Measuring Similarity Similarly: LDA and Human Perception

W. BEN TOWNE, CAROLYN P. ROSÉ, and JAMES D. HERBSLEB,
Carnegie Mellon University

Several intelligent technologies designed to improve navigability in and digestibility of text corpora use topic modeling such as the state-of-the-art Latent Dirichlet Allocation (LDA). This model and variants on it provide lower-dimensional document representations used in visualizations and in computing similarity between documents. This article contributes a method for validating such algorithms against human perceptions of similarity, especially applicable to contexts in which the algorithm is intended to support navigability between similar documents via dynamically generated hyperlinks. Such validation enables researchers to ground their methods in context of intended use instead of relying on assumptions of fit. In addition to the methodology, this article presents the results of an evaluation using a corpus of short documents and the LDA algorithm. We also present some analysis of potential causes of differences between cases in which this model matches human perceptions of similarity more or less well.

CCS Concepts: • **Information systems** → **Users and interactive retrieval**; **Similarity measures**; **Relevance assessment**; *Collaborative and social computing systems and tools*; *Association rules*; Clustering; Nearest-neighbor search; • **Human-centered computing** → *Laboratory experiments*; *Empirical studies in HCI*; Hypertext/hypermedia; • **Computing methodologies** → Language resources

Additional Key Words and Phrases: Perceived similarity, similarity metrics, algorithm validation

ACM Reference Format:
W. Ben Towne, Carolyn P. Rosé, and James D. Herbsleb. 2016. Measuring similarity similarly: LDA and human perception. ACM Trans. Intell. Syst. Technol. 8, 1, Article 7 (September 2016), 28 pages.
DOI: http://dx.doi.org/10.1145/2890510

1. INTRODUCTION

Platforms for collaborative work involving textual contributions from a large diversity of contributors are becoming increasingly prevalent and productive. Examples include platforms for online deliberation [Davies 2011], discussions around collaboratively edited products such as Wikipedia articles [Viégas et al. 2007] or open-source software patches [Tsay et al. 2014], blogs as a large-scale public venue for large-scale public discussion [Sunstein 2006], forums, venues for social support [Hwang et al. 2010; Dinakar et al. 2012], and idea management systems [Flynn et al. 2003]. Such platforms can capture innovative ideas from large, diverse populations, with potential to generate billions of dollars in value and address large-scale urgent problems such as disaster relief [Bailey and Horvitz 2010; Goggins et al. 2012]. Organizing these contributions so that readers can navigate through them is a challenge. This article provides a method and example evaluation of an automated link-creation algorithm, with some insight into its particular strengths and weaknesses.

This work is supported by the National Science Foundation's grants HCC-1302522 and HCC-1111750.
Authors' address: School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue Pittsburgh PA 15213.

2157-6904/2016/09-ART7
DOI: http://dx.doi.org/10.1145/2890510

Today, an increasingly large volume of information, work, and communication is based on, accomplished through, and/or represented by collections of text documents. These collections are increasingly often too large to read through one document at a time. Finding or synthesizing value from a subset of closely related documents can be especially difficult when navigation between documents is based on metadata such as a timestamp of last update, rather than topics contained within the documents. It can be extremely difficult for users, especially new users, to find desired information on such platforms, especially when the users themselves do not understand enough about the corpus or the pieces that they would find most interesting to specify a keyword query in a search engine. This creates a pressing need for tools supporting navigation within, and analysis of, such collections [Candan et al. 2012, pp.1–2; Liu et al. 2012, pp. 1–2; Cui et al. 2012, p. 1; Zhao et al. 2011, p. 2; Gretarsson et al. 2012, p. 2].

Research efforts to meet this pressing need are producing advanced technologies that facilitate text analyses [Liu et al. 2012, p. 2] and provide recommendations to support navigation and the discovery of desired information on sites with large volumes of text content [Zhao et al. 2011, p. 2]. Latent Dirichlet Allocation (LDA) appears to be at present one of the most commonly used—perhaps *the* most commonly used—analytic tool for this purpose. This use makes sense, however, only if the representation of texts that it yields enables algorithmic similarity computations (such as cosine similarity) having results that closely resemble similarity as experienced by human users when they compare the original texts. Yet, many of these technologies are based on measures of similarity between documents without clear validation that the algorithmic measures being used match human perceptions of similarity. Measuring how well various similarity measures match human perceptions is important when those measures are being used to help humans navigate through a large set of elements (e.g., in online communities [Spertus et al. 2005]). Understanding the limitations of a model also helps inform applications of it.

1.1. LDA

Topic models such as LDA [Blei et al. 2003] are often used to reduce dimensionality to themes that are useful building blocks for representing a gist of what a collection contains, statically or over time (e.g., Liu et al. [2012]). Without validation against human perceptions of similarity, the algorithm's usefulness for increasing navigability in large text collections by connecting documents that have similar mixtures of topics has been asserted (e.g., Steyvers and Griffiths [2007]). For example, LDA has been used to find connections between stories written by distressed teens, helping authors see that they are not alone in their plights [Dinakar et al. 2012]. In a 12-person evaluation, LDA selections were seen as closer matches and more helpful to original authors than those based on TF*IDF [Dinakar et al. 2012].

Prior work in conferences such as *CHI* and *CSCW* and journals such as this one has featured advanced technologies and visualizations supporting text analysis and better navigation. However, it is difficult to rigorously evaluate these tools against human perceptions of what the analysis is supposed to measure or support; many of these tools are described as supporting exploratory analysis. The authors in those cases generally do not focus on specific tasks, nor do they evaluate tools with tasks that have measurable, comparable outcomes. These tools also generally incorporate a host of other design decisions that may impact their usefulness toward the stated goal, assuming without testing that the document comparisons are a good match for human perceptions. Here, we build on prior work by providing such a validation method in the task context of hyperlinks for navigation among topically similar documents, and applying it to a state-of-the-art text similarity algorithm.

The literature describes many systems that use similarity comparisons based on LDA, whose utility depends on the match between LDA similarity and human similarity judgments. The Stanford Dissertation Browser, for example, uses cosine similarity of LDA topic vectors to compute text similarity and support navigation based on topic similarity to try to help users explore over 9,000 thesis abstracts. The article describing that work cited one of the field's shortcomings as a lack of validation mechanism or external ground truth to assess similarity measures [Chuang et al. 2012, p. 447]. Boyack et al. [2011] use LDA and other measures to compute pairwise similarity among 2.1 million biomedical publications (medical subject headings, titles, and abstracts), and acknowledged the value of human-based validation measures, operationalized there as connections through grant acknowledgments.

TopicNets also uses LDA to support exploration of a document corpus by providing an interactive graphical environment showing LDA topics that the documents are connected to and through. Evaluation of the human usefulness of the system is, in that article, done by the authors' generation of graphs, interacted with and explored by a few individuals with expert knowledge of the datasets (including authors) [Gretarsson et al. 2012].

Some systems use LDA to identify latent communities on platforms where users interact via text, and display the relationships among items based on these latent topic-based communities. For example, a system called Pharos uses LDA to model latent topics of text content and clusters documents based on their most-probable topic, illustrating communities and changes in them over time. In that work, evaluation consisting of ten users from IBM completing two identification tasks (of top authors and blog posts on a specific topic) with and without the tool being available to evaluate its usefulness [Zhao et al. 2011]. Yin et al. [2012] somewhat separate the concepts of topic and community with a many-to-many relationship between them, modeling community-based latent topics in user-generated content on social media sites. This method is evaluated by the authors' qualitative impressions and comparison to the results of other methods. The work by Zhang et al. [2007] discovers communities by modifying LDA for application to social graphs, testing the new method in research co-authorship networks and successfully identifying groupings of researchers within the same institution or research area. Introne and Drescher [2013] note that topic-modeling techniques do not support analysis of multiparty dialog well, in part because of the greater dynamism of dialog, and design an extension of its concepts to model communities of words over sequences of replies.

Relational Topic Models (RTMs) for Document Networks [Chang and Blei 2009] model documents and the links between them in a shared latent space of topics, with documents represented by standard LDA topic distributions, and links represented by distributions over the same LDA topics, numerically computed using the element-wise product of the documents' topic vectors. RTMs can be used to analyze linked corpora such as citation networks, hypertext, and explicitly networked user profiles (e.g., in social media). One of the RTM's novel contributions is that it can predict new links given words in a document new to the network without assuming a separate vocabulary for the links; RTMs constrain a document's words and its links to be explained by the same topics. Our work returns to the LDA origins of this model (i.e., LDA without other explicit connections between documents), looking at the use of these concepts on datasets that do not have links between documents to begin with.

Sizov [2012] extended LDA to estimate resource similarity based on both tags and geospatial data in support of automatically organizing, filtering, and recommending content on social media. A panel of five computer scientists working in the field of social media research found the new system to support those tasks well [Sizov 2012]. Other work in this journal has applied LDA to mobile phone location/activity data

to find patterns [Farrahi and Gatica-Perez 2011] or clusters of sociological interest [Joseph et al. 2014]; still other work in the same venue has focused on speeding up parallel computation of LDA models [Liu et al. 2011].

LDA is used widely, but assumptions about the extent to which the meaning attributed to topics is consistently represented in the texts that the model assigns the topic to are often unchecked. This motivates a need for new quantitative methods for measuring semantic meaning in inferred topics, based on human perception, as it is a quality not well measured by traditional algorithmic metrics [Chang et al. 2009].

1.2. TF*IDF/Cosine Similarity

Walter and Back [2013] use cosine of TF and TF*IDF vectors (described in more detail later), after stemming and stopword removal, to compute similarity of short ideas (average length 25 words) submitted on a crowdsourcing/ideation platform as a basis for "second level" K-means hierarchical clustering. They analyzed over 40,000 ideas in 112 contests and claim that text mining can be used to help automate the long expert-driven process of submission evaluation by identifying the most unique term distributions (equated to high quality). That article gives evidence that selections based on uniqueness of term distributions are somewhat predictive of contest winners selected by expert judges (maximum F_1 score 0.639). The authors suggest future work exploring "more sophisticated clustering" than simple term frequency (TF) or TF*IDF measures.

