−1.208	.227
Creativity as dependent variable					
D1	.2188	.41353	−.064	−2.081	.038
D2	.5451	.49807	−.120	−3.671	.000
D3	.0728	.25985	−.064	−2.539	.011
D4	.0403	.19669	−.031	−1.308	.191

Note. In these three regression models, category of users were coded into 4 dummy variable (D1, D2, D3, D4), academic scholars and writers were used as reference group (0, 0, 0, 0 on 4 dummy variables), and D1 means the difference between business leaders (1, 0, 0, 0) and academic scholars and writers, D2 means the difference between entertainment celebrities (0, 1, 0, 0) and academic scholars and writers, D3 means the difference between sport stars (0, 0, 1, 0) and academic scholars and writers, and D4 means the difference between others (0, 0, 0, 1) and academic scholars and writers.

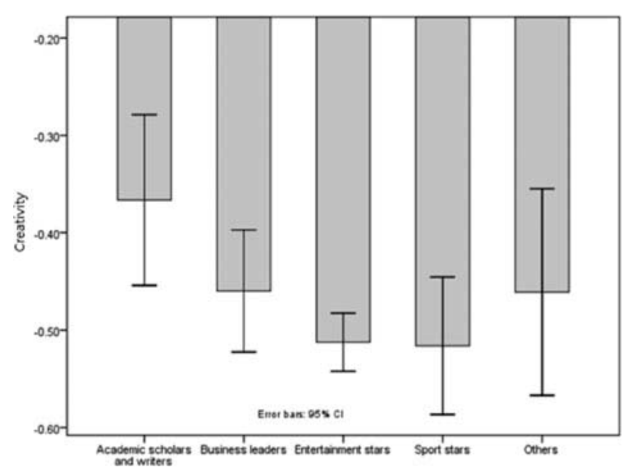

FIGURE 3 Creativity of different groups.

FIGURE 4 Tag cloud examples of high and low creative users.
Note. The first word cloud is a professor of economics, the words are: Economics, China, Finances, Markets, Money, Future, Company, Political System, Investment, Germany, Japan, Prices, Fluid, Pressure, Stokes, Needs, Relations, Yuan, Blogs, Factors, etc. The second cloud is a singer, the words are: Beijing, Thanks, Success, Good afternoon, Healthy, Search, Safety, Happiness, Kiss, Management, Blogs, System, Forward, Taipei, Hong Kong, Xi'an, Earth, World.

used. In Table 2 we listed the most frequent words used by different categories of users. In this table, we removed prepositions (e.g., on, in the middle, etc.), degree adverbs (e.g., very, greatly, etc.), words with negative connotations (e.g., no), and certain verbs used in colloquialism (e.g., do, want, etc.). In this table, different categories had their own features, academic scholars and writers are more likely to use the words associated with writing, like *writing, book*, etc.; business leaders are more likely to say *company, corporate*, etc.; entertainment celebrities are more likely to say *shooting, movie*, etc.; and sport stars are more likely to use words related to games, such as *match, champions*, etc.

DISCUSSIONS

Our results supported the hypothesis that the linguistic creativity of academic scholars and writers was higher than that of entertainment and sport stars. The results were consistent with our assertion that scholars and writers are more creative in the literal world when compared to business leaders, entertainment celebrities, and sport stars.

On the density dimension, academic scholars and writers were the highest, and entertainment celebrities and sport stars were the lowest. In contrast, on the coverage dimension, the results were the opposite. It is interesting how the word coverage of entertainment celebrities and sport stars were the highest among all these categories, and the word coverage of academic scholars and writers were the lowest. This might be because stars used fresh, popular words more than scholars, but the latter were more likely to use academics terms and languages that only cover a small range of topics. An additional interpretation of these results could be that the stars prefer to change their topics in microblogs frequently, but scholars and writers tend to stick to the few topics that they care about.