We include unigram TF*IDF cosine similarity as a measure for comparison later. As discussed in Section 5.7.2, we also include measures for comparing vectors other than the common and quickly computable cosine similarity measure. In a related setting helping users navigate through a large set of online communities, Spertus et al. [2005] empirically found that the cosine measure showed the best empirical results when compared against other measures. Informed by that work, this article uses the measure previously found to be best. In our own evaluation, we find that methods other than cosine still yield similar results for the comparisons evaluated here (see Sections 7.2 and 9.3 for details).

1.3. Outline of this Article

The problem of navigation in large text collections is important; topic modeling approaches such as LDA are frequently used to address that problem. What this body of work is missing is rigorous analysis of how the algorithmic similarity measures being used match up with human perceptions.

We contribute to this body of work by presenting a methodology for doing this evaluation as well as by applying that methodology to a medium-scale human validation of a state-of-the-art technique for analysis of similarity between documents in a corpus within a particular scope and setting. Specifically, the state-of-the-art analysis technique that we chose to evaluate here is a cosine similarity measure based on LDA topic modeling. We present results of this analysis and caveats to consider when using the algorithm. We also present results of methodology validation steps (e.g., interrater reliability), so that other researchers might be able to more easily evaluate proposed advances for this state-of-the-art analysis technique.

In the following sections, we review examples of idea management systems as one class of tools illustrating the problem that we hope to address and for which we hope to motivate a technical solution. We review work toward that technical solution and toward evaluation of those solutions. As graphically illustrated in Figure 1, we then describe a method for experimentally evaluating a technical solution, and apply this method to evaluate similarity judgments based on LDA. Further experiments and analyses help identify more specific conditions under which the algorithm performs better or worse. We also analyze the topics, documents, and reasons for which human

similarity judgments were most different from the LDA-based cosine similarity measure in order to understand the reasons for divergence. This evaluation is useful for others who may use or wish to use similar algorithms in their own work, as well as for those who wish to develop and test improvements to those algorithms, especially as they apply to increasing navigability.

2. A MOTIVATING EXAMPLE: IDEA MANAGEMENT

Idea management systems—one example for which topic modeling can be used to add link structure for navigability—are popular and growing. Competitive providers each report millions of users and hundreds of thousands of submitted ideas [IdeaScale 2013]. Microsoft, IBM, Dell, Whirlpool, and UBS have used this approach to tap the knowledge of their employees and customers [Bailey and Horvitz 2010]. The International Conference on Communities and Technologies focused a workshop on large-scale idea management and deliberation systems specifically [Convertino 2013].

One common approach is to have a diverse crowd first generate many contributions through an open brainstorm-like process dominated by generation and divergent thinking, followed by a review and sometimes narrowing of those ideas. The convergent data reduction stage is made easier by tools for organizing and grouping or connecting related ideas, usually based on the ability to measure similarity. In some platforms, there is no specific timeframe distinguishing divergent and convergent phases, which may happen in different parts of the discussion space simultaneously.

For example, IBM hosts multiday online "Jams" to solve business challenges, clarify company values, and produce ideas for new initiatives or improved operations. A 2006 Innovation Jam, described as the largest-ever online effort to advance technological innovation, involved 150,000 people from 104 countries, investing $100 million in the best ideas from the event [Bjelland and Wood 2008; IBM 2012]. It generated billions of dollars in revenue, followed by another Innovation Jam in 2008 and a spin-off consulting service [Bjelland and Wood 2008; IBM 2012].

Primary goals for that Jam included connecting people in an exciting way that helped them build on each other's ideas and create something new and innovative [IBM 2012]. The actual experience, however, fell short. Contributors were not constructively building on each other's postings, and new visions and connections did not emerge until the manual review process after the event [Bjelland and Wood 2008].

This review process required teams of specialists and senior executives to spend weeks sorting through gigabytes of text, to pick a few new ideas from tens of thousands of postings. This mostly-manual process did help IBM listen to already-circulating ideas and combine related ideas in major new initiatives, producing business success. However, extracting value from the ideas required a great deal of management time, and participants were not readily able to find and build on one another's ideas or connect directly to implement them [Bjelland and Wood 2008].

After Japan's triple disaster in the spring of 2011, IBM adapted their jam to an issue-specific forum that drew voluntary contributions from 275 employees in 23 countries on seven topics, each at the root of a tree-structured, text-based discussion, with an average of 100 responses per topic, in just a few days [Muller and Chua 2012]. Leveraging the power of distributed knowledge and efforts to respond to natural disasters is becoming increasingly necessary, and requires innovation with empirical grounding toward rapidly, dynamically structuring information contributed after such events, so that contributors and officials can more easily make sense of the situation and act appropriately [Palen et al. 2010, p. 6].

Bailey and Horvitz [2010] specifically note that comments on ideas commonly try to link the author/idea and other people, efforts, or ideas related to the one posted, that these connections are an important but often unmet expectation motivating

contribution, and that automatically attaching this kind of link information to ideas could be useful. Towne and Herbsleb [2012] identify navigability as one of the key challenges for online deliberation.

In many idea-management systems, including the one that we implemented to collect data used in this article, ideas are often organized by post time. Another common option is to list by upvote count, but this may not correlate with quality if the higher vote count is also a result of higher list position [Salganik et al. 2006]. Tagging or categorization is sometimes done by the contributor at the time of the submission [Bailey and Horvitz 2010]. However, this suffers from polysemy (a tag may have many meanings) and synonymy (different tags may mean the same thing), and requires that the contributor know which tags may be appropriate in the corpus and take the time to apply them. In each of these approaches, it is difficult to find, connect, consider, and/or synthesize ideas based on topic relationships.

Processing through such ideas, the aim is often data reduction, such as finding the best ideas, or identifying cross-cutting concerns or themes that may be present in many ideas. Another key aim is discovering deep knowledge embedded in the data, not just the answers to questions that a user thought to ask [Nahm 2004, pp. 1–2]. Synergy between already-contributed ideas, if discoverable, can lead to new and better contributions, and is arguably the source of most good ideas [Johnson 2010]. Each of these goals is supported by better navigation between related contributions.

Algorithmic help for navigating along conceptual adjacencies can be useful [Johnson 2010] as long as human readers perceive conceptual connection in those links. The links would also need to be dynamically generated in order to rapidly incorporate new information, which may arrive in the form of new documents and confirmation or disconfirmation of machine-generated link suggestions by humans who are navigating through, curating, and improving on the dataset. Other work assumes that users have well-defined views of similarity for the purposes of organization, and assumes that the computer can learn these (e.g., Huang and Mitchell [2007]).

3. EVALUATING SIMILARITY MEASURES

Observation that statistical tasks such as unsupervised classification are "still largely based on ad-hoc distance measures with often no explicit statistical justification" has previously motivated works exploring statistical properties of those measures and mathematically demonstrated suitability of use for certain applications (e.g., El-Yaniv et al. [1997] and Lin [1991]). We observe that, analogously, tools supporting human tasks such as navigation in a text corpus are still largely based on computed distance measures with often no explicit experimental validation against human perceptions, and are motivated to make the present contribution sharing a method for such testing as well as results for a commonly used measure.

Automatic methods for adding structure by linking related documents rely on having a similarity measure that matches human perceptions of similarity reasonably well. This match should be empirically validated for the intended application and tested against reliable human judgments of similarity, with differences investigated and understood. We develop and apply a method to investigate how well an algorithmic measure correlates with the links that humans would make, and under what conditions it performs more or less well. We illustrate the method with evaluation of a commonly used LDA-based similarity measure, the cosine similarity of LDA representations of the original texts.

Some studies have examined the match between algorithmic and human perceptions of similarity using a stand-in such as a human-built hierarchical Open Directory instead of direct evaluation (e.g., Haveliwala et al. [2002]). Others have used the

DARPA[1]-organized TREC[2] collections with an abstracted retrieval task ranking document sets as more or less relevant (according to human judges at NIST[3]) to given information-need statements [Voorhees 2007] rather than to each other. Mani [2001] presents several methods for evaluating the match between a long document and the summary of it, or between two summaries. Dinakar et al. [2012] state that an approach using LDA performed better than TF-IDF in a story-matching task for which participants completed several two-item pairwise similarity evaluations.

At the individual word level, Faruqui et al. [2015] computed the cosine similarity between feature vectors that incorporated information from lexicons and large corpora, and compared these algorithms against benchmark datasets containing pairs of English words that had been assigned similarity ratings by humans. The Spearman correlation of these two similarity scores (human and algorithmic) was used as the algorithms' primary evaluation measure, and feature vectors with a higher correlation were said to produce better results than those with a lower correlation. The interquartile range of Spearman correlations reported in that article's Tables II through IV is (0.580,0.737).

The work closest to ours in task and goals is the investigation by Lee et al. [2005], in which students rated pairs of short news stories on a 5-point relatedness scale. The authors report a human-rater-agreement level of 0.605, computed as the mean correlation between each rating and the others for that pair. They evaluate a number of similarity algorithms and find that the best simple models have a correlation of about 0.5 with human judgments. Their most complex comparison is cosine similarity based on Latent Semantic Analysis (LSA), trained on a 364-document corpus. The best LSA models correlated about 0.6 with human judgments. In general, they find that the best models can detect only a subset of the highly similar document pairs, which suggests a need for alternative, more nuanced, models of text document similarity. Our work measures human perception of similarity in more detail and with a larger, more diverse group of human evaluators, and evaluates an LDA-based cosine similarity measure against human judgment.

4. EXPERIMENTS: OVERVIEW

We present a series of experiments designed to explore the connection between LDA and human conceptions of similarity, as illustrated in Figure 1. Following an overview here, Section 5 describes models and materials selection. Sections 6 through 10 provide more detail about each experiment.

We begin with a task that models selection of links between ideas, as might be used to help address the problem discussed in Section 2: given three documents, we ask people which two are most similar and to explain why in free text. We construct the sets of three in ways that let us assess agreement, but more importantly, explore the sources of failure of algorithmic similarity to match human judgment. We construct sets of three documents that always contain a second document that is highly algorithmically similar to the first, and a third that the algorithm ranks as rather different from the first. However, those three documents are randomly sequenced for presentation.

Qualitatively exploring the dimensions of similarity that people used, our results suggest that some ratings are not based on topic content. We explore these other bases of similarity judgments and the impact that they have on human–algorithm agreement.

Based in part on these observations, we sought to explore human perceptions of similarity from multiple perspectives, producing a highly reliable multi-item human

[1]Defense Advanced Research Projects Agency (USA).
[2]Text REtrieval Conference.
[3]National Institute of Standards and Technology (USA).

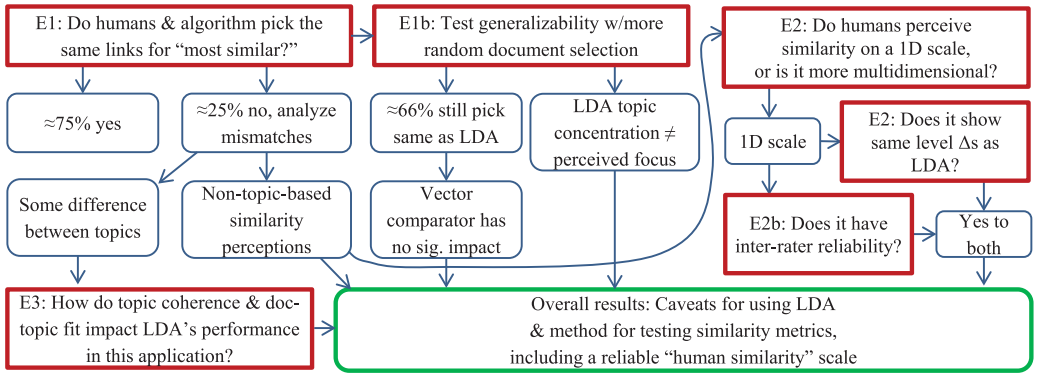

Fig. 1. Overview of the Experiments.

similarity measurement scale, further validating interrater reliability in a smaller test (Experiment 2b). We compared human similarity ratings of pairs of documents with algorithmic similarity scores and found generally high correlations.

In the first experiment, when the humans and algorithm disagreed about choosing which two of three documents were most similar, those disagreements were not uniformly common across all of the topics. We sought to better understand what conditions or aspects of those topics might affect the agreement between algorithmic and human similarity. We conducted a third experiment to gain this understanding, and found that agreement is impacted both by topic coherence and the strength of fit between a document and its most closely associated model topic.

5. MODELS AND MATERIALS SELECTION

5.1. Data

The input dataset for this article comprises all 10,331 ideas submitted to the 2012 President's SAVE (Securing Americans' Value and Efficiency) Award, which, at the time of the study, were publicly visible at saveaward2012.ideascale.com. U.S. President Obama began the annual SAVE award in 2009 as "a process through which every government worker can submit their ideas for how their agency can save money and perform better" [IdeaScale 2012]. Most ideas in the corpus are a paragraph or a few paragraphs, written by government employees during a short contest-based solicitation period. They span a wide range of quality and feasibility. Contributors have the opportunity to have their idea heard, recognized, and possibly implemented. Contest administrators read through the submitted ideas, selecting four finalists put to popular vote at WhiteHouse.gov/Save-Award, and the winner is brought to Washington, DC for presidential recognition.

We chose this dataset because it is a real-world instance of our motivating example, which realistically represents the scale and diversity of results that might be expected from an ideation process in a population the size of the US Federal Government. The motivation behind the ideation and award (reducing expenditures of public money) are noncontroversial, especially compared to other government-related topics, even while particular submitted ideas might be less neutral, as might be expected among contributions to other large-scale collaboration platforms. The corpus scale is comparable to the Stanford Dissertation Browser's dataset of just over 9,000 thesis abstracts [Chuang et al. 2012].

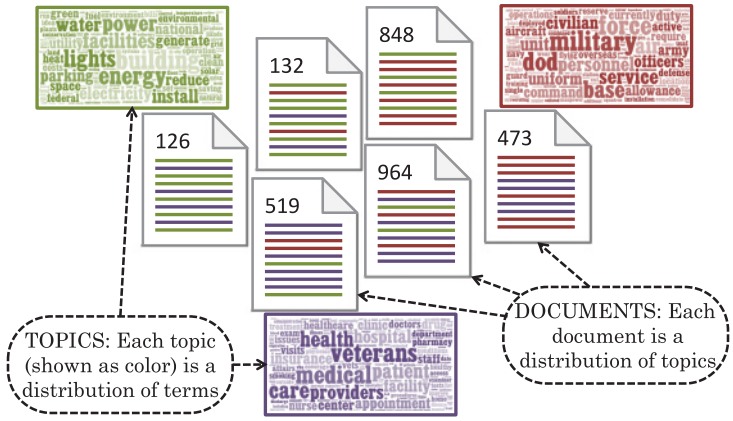

Fig. 2. LDA generative model. Each topic (here, a color) is a distribution of terms; each document is a distribution of topics. Information embedded in the ordering of terms, sentences, and so on, is generally discarded; the model instead focuses on information about relative proportions ("bag-of-words").

5.2. Preprocessing

Our documents each contain the title and body of one of the 10,331 ideas, excluding any comments, author information, or tags that were manually added in the original data. Knowing that our results do not rely on such additional features, which may not be present in all other systems that can build on this work, increases generalizability. We set text to lowercase and removed stopwords before algorithmic processing.

5.3. Topic Modeling Algorithm

We chose a statistical technique that models a whole corpus of text, including its commonalities and themes. This technique supports comparing documents to each other within the context of that corpus, which is appropriate for the task of generating within-corpus navigational links based on topic similarity. This is contrasted with representations used to compare documents independently of context, such as term frequency vectors. Probabilistic topic modeling such as LDA [Blei et al. 2003; Steyvers and Griffiths 2007] is one such state-of-the-art technique, and widely used, including in applications described earlier. More advanced techniques, including those not yet invented, might be evaluated using the human-centered methods described in this article.

As graphically illustrated in Figure 2, LDA models each "document" (shown as a page) as a multinomial distribution over "topics" (shown as colors) and each "topic" (shown in a colored box) as a multinomial distribution over terms. It assumes a model of document generation in which, for each term, a topic is first sampled from the topic distribution for that document, then a term is sampled from the term distribution for that topic. Users of LDA often assume that the distribution of topics provides a useful view either instead of or in addition to the word distribution. Similarity between two documents can be measured in terms of those topics [Steyvers and Griffiths 2007]. This measure may be more accurate for documents that focus on those topics than for documents covering unusual content [Dinakar et al. 2012].

We ran an LDA model using program default hyperparameters. These were $\beta = 0.01$, as suggested in Steyvers and Griffiths [2007], and $\alpha = 1$, which is in the same range as suggested by Steyvers and Griffiths [2007] and which does not bias the model toward smoothness nor sparsity [Steyvers and Griffiths 2007].

As noted by Singh et al. [2012, p.143], "Selecting the right number of topics is an important problem in topic modeling." Consistent with the goal of Xu and Ma [2006, p. 303] of maximizing dissimilarity between clusters, we selected the number of topics by choosing the model with the lowest average percentage of documents that has neither or both topics in each possible topic pair. This evaluation metric readily distinguished a 21-topic model from the tested range of 5 to 300 topics as the model which maximally separated documents into different topics.

5.4. Sampling Strategy

Our strategy for sampling documents was designed to serve several needs. First, our motivating application—linking documents in a corpus—suggests that we focus primarily on how well high-similarity as compared to low-similarity ratings match human judgment, because automatically generated navigational links are likely to be selected from the highest-similarity ratings. We selected pairs from qualitatively different levels of "high" and "low" algorithmic similarity, as well as "low similarity with the same most-probable topic," to have clear differences between levels of LDA-based similarity and thus clarify interpretation of differences in experimentally measured results. We relaxed this in Experiment 1b, which tests for generalizability.

Intentionally choosing pairs with widely different levels of similarity makes our results less sensitive to particulars of the topic model, corpus, and similarity measure's distribution in the middle of its range. These larger differences between tested similarity levels increase the chance that human similarity judgments would agree with the algorithm, producing a rough upper bound on that agreement and giving us an interesting set of failure cases (i.e., more clearly identifying where algorithmic and human similarity judgments substantially diverged). We investigated these failure cases to improve understanding of LDA's fundamental limitations. Experiment 1b steps back from the rough upper bound to provide a more general test that is more sensitive to mid-range values.

Second, we wanted to sample enough documents within each topic to investigate interesting within-topic phenomena, while having sufficient breadth across the range of topics in the corpus. This was also a goal of the sampling strategy used by Lee et al. [2005], in which a total of 50 news articles were chosen to represent a handful of topic clusters "in an attempt to ensure a broader spread of human judgments of document similarity" [Pincombe 2004, p. 11].

In pursuit of this goal, for experiments other than Experiment 1b, we chose documents strongly associated with a small but diverse set of distinct topics, as described in Section 5.5. This aspect of the method helps ensure that results of testing against human similarity on a subsample covers the breadth of topics covered by the corpus. (Experiment 1b does this by random selection.) This aspect also helps clarify interpretations in Experiment 3; had we selected topics by other criteria (e.g., randomly, as in Experiment 1b), we might have wound up with two or more similar topics and not been able to get as much unique value out of testing each one. When the marginal benefits (potential knowledge gains) from exploring an additional topic are proportional to how unique that topic is compared to what has already been explored, this diversity-based sampling strategy maximizes expected knowledge gain in a limited experimental budget.

Choosing documents most strongly dominated by those topics allowed us to more easily look for relationships between characteristics of topics and the degree of agreement between algorithmic similarity and human judgment, our third sampling need. We acknowledge that this strategy limits generalization away from documents that may not have a clear dominant topic in the model (effects explored in Experiments 1b and 3), but feel that the gain in clarity for how judgments about documents reflect on their

dominant topics was worth the trade-off. In Experiment 1b, we removed consideration of dominant topics from the selection process to test generalizability.