This study had some limitations. First, because each occupation had different terms or shared languages, the

TABLE 2
Content Analysis of Different Categories of Users

Academic scholars and writers			*Business leaders*			*Entertainment celebrities*			*Sport stars*			*Others*		
Words	*English*	*Frequency*	*Word*	*English*	*Frequency*	*Word*	*English*	*Frequency*	*Word*	*English*	*Frequency*	*Word*	*English*	*Frequency*
中国	China	0.001504	中国	China	0.000866	朋友	Friends	0.000829	比赛	Match	0.003160	生活	Living	0.004923
美国	USA	0.000537	人生	Life	0.000473	喜欢	Liking	0.000795	中国	China	0.002698	人生	Life	0.004850
孩子	Kids	0.000449	生活	Living	0.000463	笑	Smile	0.000793	球	Ball	0.002333	独家	Exclusives	0.004206
世界	World	0.000402	心	Heart	0.000454	快乐	Joy	0.000739	队	Team	0.001871	别人	Others	0.002803
社会	Society	0.000397	朋友	Friends	0.000426	家	Family	0.000656	北京	Beijing	0.001676	世界	Worlds	0.002654
家	Family	0.000394	时间	Time	0.000381	中国	China	0.000655	朋友	Friends	0.001623	幸福	Happiness	0.002554
生活	Living	0.000391	世界	Worlds	0.000379	拍	Shooting	0.000592	希望	Hope	0.001449	快乐	Joy	0.002299
写	Write	0.000385	公司	Company	0.000365	谢谢	Thanks	0.000588	明天	Tomorrow	0.001386	生命	Life	0.002186
朋友	Friends	0.000379	家	Family	0.000355	工作	Work	0.000579	时间	Time	0.001384	孩子	Kids	0.002135
日本	Japan	0.000376	成功	Success	0.000355	时间	Time	0.000550	笑	Smile	0.001345	朋友	Friends	0.001856
北京	Beijing	0.000376	工作	Work	0.000330	分享	Sharing	0.000547	快乐	Joy	0.001311	时间	Time	0.001778
心	Heart	0.000364	企业	Corporate	0.000322	北京	Beijing	0.000537	足球	Soccer	0.001211	中国	China	0.001759
钱	Money	0.000356	钱	Money	0.000318	开心	Fun	0.000529	家	Family	0.001209	工作	Work	0.001700
国家	Country	0.000339	美国	USA	0.000304	生活	Living	0.000512	球迷	Fans	0.001172	希望	Hope	0.001700
时间	Time	0.000335	孩子	Kids	0.000301	希望	Hope	0.000505	训练	Training	0.001170	烦恼	Trouble	0.001495
名	Fame	0.000334	北京	Beijing	0.000274	睡	Sleep	0.000478	加油	Cheering	0.001162	嘉宾	Guests	0.001390
人生	Life	0.000329	幸福	Happiness	0.000269	幸福	Happiness	0.000442	世界	World	0.001091	成功	Success	0.001390
经济	Economics	0.000326	分享	Sharing	0.000262	世界	World	0.000431	泪	Tears	0.000976	首播	Premiere	0.001383
工作	Work	0.000311	快乐	Joy	0.000256	电影	Movie	0.000403	冠军	Champions	0.000952	爱情	Love	0.001251
书	Book	0.000301	水	Water	0.000250	孩子	Kids	0.000401	分享	Sharing	0.000935	花	Flower	0.001249
教育	Education	0.000298	女人	Woman	0.000235	明天	Tomorrow	0.000401	睡	Sleep	0.000934	佛	Buddha	0.001155
政府	Government	0.000295	男人	Man	0.000228	可爱	Cute	0.000366	感谢	Thanks	0.000924	选择	Choice	0.001117
死	Death	0.000279	社会	Society	0.000224	歌	Songs	0.000355	决赛	Finals	0.000922	激情	Passion	0.001052
发现	Explore	0.000251	关注	Focus	0.000223	感觉	Feelings	0.000354	体育	Sports	0.000921	学会	Learned	0.001027
公司	Company	0.000249	车	Car	0.000220	生日	Birthday	0.000350	名	Ranking	0.000920	简单	Simple	0.001009