Finally, we wanted multiple human judgments per document pair in the experiment so that (1) we could better perform reliability assessment of our similarity measurement scale, and (2) so that the free text fields could be completed multiple times, allowing us to separate common from idiosyncratic content. Accepting costs associated with human subject experiments meant that we would have to select a relatively small minority of topics and documents. Again, we judged this a trade-off worth making: while reducing the number of topics or documents sampled in the experiment may reduce generality, this concentration increases our ability to look at particular judgments in much more detail and more reliably. This was the same decision and reasoning as used by Lee et al. [2005], for the same reasons [Pincombe 2004, p. 11]. This approach is complemented by our generalizability test in Experiment 1b, which optimizes for covering more content rather than getting repeated measurements of the same document sets.

5.5. Focal Topic Selection

Because a human subjects experiment does not allow us to efficiently explore all 10,000+ ideas in depth, we needed to choose a smaller number of documents to include in our test set. We wanted to ensure topic diversity while also having enough documents within each specific topic to be able to investigate any differences between those topics. This aspect of the selection strategy allowed us to test for and observe between-topic differences in Experiments 1 and 2 that we could investigate further in Experiment 3. This section describes the method we used to select a diversity of distinct topics.

Each topic is a vector of weights for the unique terms in the corpus. We computed the cosine similarity measure between these term-based probability distributions and, for each topic, computed its average cosine similarity with the other topics. We chose the five topics lowest on this measure, most unique from the other topics, on average. In order, these five topics were about veterans' health care (topic ID #4), the military (#14), public social services (for food, education, and the like, #1), reducing building energy use (#3), and Social Security benefits (#20), as labeled by manual inspection of the topics' term distributions.

5.6. Focal Document Selection

Having selected topics as described in the previous section, we then selected "focal" documents from only the 27.4% of documents in which the highest-weighted topic was one of those five. We focused on documents that were strongly dominated by the focal topics, as measured by the difference in probability between the highest-weighted topic (by definition, one of the focal topics) and the second-highest-weighted topic. For each of the five focal topics, we chose the five documents highest in this difference measure. This produced a set of 25 focal documents, each clearly associated with one of the topics picked in the prior step, so that the set of 25 focal documents represented a small set of distinct topics from across the dataset.

Documents that were duplicates of other documents or had no understandable English content (one document) were eliminated and replaced with the next document on the list, iteratively. These quality filters caused 44 documents from one author, 11 from a second author, and 6 from a third author to be disqualified from use in this experiment.

5.7. Computing Document–Document Similarity

5.7.1. LDA/Cosine. After selecting focal documents, we needed to select documents that the similarity algorithm computed were very similar as well as those the algorithm

computed were not very similar, at qualitatively different levels, to be able to compare these differences with humans' judgment and measure whether or not the algorithmic measure matched human perceptions, at least at these coarse levels.

We represented each document in the corpus as a vector of 21 topic probabilities, which summed to 1 and was not smoothed (i.e., if a topic was not assigned to any tokens in that document, it had a probability = 0). We computed the cosine similarity between all possible pairs of document vectors. For each document, this produced a scored ordering of the other documents based on the degree to which they contained the same distribution of topics.

For most documents, a few others were fairly similar, followed by many documents with lower similarity scores. Giving all pairs equal chance of selection could produce mostly low similarity ratings because of this "long tail." In order to follow our criteria of creating a task with clear differences in the LDA similarity measure, and also focusing on high similarity ratings, we first chose the document that was computed as *most* similar to each of the focal documents, labeling these as "near" documents. If the selection was a focal document or met the disqualification criteria stated earlier, it was removed and replaced with the next in sequence, as before. A single "near" document could be selected multiple times.[4] Since the task focuses on similarity judgments of a document pair, we felt that allowing a document to appear in more than one pair was not a problem.

For each focal document, we then separated the corpus into two groups: one group of documents containing the same highest-weighted topic, and a larger group of all other documents. To identify documents with clearly different levels of LDA-based similarity, we filtered to just the 20% (plus any documents tied with those in that bottom 20%) of each group that was *least* similar to the focal document (by the same measure), and randomly selected one document from each group. We call these the "distant/same lead topic" and "distant/different lead topic" documents, respectively. These choices allow us to examine how much the single highest-weighted topic helps predict document similarity, simulating applications that link documents based on shared membership in a single category or cluster. The full experimental set includes 90 documents, with a median length of 39 tokens (excluding stopwords). Experiment 1b removes the 20% filter and uses randomly selected "distant" documents, as described in Section 7.1.

5.7.2. Other Measures. We also computed document–document similarity using two *feature vector* alternatives to LDA topic weights, based on the conceptually simpler TF*IDF measure, with and without Porter stemming (as two different ways of constructing feature vectors).[5] TF*IDF, which is widely used, represents each document as a vector the size of the corpus vocabulary, in which the weight for each of those values is directly proportional to how frequently the term appears in the document (TF) and inversely proportional to how frequently a word appears in the corpus overall (IDF). The latter term means that words that are common across very many documents have less weight. Words that are common and meaningless enough to be considered stopwords (a, and, the, ...) have weight set to 0 by removal, which accelerates processing. When using these alternative *feature vectors*, we use the cosine similarity measure, like Walter and Back [2013]. Using a feature set size as large as the vocabulary means that when people use different terms to refer to the same concept (e.g., "aid," "assistance," "help"), the TF*IDF measure does not draw this connection, and when people use the same term to refer to multiple concepts (e.g., "bank," "book," "close"), a false connection

[4]Two "near" documents were most similar to three "focal" documents each; five "near" documents were most similar to two "focal" documents each, and nine "near" documents were each most similar to a single focal document.

[5]We used interactive RapidMiner for TF*IDF and source-code Lingpipe for LDA implementation.

is detected. Stemming reduces dimensionality by collapsing different forms of the same word, for example, detecting a connection between one document referring to "service" and another referring to "services." While TF*IDF is parameterless and conceptually simpler, other studies (e.g., Boyack et al. [2011]) have found topic modeling to have superior performance.

Several efforts to improve on LDA topic modeling have been made, but no specific advance seems to have yet gained comparably widespread adoption in the research literature, and to the extent that they are still based on LDA, most would be likely to share fundamental limitations of LDA explored in this study. The experimental methods described in this article can be repeated in future work with more advanced models to test new developments, though we would expect some of the inherent limitations of LDA (e.g., similarity judgments not based on topics) to still be observed in more advanced models unless the advance specifically addresses one or more of those limitations.

We also used the LDA feature vectors with four *measures for comparing vectors* alternative to cosine similarity. After cosine, we used the symmetrized Kullback-Leibler (KL) divergence (1/2 the J measure of Equation (2.2) in Lin [1991]) with additive smoothing (adding 0.000001 to all values and rescaling) to overcome the issue that these divergences are not defined when one vector has some zero values. Third, we used the related symmetrized Jensen-Shannon (JS) divergence (1/2 the L measure of Equation (3.4) in Lin [1991][6]). Fourth, we computed Euclidian distance. Steyvers and Griffiths [2007, p. 443] suggest all four of these methods[7] for computing similarity between documents.

5.8. Participants and Filters

Participants were recruited to the experiment website via Amazon's Mechanical Turk (mTurk), a paid microtask crowdsourcing platform. Since the documents discussed concepts specific to the US Federal Government, participation was limited to those who were in the United States according to their Mechanical Turk profile and a GeoIP lookup confirming access from the United States for at least one participation in a given experiment, and who had had at least 500 assignments approved by other requesters (to help with quality control). We filtered out data from surveys that were not submitted as complete, as well as any in Experiment 1 in which no pair of documents was selected. We also removed from analysis those by participants who submitted two or more blank or copy/paste form fields (idea summaries, or the explanations of similarity and difference) in the same assignment. We manually reviewed the fastest submissions, but did not find any compelling reason to remove further submissions based on speed in Experiments 1 and 2. Work was discarded from any participant who, on both pages of at least one of their Experiment 3 responses, gave the same answer to all questions.

[6]This is noted for full disclosure of exactly which formula we computed, but because we use only rank information, neither the 1/2 factor nor the square root transformation suggested by Endres and Schindelin [2003] should have any impact on our work. The 1\2 is suggested by (a) the concept of averaging two asymmetric divergence measures to create a symmetric measure, (b) Definition 1/Equation (4) in El-Yaniv et al. [1997], (c) the definition at the top of p. 1860 in Endres and Schindelin [2003], and (d) mathematical derivation from Equations (4.1) and (5.1) in Lin [1991].

[7]The printed edition of Steyvers and Griffiths [2007, p. 443] uses a different definition of symmetrized KL-divergence than the one (a) used here, (b) found in other sources, and (c) found in the version posted on an author's website. We believe this is just a print error. This source suggests dot product as well as cosine similarity; we use only the latter (dot product scaled by magnitude) to focus on comparing documents based on their mix of topics, reducing variance caused by length differences between short ideas.

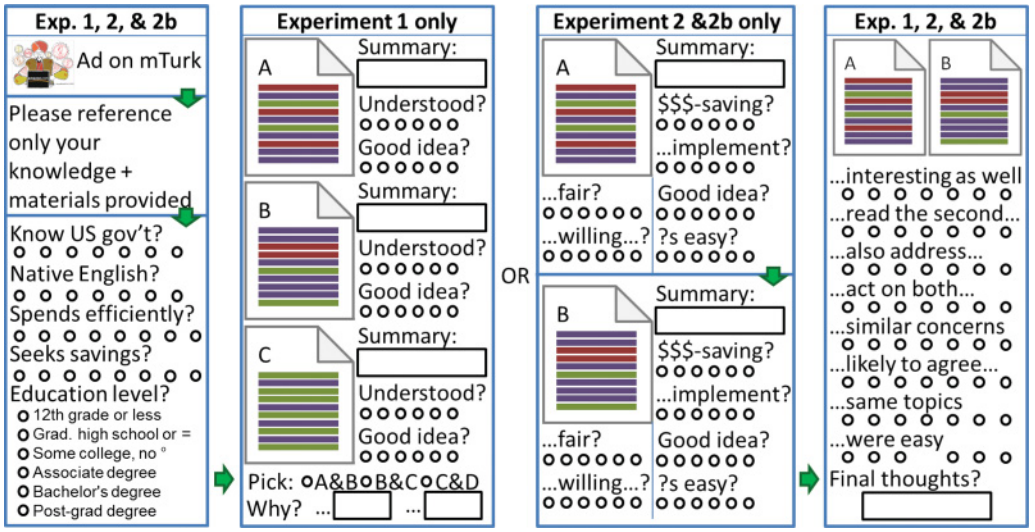

Fig. 3. Schematic of primary pages showing page sequence and question summaries, Experiments 1 to 2b. Here, the last page shows that the participant selected A&B as most similar on the second page of Experiment 1.

On arriving at our Experiments 1 or 2 for the first time, participants were asked a short page of demographic and background questions, each a 7-point Likert item plus "I don't know" for the latter two:

—I am knowledgeable about the US Federal Government.
—English is my native language.
—The US Federal Government spends money efficiently.
—The US Federal Government looks for opportunities to save money.

A standard 6-level educational attainment item followed. Experimental questions, as described later, followed on separate pages (see Figure 3).

At the end of all experiments, participants had an optional free text box for anything else that they wished to tell us. The most frequent comments were variants of "thank you" and "no comments." They received a code for payment as advertised, generally US$.40 in Experiments 1 and 2 and US$.15 for Experiment 3, which had a proportionally shorter task.[8] Participants in Experiment 2b, which was designed to test interrater reliability, received 90% of their payment in the form of a bonus after having completed all 15 instances of the task that we made available, as advertised.

Likert items in all three experiments were coded for analysis from 1 to 7, for which 1 = strongly disagree and 7 = strongly agree. In analyses in which items are dichotomized, the "agree" options are coded as "1," the "disagree" options are coded as "0," and there is no neutral. All experiments except Experiment 1b used the same 90 documents selected via the process described earlier.

[8]Because of the additional questions and slightly longer documents resulting from random selection, we increased pay by $.10 for most participants in Experiment 1b, with a 10x participation limit in that experiment.

6. EXPERIMENT 1

6.1. Experiment 1: Methods

Our first experiment explores human-perceived dimensions of similarity and the degree to which the algorithmic measure captures that perception of similarity, for use in the application environment of links for navigability. In Experiment 1, participants were presented with a focal document, its associated "near" document, and one of the two associated "distant" documents, chosen randomly. These were presented in random order but labeled A, B, and C in presentation order, as schematically illustrated in Figure 3. For each, participants were asked to summarize "What is this about?" as an incentive to read the documents carefully [Kittur et al. 2008], and respond to two six-point Likert-style items: "I understood this passage" and "I think this is a good idea" with options of {strongly, moderately, slightly} {agree, disagree}. They were then asked to indicate "which pair of documents seems most closely related and different from the third," analogous to the link creation task that we are most interested in. This task is often called triading when employed in usability testing [Bank and Cao 2015]. The number of times that participants select a particular pair as being the most similar, as a percentage of the times they could have picked that pair, is a primary dependent variable of interest. We experimentally compare it to the LDA-based predictions to see how well LDA performs.

In two free-text fields immediately following this selection, participants were asked (1) what the two have in common as opposed to the third, and (2) what makes the third different from the other two. These questions and all three documents were visible on a single webpage. On one following page, participants answered a pilot version of the human similarity scale used in Experiment 2, using the two documents that they had just selected as being most similar.

6.2. Experiment 1: Topic-Based Similarity Determination?

To examine whether or not people used topic similarity as the basis for their judgments in this task, we examined the explanations people gave as to why the chosen two documents were similar and why the third one was different. A binary code was applied to responses, as a 1 if the primary distinction described that two were about the same topic and the third about another topic, for example, "A&B talk about VA hospitals. [The other document] is about phones, which is an entirely different topic." If other reasons were given, such as understandability or idea quality, or when people focused on the beneficiaries of an idea, or its level of detail, scope, or the approach that the ideas were taking, that was coded as 0. For examples, "The first 2 are avoiding paying more money out. The third is a way to generate new income" shows a perception of similarity based on approach, and was coded as 0. "Brevity. The third is different from the other two in that it has a long and thorough explanation" shows a perception based on style, and was coded as 0. One author of this article coded all responses for analysis. To check for interrater reliability, another received codebook instructions and coded 115 response pairs, with 102 agreements and an acceptable Cohen's Kappa of 0.750. Results are shown in Table I.

6.3. Experiment 1: Results

In Experiment 1, we had 562 results from 413 unique participants. Participants indicated that they used topics as their primary basis for perception of similarity in 72.8% of responses. As shown by the boldfaced number in the upper right corner of Table I, just over 75% chose the same pair as the LDA-based similarity metric as being the most similar; a higher percentage agreed with LDA when choosing based on topics.

Table I. Responses to "Which of These Two Seem Most Closely Related and Different From the Third?"

Document pair	Topic-based assessment	Other aspects	Overall
"Focal" w/"Near" (matches LDA)	344 (84.1%)	78 (51.0%)	422 (**75.1%**)
"Distant" w/"Near"	37 (9.0%)	35 (22.9%)	72 (**12.8%**)
"Distant" w/"Focal"	28 (6.8%)	40 (26.1%)	68 (**12.1%**)
Total	409	153	562

Note: Percentages are of column totals. If responses were random, all cells would be ≈33%.

This might be seen as an upper bound for this similarity measure's predictive validity in building links.

Also as shown in bold in Table I. Those who did not choose the same pair as LDA were about evenly split between pairing the "distant" document with the "focal" or the "near" document. Interestingly, even among those who described the basis for their similarity choices as something other than topics, a slight majority still selected the same pair as LDA, suggesting that the topic model may still pick up vocabulary indicative of similarity that people do not semantically consider primarily about "topics" (e.g., beneficiaries or approaches). There were no significant differences in the demographic questions between people who chose the same pair as LDA and those who did not.

In the next two subsections, we analyze the possible roles of document understandability and perceived idea quality in explaining selections that did not match LDA predictions, also motivating other experiments as illustrated in Figure 1.

6.4. Experiment 1: Understandability

When reading through descriptions of why people selected a pair as most similar, we found that some people wrote that the selected pair were similar because "They are easy to understand" while the other was different because "I have no idea what they're saying" or inversely, that the selected pair were similar because "they're both incomprehensible" and the third was different because "I actually understood what the person was suggesting." A rough coding indicated that about 25 of the description pairs were related to understandability, like these examples. We then analyzed responses to "I understood this passage" and found that most people understood most documents: of the 1666 responses to this question, only 3.3% were "strongly disagree," with less than 5% in each of the other "disagree" categories; the highest levels of understanding were the most frequent. Across the 90 documents, average understanding scores had a mean of 5.888 and a standard deviation of 0.94582. Free-text summaries indicated that even documents lower on this measure were reasonably well understood.

There were 69 cases in which participants understood exactly one or two of the documents (as measured by the dichotomized "understanding" item), and the two matched their selection of which two documents were the most similar. More of these were cases in which the person disagreed with LDA than would be expected if the variables were independent, by a Pearson chi square test ($p = 0.016$).

6.5. Experiment 1: Idea Quality

Other descriptions explaining participants' similarity choices focused on the idea quality or some aspect of it. For example, one response noted that "A and B seemed like more thought out ideas" and the third is different because "It is not written well and is not a clear idea"; another wrote the inverse, that the selected two are similar because "Both are harbrained [sic] crazy ideas that really don't make much sense." Approximately 43 comment pairs appeared to focus primarily on idea quality. We then analyzed dichotomized responses to "This is a good idea" and found 119 cases in which exactly one or two of the ideas were seen as good and the two matched their selection for "most similar." As with understanding, a disproportionately large share of these were in cases

in which the selection disagreed with LDA, according to a Pearson chi square test ($p = 0.045$).

Of the 140 "most similar" pair selections that differed from the topic-based LDA prediction, 51 could potentially be explained by these measures of similarity in understandability or idea quality. This analysis helps us understand some factors that limit how well LDA can agree with human perceptions of similarity.

7. EXPERIMENT 1B: GENERALIZABILITY

7.1. Experiment 1b: Methods

To measure human and algorithmic agreement across a broader range, we randomly selected 10% of the documents as focal documents for a generalizability experiment (1b). We did this by randomly shuffling documents and taking the first 1,033, skipping over documents disqualified as described in Section 5.6. For the "near" document, we selected the most closely related document according to LDA, randomly choosing one of the four measures for comparing vectors discussed at the end of Section 5.7. In this experiment, we randomly selected the "distant" document from among all nonfocal documents instead of from the most-distant 20% so that the comparison between human and algorithmic similarity might be made over the range more broadly, and to remove any dependence on the algorithmic similarity measure for this part of the triad selection. We then repeated the experiment described earlier.

In our other experiments, we only selected documents that are strongly associated with a particular topic, for reasons described in Section 5.4 (similar to Lee et al. [2005] and Pincombe [2004]) but which produce a rough upper bound on agreement. Randomly selecting focal documents helps assess agreement more generally by including documents that have a more "flat" or uniform distribution of topics. If this "flatness" means that the documents are less focused, it may be hard for humans to compare with others. If, on the other hand, it means that the document is focused but the focus is not aligned with the topic model, humans might be able to make reasonable comparisons even when the algorithm might have difficulty; it is at least a different range of values than was tested earlier. To help distinguish, we asked people to indicate their level of agreement/disagreement with "This focus of this idea is clearly expressed" as with understandability and idea quality.

To test whether or not participants' selections (of which pair in three is most closely related) depends on the method used to compare topic vectors and select the "near" document, we used Pearson's chi-square to test a null hypothesis that those two factors are independent. The chi-square test differs from other tests in that it can be used to confirm that a null hypothesis is correct [Privitera 2015, p. 298].