Note. The "Frequency" in the table = the total frequencies of a specific world of a specific category /the total frequencies of all words of a specific category.

total word coverage we used in this study did not eliminate the differences of occupation. Future research could focus on improving the algorithms of coverage. Second, our method in this study only tested Weibo microblogging and used intuitive occupation differences as the criterion. The real world creativity measurement (for a review and index, please see Batey, 2012; Pinheiro & Cruz, 2014) could be considered to test the validity of this method.

In summary, we employed a connectionist network viewpoint to reconstruct the usefulness and novelty dimensions of creativity, and found a method to calculate them with social media and Big Data in this study.

FUNDING

This research was supported by National Natural Science Foundation of China (Grant No. 71501105) and Beijing Social Science Fund (Grant No. 15ZXC020).

REFERENCES

Batey, M. (2012). The measurement of creativity: From definitional consensus to the introduction of a new heuristic framework. *Creativity Research Journal*, *24*, 55–65. doi:10.1080/10400419.2012.649181

Cohen, J. R., & Ferrari, J. R. (2010). Take some time to think this over: The relation between rumination, indecision, and creativity. *Creativity Research Journal*, *22*, 68–73. doi:10.1080/10400410903579601

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*, 407–428. doi:10.1037/0033-295X.82.6.407

Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. C. (2015). Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts*, *9*, 35–40. doi:10.1037/a0038688

Gardner, H. (1985). *Frames of mind: The theory of multiple intelligences*. New York, NY: Basic Books.

Glück, J., Ernst, R., & Unger, F. (2002). How Creatives define creativity: Definitions reflect different types of creativity. *Creativity Research Journal*, *14*, 55–67. doi:10.1207/S15326934CRJ1401_5

Guilford, J. P. (1967). *The nature of human intelligence*. New York, NY: McGraw-Hill.

Kogan, N. (2002). Careers in the performing arts: A psychological perspective. *Creativity Research Journal*, *14*, 1–16. doi:10.1207/S15326934CRJ1401_1

Lin, C. (2001). Syncretizing east-west education patterns, cultivating “T” talents. *Journal of Beijing Normal University*, *163*, 5–11.

Lin, D. (1998). An information-theoretic definition of similarity. *ICML*, *98*, 296–304.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge, UK: Cambridge University Press.

Paletz, S., & Peng, K. (2008). Implicit theories of creativity across cultures: Novelty and appropriateness in two product domains. *Journal of Cross-Cultural Psychology*, *39*, 286–302. doi:10.1177/0022022108315112

Paletz, S., & Peng, K. (2009). Problem finding and contradiction: Examining the relationship between naive dialectical thinking, ethnicity, and creativity. *Creativity Research Journal*, *21*, 139–151. doi:10.1080/10400410902858683

Park, G., Schwartz, A., Eichsteaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., … Seligman, M. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952. doi:10.1037/pspp0000020

Petocz, P., Reid, A., & Taylor, P. (2009). Thinking outside the square: Business students’ conceptions of creativity. *Creativity Research Journal*, *21*, 409–416. doi:10.1080/10400410903297998

Pinheiro, I. R., & Cruz, R. M. (2014). Mapping creativity: Creativity measurements network analysis. *Creativity Research Journal*, *26*, 263–275. doi:10.1080/10400419.2014.929404

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*, 92–96. doi:10.1080/10400419.2012.650092

Sowa, J. F. (1991). *Principles of semantic networks: Explorations in the representation of knowledge*. New York, NY: Morgan Kaufmann.

Verhaeghen, P., Joormann, J., & Khan, R. (2005). Why we sing the blues: The relation between self-reflective rumination, mood and creativity. *Emotion*, *5*, 226–232. doi:10.1037/1528-3542.5.2.226