Conventionally, assessments of Type II error (incorrectly retaining a null hypothesis) target an 80% power (probability of detecting an effect if present) [Cohen 1992], considering mistaken rejection of a null hypothesis to be four times as serious as mistaken acceptance [Cohen 1988, p. 5]. Here, we target 95% power ($\alpha = \beta = 0.05$) and report the smallest effect size that we are 95% confident we would have observed if it existed, given post-filter sample size. That effect size, for the chi-square test, is reported in terms of an index w [Cohen 1988, chap. 7.2]. The values of w corresponding to adjectives such as "small," "medium," and "large" depend on the particular problem or field. For the psychology and behavioral science fields that he was writing about, Cohen suggested that $w = 0.1$ corresponds to a "small" effect, $w = 0.3$ to "medium," and $w = 0.5$ to "large," cautioning investigators that these should be treated as a general frame of reference and not taken too literally [Cohen 1988, pp. 224–225]. Researchers often aim to detect a medium effect size of $w = 0.3$ [Newton and Rudestam 1999, p. 76], which was intended to represent "an effect likely to be visible to the naked eye of a

Table II. Experiment 1b Results

Document pair	Cosine	KL	JS	Euclidian	Overall
"Focal" w/"Near" (matches LDA)	151 (62.1%)	145 (62.5%)	159 (71.0%)	159 (67.1%)	614 (**65.6%**)
"Focal" w/"Distant"	39 (16.0%)	43 (18.5%)	33 (14.7%)	39 (16.5%)	154 (**16.5%**)
"Distant" w/"Near"	53 (21.8%)	44 (19.0%)	32 (14.3%)	39 (16.5%)	168 (**17.9%**)
Total	243	232	224	237	936

Note: Percentages are of column total.

Table III. Spearman Correlation (ρ) Between Various "Flatness" Measures

	"Focus clear"	Gini	Standard deviation	Word count
Human: "Focus clear"	1	0.004, p = 0.854	−0.008, p = 0.674	−0.011, p = 0.570
Gini coefficient	0.004, p = 0.854	1; 1	0.921*; 0.933*	−0.460*; −0.418*
Standard deviation	−0.008, p = 0.674	0.921*; 0.933*	1; 1	−0.300*; −0.272*
Word count	−0.011, p = 0.570	−0.460*; −0.418*	−0.300*; −0.272*	1; 1

Note: *: $p < 0.0005$; $n = 2752$; $n = 10331$.

careful observer" and which has since been found to approximate the average size of observed effects in various fields [Cohen 1992, p. 156].

7.2. Experiment 1b: Results

After filtering as described in Section 5.8, we had 936 tasks completed by 458 Turkers, including "focus clear" assessments for 2752 documents. As shown in the boldfaced numbers in Table II, humans chose the same pair of documents as the algorithm **65.6%** of the time, with the remainder about evenly split between the other two options.

The hypothesis that participants' selections are independent of which method was used to compare topic vectors was *retained* with $p = 0.320$. A power analysis using G*Power 3.1.9.2 [Faul et al. 2009] showed that this test would have a 95% chance of finding even a fairly small effect of size $w \geq 0.149$. The null hypothesis was also retained when combing the middle two rows of Table II and with 95% probability would have detected an effect size with index $w \geq 0.135$. Based on these observations, we conclude that the method used for comparing topic vectors does not make a significant difference to which pair of documents participants selected as most closely related. We therefore combine all four methods for further analyses.

After people chose which pair they thought was most similar, they rated the similarity of the pair along the scale described in Section 8.1. For the document pairs that were rated, the human similarity score Spearman-correlated $\rho = 0.267$ ($p < 0.0005$) with the LDA cosine similarity of the two documents. As expected from differences in document selection, this is lower than the result reported in Experiment 2, to follow.

We compared average responses to "This focus of this idea is clearly expressed" with the "flatness" of the distribution of topics and found no correlation between what ideas people rated as less focused and "flatter" distributions. Here, "flatness" is measured by the standard deviation and/or Gini coefficient of the percentage weights in each document's topic vector. In both measures, numbers closer to 0 represent "flatter" distributions. As shown in the "word count" row of Table III, we also tested to see if longer documents (measured by total number of words assigned to any topic) were "flatter" or less focused. The absence of significant correlations in the first row of Table III should caution readers against interpreting flatter LDA topic distributions to indicate that a human would perceive the document to lack clear topical focus; a perceived focus might just not align with one of the relatively few topic dimensions that the model has computed as primary for dimensionality reduction.

8. EXPERIMENT 2

Experiment 2 uses a multi-item scale to measure perceived similarity in order to test the reliability and distinctiveness of aspects of human similarity judgments.

8.1. Experiment 2: Methods

In Experiment 2, we randomly selected one focal document, and then selected with equal probability the "near," "distant/same lead topic," or "distant/different lead topic" documents, presenting the two documents across the tops of separate pages, in random order. Participants summarized each document, as before, and {strongly, moderately, slightly} {agreed, disagreed} with each of the following statements:

—This idea is likely to save money.
—This idea would be easy to implement.
—This idea is fair.
—The relevant government decision-makers would likely be willing to do this.
—I think this a good idea.
—The questions on this page were easy.

The first four of these were intended to align with four commonly used policy analysis criteria: effectiveness, efficiency, equity, and political feasibility, respectively. These were added after it became apparent (reading the free-text explanations of similarity) that some participants in Experiment 1 were using these factors to measure similarity between documents. The fifth question is conceptually similar to the four more detailed criteria (overall "good idea") and matched Experiment 1. Experiment 1's question about understanding was omitted, since participants were not choosing a pair of documents, so that pair selection could not be influenced by differential understanding.

Participants were then shown the two documents together, in the same order, and {strongly, moderately, slightly} {agreed, disagreed} or were neutral with respect to the following statements:

—If I had just read the first idea and found it interesting, I would find the second idea interesting as well.
—It would be good for the author of the first passage to read the second passage.
—Addressing the ideas expressed in the first passage might also address the ideas in the second passage.
—It would make more sense to act on both ideas together than to act on them separately.
—These two passages address similar concerns.
—The author of the first passage is likely to agree with the second passage.
—These two ideas are about the same topics.
—The questions on this page were easy. [No neutral option.]

8.2. Experiment 2: Results

Questions were first analyzed individually. Each of the similarity questions is significantly correlated with the LDA-based similarity score for that pair (Pearson, $p < 0.05$). Using ANOVA as well as Welch and Brown-Forsythe statistics (which do not require homogeneity of variance), we found that each question has a statistically significant ($p < 0.05$) difference between ratings of pairs for which both documents had the same lead topic as compared to pairs for which documents did not share the same lead topic.

We then examined whether these questions addressed different dimensions of similarity or loaded onto a single scale. An exploratory principal-components analysis showed that the seven questions asking in various ways about document similarity loaded on a single factor, with all items intercorrelated significantly ($p < 0.0005$) and strongly (average strength by Pearson = 0.578; by Spearman = 0.579). We analyzed

those seven items as part of a single similarity construct, as suggested by Spector [1992], and found a Cronbach's α of 0.905 over the 457 surveys with answers to all these questions. This became our summated human similarity construct for further analyses (range: 7–49). For analysis of document pairs, we average the human similarity scores received for that pair.

In Experiment 2, the LDA-based distance between a pair of documents was controlled to three levels that are significantly different from each other. If this distance metric was measuring the same concepts as our scale, we would expect to see those same significant differences in the ratings. The human similarity measure was, indeed, significantly different across those three levels of LDA-based similarity. Further, the mean of the per-pair human similarity measure correlated significantly ($p < 0.0005$; $n = 75$) with the LDA-based cosine similarity score ($r = 0.544$; $\rho = 0.553$).

In order to see if some topics produced higher human similarity scores than others, we tested to see if the per-pair rating was different across different focal topics. A difference was observed: focal topics 1 and 20 led to significantly lower human similarity scores than the other three ($p < 0.0005$; 23.3 vs. 29.6 points; $n = 75$). As a point of reference, the LDA-based cosine similarity score was not significantly different across document pairs in different focal topic groups. We investigate differences among the topics further in Experiment 3, to follow. From Experiment 2, we have a "human" measure of similarity between two documents, on a 7-item scale that has desirable properties.

9. EXPERIMENT 2B

9.1. Experiment 2b: Methods

Experiment 2b was designed to measure interrater reliability for the scale in Experiment 2, because that experiment did not have enough examples for which multiple people rated the *same set of pairs*, as most participants in Experiments 1 and 2 rated only a single pair and interrater reliability requires having different people rate the same set of materials. We made Experiment 2b an independent test by allowing only participants who had not been part of Experiment 2. The procedure was the same, but used only 15 document pairs to be rated by all participants, randomly selected with balance across the five focal topics as well as the three LDA-computed similarity categories.

9.2. Experiment 2b: Results

A total of 31 participants each rated all 15 document pairs, for 465 additional results. Cronbach's alpha for the seven-item scale in this dataset is exceptionally high at 0.953 and would be lower if any item on the seven-item scale were omitted, giving us further confidence in interitem reliability. The average of (31 choose 2 = 465) Pearson correlations ($n = 15$) between raters' composite similarity scores is $r = 0.747$. Omitting the two raters whose scores sometimes did not correlate significantly with other raters, the average correlation is $r = 0.790$. All remaining correlations were significantly positive ($p < 0.05$; more than half with $p < 0.0005$). Therefore, we have confidence in the interrater reliability of this seven-item human similarity scale, even in the presence of noise from some raters.

9.3. Experiments 1 and 2: Joint Results

In Experiment 1, participants saw three documents and selected which pair was the most similar. They could choose the focal document paired with the "near" document, or the focal document paired with the "distant" document (which may have had the same highest-weighted topic or not), or they could choose the "near" and "distant" document pair. We computed the percentage of times each pair was selected, as a fraction of the

Table IV. Spearman Correlation (ρ) Between Various Similarity Measures

	Human		LDA				TF*IDF, Cosine	
	Pick 2/3	Scale	Cosine	KL	JS	Euclidian	Unstemmed	Stemmed
Human Pick 2/3	1	0.623	**0.730**	0.752	0.745	0.744	*0.504*	*0.577*
Human Scale	0.623	1	**0.553**	0.547	0.535	0.480	*0.441*	*0.504*
LDA Cosine	**0.730**	**0.553**	1	**0.980**	**0.983**	**0.919**	**0.511**	**0.639**
LDA KL	0.752	0.547	**0.980**	1	0.994	0.936	0.543	0.665
LDA JS	0.745	0.535	**0.983**	0.994	1	0.937	0.520	0.649
LDA Euclidian	0.744	0.480	**0.919**	0.936	0.937	1	0.525	0.632
TF*IDF	*0.504*	*0.441*	**0.511**	0.543	0.520	0.525	1	0.917
TF*IDF stemmed	*0.577*	*0.504*	**0.639**	0.665	0.649	0.632	0.917	1

Note: $p < 0.0005$; $n = 75$.

times it was available for selection, as a dependent variable in this analysis. This is labeled "Human Pick 2/3" in Table IV.

In Experiment 2, humans rated the similarity of the focal document paired with either the "near" document, the "distant/same lead topic" document, or the "distant/different lead topic" document, and the mean score was computed for each of the 75 document pairs. This is labeled "Human Scale" in Table IV.

Combining these two, we see how the percentage of time that a pair was picked correlates significantly ($p < 0.0005$; $n = 75$) with the human similarity metric ($r = 0.652$, $\rho = 0.623$) and the LDA similarity metric ($r = 0.772$, $\rho = 0.730$). As we mentioned before, this can perhaps be interpreted as a rough upper bound on the degree of agreement between the LDA measure and human similarity judgments. This is also consistent with the Spearman correlation between human and algorithmic judgments of similarity that have been found at the word level [Faruqui et al. 2015].

Table IV shows the correlation between these alternative measures and the human similarity scores for tested pairs of documents (as well as each other). Because the triads tested in the human similarity measures were constructed based on the LDA cosine similarity measure, those numbers are in bold.

We suspect that our basis for triad construction is why the "percentage of time that a pair was picked" measure from Experiment 1 correlates slightly stronger with the LDA measure than even with the highly reliable, independently measured human similarity scale of Experiment 2, and caution against drawing conclusions from this numerical difference. In the size of this dataset, and in other datasets we hope what has been learned from this work can apply to, it is infeasible to collect human similarity judgments for all possible pairings as an input to triad selection. Lower correlations between TF*IDF and human perceptions of similarity, shown in italics, are consistent with prior work that finds LDA to have superior performance [Dinakar et al. 2012], but the correlations between human similarity scores and TF*IDF scores may have differed if we had selected the triads based on TF*IDF.

We then measured the extent to which our triad selection based on cosine similarity (see Section 5.7.1) still tests coarse differences in similarity scores under alternative measures of comparing LDA topic vectors, and found that our selection of document sets reflecting coarse-level differences in the cosine measure also reflects comparable coarse-level differences in scores produced by these other methods for comparing vectors. The rank-ordering of similarity scores was not highly sensitive to the method used to compare vectors, as seen by the underlined correlations (average 0.958). For all four of the measures (Cosine, KL, JS, and Euclidian), the "near" documents were within the 0.62% of documents most similar to the focal document. Separating the corpus into one group of documents containing the same highest-weighted topic and a larger group of all other documents, most of the "distant" documents fall into the half of those sets that

was "least similar" according to each measure. There are only four exceptions: two of the "distant/same lead topic" documents under the KL divergence measure, and two of the "distant/different lead topic" documents under the Euclidian distance measure. These observations about coarse-level differences and rank-order correlation further support the conclusion from Experiment 1b that using another method for comparing LDA feature vectors does not make a significant difference in these results.

For the TF*IDF-based measures, to which the separation in selecting "distant" based on "same lead topic" or not does not apply, we examined the positions of our selected "near" and "distant" documents in a list ordered by similarity to the respective focal document. The "near" documents used in our experiments had an average percentile ranking (1 being very similar and 99 being very dissimilar) of 20 (unstemmed) and 14 (stemmed). The "distant/same lead topic" documents had an average percentile ranking of 33 (unstemmed) and 32 (stemmed). The "distant/different lead topic" documents had an average percentile ranking of 55 (unstemmed) and 62 (stemmed). The differences between those levels is significant ($p < 0.05$) for the stemmed measure, and the latter difference is significant for the unstemmed measure.

10. EXPERIMENT 3

Experiment 3 evaluates the cohesiveness of topics and the perceived fit between documents and topics, motivated by observed differences between documents by most-probable topic as described earlier (see Figure 1). We then investigate the extent to which the performance of the topic-modeling algorithm in this application context depends on the perceived coherence of the topics and the degree to which documents match the latent topics.

In Experiment 1, we found that which pair people chose as the most similar was not independent of the focal topic. We also found that, of the 48 responses by 47 workers in which documents in the pair selected as "most similar" did not have the same highest-weight topics, nearly half (23/48) came from three focal documents and 44% (21/48) came from documents with two particular highest-weight topics (1 and 3). Combined with some of the free-text comments in Experiment 1, we were led to explore whether or not some of the cases in which LDA did not predict the "pick two of three" task as well might be due to less coherent topics or documents that did not fit well with those topics. As algorithmic metrics do not capture topic coherence well [Chang et al. 2009], we decided to use a human scale measure, and designed a direct experiment drawing from Newman et al. [2010]. This experiment also tests the observation by Dinakar et al. [2012] that LDA-based algorithmic links seemed less helpful when they involved stories with unusual vocabulary that may not have clearly fit a single topic well.

10.1. Experiment 3: Methods

In this experiment, each of the five focal topics were shown in two different ways, assigned independently of topic: a word cloud with form similar to those in Figure 2 or a two-column list of the top 15 words and their probabilities, which included all words with probabilities >1%. Two visualizations were used in order to investigate the possible role of presentation in human judgments of coherence. Because results for both were virtually identical, we report only the aggregated results.

Participants were asked to come up with a short title for the collection, then answered five items about topic coherence, each on a six-point agree/disagree scale:

—Coming up with the title was easy.
—This collection of words is coherent.
—If I put these words into a search engine, it's pretty clear what the returned documents would be about.

—These words all belong to the same topic.
—This collection of words is meaningful.

Participants could complete the task up to 5 times, seeing a different topic each time but always the same representation, randomly assigned on their first visit.

On a second page, they were shown the same set of words in the same form, shown one of the short documents, and asked to answer another six items on the same six-point agree/disagree scale, the first five measuring how well the document fit into the topic:

—If I searched for these words, I would expect to find this document.
—If someone subscribed to the collection or feed of documents that generated this list of words, it would be appropriate to send them this document.
—This document fits into the topic described by the list of words.
—I would expect to find these words in other documents related to this one.
—This document belongs in the same collection as the documents that generated this word list.
—The questions on this page were easy.

The document shown was either one of the focal documents associated with the displayed topic or one of the "near" or "distant" documents that had been presented with that focal document in Experiment 1. We expected (1) that the "distant/different topic" documents would not be seen as fitting in well with the focal document's highest-weighted topic; (2) that selected documents might, on average, be seen as fitting less strongly with less cohesive topics; and (3) that if some of the documents did not fit particularly well with their focal topics, this might affect how people selected the "most similar" pairs in Experiment 1.

This experiment draws on work by Newman et al. [2010] on human evaluation of topic coherence for comparison to automated methods of evaluating topic coherence, but uses a much larger crowd of raters instead of trained specialists. We evaluated coherence using the five-item scale presented earlier, drawn from Newman et al.'s description of their measure.

10.2. Experiment 3: Results

In Experiment 3, we had 993 complete results from 309 unique workers after the filtering described earlier, for an average of 11 ratings per document. A reliability analysis similar to that of Experiment 2 was done on each of the two constructs in Experiment 3. The "topic-coherence" scale had a Cronbach's alpha of 0.839 after removing the "coming up with the title was easy" item, which was not as strongly correlated with the other items as suggested by our source for the scale [Newman et al. 2010]. The five-item topic-document fit scale had a Cronbach's alpha of 0.976. That would have been slightly reduced by removing any item; thus, all were retained.

To explore document-topic fit, we dichotomized the mean document-topic fit score around its neutral point of 20. All but two of the focal documents (both in topic 1), all but one of the "most similar" documents (from topic 1), and all but eight of the "distant/same lead topic" documents (half of those eight from topic 1) were seen as fitting into the highest-weighted topic. This human measure of document-topic fit correlated $r = 0.506$ ($p < 0.0005$, $n = 65$) with the fraction of a document's tokens assigned to its dominant focal topic. Of the "distant-different" documents, 24 of 25 were seen as not fitting into the focal document's highest-weighted topic. To the degree tested, the vast majority of documents fit their most-probable topics well, and did not fit other topics well, as expected. One topic accounted for most of the exceptions, with significantly lower document-topic fit ($p < 0.0005$, $n = 672$, contrast value of -5.23, on

a scale from 5 to 35), and it was one of the topics with significantly lower agreement with LDA and lower human similarity scores, above. Manual examination suggests that this topic covered a wider range of concepts (within the Department of Health and Human Services) than the other focal topics; it also scored significantly ($p < 0.0005$) lower than other topics on the coherence scale.

In general, we noticed a positive association between whether or not a participant's selection in Experiment 1 matched LDA's, and how well the focal document was perceived to fit its highest-weighted topic. The mean document-topic fit score was 27.2 in cases in which people agreed with LDA and 24.6 when they did not, a statistically significant difference ($p < 0.0005$; $n = 562$). The dichotomized document-topic fit for the focal document is not independent of whether a person selected the same pair as LDA ($p < 0.0005$).

Limiting analysis to cases in which the "distant" document had the "same lead topic" as the "focal" and "near" documents, participants were more likely to disagree with LDA and more likely to choose the focal and "distant" pair when the "distant" document fit well into its topic. In a majority of the cases in which the participant chose the "focal" and "distant/same lead topic," the "distant/same lead topic" document was seen as a better fit with that topic ($p < 0.0005$).

In 416 of the 562 cases (74%), Experiment 1 participants selected the two documents most closely associated with the focal topic. In 28 of the 140 cases in which participants disagreed with LDA, their selection was consistent with choosing the two that best fit the focal topic (according to ratings from Experiment 3).

11. LIMITATIONS

We have done this work on only one dataset, as well as a subset of that corpus constructed to compare pairings that have very high levels of LDA-based similarity scores with pairings that have lower algorithmic similarity scores, based on the motivating application. Except when testing generality in Experiment 1b, we deliberately chose conditions favorable for observing agreement between humans and LDA, such as selecting documents with a single very dominant topic and selecting a number of topics that maximally distinguished between documents. We did this in order to see how well LDA could perform under favorable circumstances—roughly, an upper bound—and to generate exceptions for further investigation about what LDA seems unable to handle.

12. CONCLUSIONS

In this article, we have presented a method for evaluating a measure of document similarity, including a "choose two most similar of three" task setup as well as an independent human similarity rating on a novel and reliable seven-item scale that captures a single conceptual dimension of similarity. The methods that we have presented in this article for preparing data and testing it against an algorithm can be adapted for use with other textual datasets. The methods can also be used to evaluate more advanced topic models or other similarity measures.

We have applied these methods to evaluate an LDA-based cosine similarity measure. We find that, for the most part, under deliberately chosen conditions favorable for observing agreement, human similarity judgments and the "simple" LDA-based cosine similarity measurement often ($\approx$75%) agree, and that agreement is closer to 66% when documents are selected more generally. There is still room for improvement of the algorithm in being able to automatically generate links that match selections that humans would make.

Although we built our experiments using one specific method for comparing topic vectors (cosine similarity), the rank orderings produced by other comparison methods are quite similar. More important, the limitations of the LDA model that we found

and explored here are limitations of the LDA model regardless of the approach used to compare feature vectors and regardless of certain improvements that might be made on the basic LDA model. Even a highly refined factorization or data-reduction model attends to some features (e.g., topics) and obscures others; it is important for practitioners who hope to help human readers with an algorithmic similarity measure to understand those limits better than what prior work permits.

Exploring human perceptions of similarity more deeply, we qualitatively and quantitatively identified aspects of documents, such as understandability and idea quality, that LDA does not pick up on, nor could it be expected to, from its "bag of words" model. We find that, at least in our setup, topic coherence and perceived fit between documents and highest-weighted topics appear to be important factors in predicting what humans will choose on the link analogy task; this fit may be generally important when using dimensionality reduction techniques. These are caveats for the use of LDA for link generation based on topical similarity.

We think that our work suggests the value of working on techniques that can correct for the observed shortcomings, for example, by incorporating features beyond word or topic vectors (such as tags, authors, and connections between authors) in measuring similarity, using sources of prior knowledge to constrain the topic model (e.g., Andrzejewski et al. [2011], Andrzejewski and Zhu [2009], and Zhai et al. [2011]) or using a human-in-the-loop mixed initiative approach (e.g., Huang and Mitchell [2007]), so that the dynamic structuring of content benefits from both automation and human knowledge.

In future work, we would also like to see how human measurements of topic coherence match algorithmic measures of topic quality, such as the "flatness" of the term distribution that makes up the topic, expanding on Chang et al. [2009]. If an algorithm can automatically understand when its similarity measure will perform well versus poorly, it can more appropriately add/not add links and structure, differentially weight ensemble methods, and better incorporate human feedback, leading to more effective use of these large-scale collaboration support tools.

ACKNOWLEDGMENTS

Thanks also to staff programmer Eric Rosé, who provided logistics and technology support. Thanks to anonymous reviewers for their suggestions. Thanks also to the many people who submitted ideas for, or facilitated the original collection of, the SAVE award ideas used in these experiments, and the many Mechanical Turk workers who assessed and compared them.

REFERENCES

David Andrzejewski and Xiaojin Zhu. 2009. Latent Dirichlet allocation with topic-in-set knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing (SemiSupLearn'09)*. Stroudsburg, PA: Association for Computational Linguistics, 43–48.

David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. 2011. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence - Volume Two (IJCAI'11)*. Barcelona, Catalonia, Spain: AAAI Press, 1171–1177. DOI:http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-200

Brian P. Bailey and Eric Horvitz. 2010. What's your idea?: A case study of a grassroots innovation pipeline within a large software company. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. New York, NY: ACM, 2065–2074. DOI:http://dx.doi.org/10.1145/1753326.1753641

Chris Bank and Jerry Cao. 2015. *The Guide to Usability Testing*, Mountain View, CA: UXPin.

Osvald M. Bjelland and Robert Chapman Wood. 2008. An Inside View of IBM's "Innovation Jam." Retrieved June 29, 2016 from http://sloanreview.mit.edu/article/an-inside-view-of-ibms-innovation-jam/.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022.

Kevin W. Boyack et al. 2011. Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoSONE* 6, 3, e18029. DOI:http://dx.doi.org/10.1371/journal.pone.0018029

K. Selçuk Candan, Luigi Di Caro, and Maria Luisa Sapino. 2012. PhC: Multiresolution visualization and exploration of text corpora with parallel hierarchical coordinates. *ACM Transactions on Intelligent Systems Technology* 3, 2, 22:1–22:36. DOI:http://dx.doi.org/10.1145/2089094.2089098

Jonathan Chang and David Blei. 2009. Relational topic models for document networks. In *AISTATS*. Clearwater Beach, FL, 81–88.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of NIPS*. Vancouver, BC, Canada.

Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. 2012. Interpretation and trust: designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*. New York, NY: ACM, 443–452. DOI:http://dx.doi.org/10.1145/2207676.2207738

Jacob Cohen. 1992. A power primer. *Psychological Bulletin* 112, 1, 155–159. DOI:http://dx.doi.org/10.1037/0033-2909.112.1.155

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, New York, NY: L. Erlbaum Associates.

Gregorio Convertino. 2013. Large-Scale Idea Management and Deliberation Systems Workshop. Retrieved June 29, 2016 from http://comtech13.xrce.xerox.com/comtech13.html.

Weiwei Cui, Huamin Qu, Hong Zhou, Wenbin Zhang, and Steve Skiena. 2012. Watch the story unfold with textwheel: Visualization of large-scale news streams. *ACM Transactions on Intelligent Systems Technology* 3, 2, 20, 1–20:17. DOI:http://dx.doi.org/10.1145/2089094.2089096

Todd Davies. 2011. Online Deliberation Conferences. (March 2011). Retrieved May 13, 2011 from http://online-deliberation.net/.

Karthik Dinakar et al. 2012. You too?! mixed-initiative lda story matching to help teens in distress. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*. Dublin, Ireland: AAAI, 74–81.

Ran El-Yaniv, Shai Fine, and Naftali Tishby. 1997. Agnostic classification of Markovian sequences. In *Advances in Neural Information Processing Systems 10*, M. I. Jordan, M. J. Kearns, and S. A. Solla (Eds.). Cambridge, MA: MIT Press, 465–471.

D. M. Endres and J. E. Schindelin. 2003. A new metric for probability distributions. *IEEE Transactions on Information Theory* 49, 7, 1858–1860. DOI:http://dx.doi.org/10.1109/TIT.2003.813506

Katayoun Farrahi and Daniel Gatica-Perez. 2011. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems Technology* 2, 1, 3, 1–3, 27. DOI:http://dx.doi.org/10.1145/1889681.1889684

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*.

Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41, 4, 1149–1160. DOI:http://dx.doi.org/10.3758/BRM.41.4.1149

M. Flynn, L. Dooley, D. O'Sullivan, and K. Cormican. 2003. Idea management for organisational innovation. *International Journal of Innovation Management* 07, 04, 417–442. DOI:http://dx.doi.org/10.1142/S1363919603000878

Sean Goggins, Christopher Mascaro, and Stephanie Mascaro. 2012. Relief work after the 2010 Haiti earthquake: Leadership in an online resource coordination network. In *CSCW'12*. New York, NY: ACM, 57–66. DOI:http://dx.doi.org/10.1145/2145204.2145218

Brynjar Gretarsson et al. 2012. TopicNets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems Technology* 3, 2, 23, 1–23, 26. DOI:http://dx.doi.org/10.1145/2089094.2089099

Taher H. Haveliwala, Aristides Gionis, Dan Klein, and Piotr Indyk. 2002. Evaluating strategies for similarity search on the web. In *Proceedings of the 11th International Conference on World Wide Web (WWW'02)*. New York, NY: ACM, 432–442. DOI:http://dx.doi.org/10.1145/511446.511502

Yifen Huang and Tom Mitchell. 2007. A framework for mixed-initiative clustering. In *North East Student Colloquium on Artificial Intelligence (NESCAI'07)*. Ithaca, NY.

Kevin O. Hwang et al. 2010. Social support in an Internet weight loss community. *International Journal of Medical Informatics* 79, 1, 5–13. DOI:http://dx.doi.org/10.1016/j.ijmedinf.2009.10.003

IBM. 2012. A Global Innovation Jam. Retrieved Jun 29, 2016 from http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/innovationjam/.

IdeaScale. 2012. Save Award 2012. Retrieved June 29, 2016 from http://saveaward2012.ideascale.com/.

IdeaScale. 2013. The Truth About IdeaScale. Retrieved June 29, 2016 from http://dev.ideascale.com/infocomics/.

Joshua E. Introne and Marcus Drescher. 2013. Analyzing the flow of knowledge in computer mediated teams. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW'13)*. New York, NY: ACM, 341–356. DOI:http://dx.doi.org/10.1145/2441776.2441816

Steven Johnson. 2010. *Where Good Ideas Come From*. New York, NY: Penguin.

Kenneth Joseph, Kathleen M. Carley, and Jason I. Hong. 2014. Check-ins in "blau space": applying Blau's macrosociological theory to foursquare check-ins from New York City. *ACM Transactions on Intelligent Systems Technology* 5, 3, 46, 1–46, 22. DOI:http://dx.doi.org/10.1145/2566617

Aniket Kittur, E. D. H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems (CHI'08)*. New York, NY ACM, 453–456. DOI:http://dx.doi.org/10.1145/1357054.1357127

Michael David Lee, B. M. Pincombe, and Matthew Brian Welsh. 2005. An empirical evaluation of models of text document similarity. In *Stresa*, Italy: Cognitive Science Society.

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37, 1, 145–151. DOI:http://dx.doi.org/10.1109/18.61115

Shixia Liu et al. 2012. TIARA: Interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems Technology* 3, 2, 25, 1–25, 28. DOI:http://dx.doi.org/10.1145/2089094.2089101

Zhiyuan Liu, Yuzhou Zhang, Edward Y. Chang, and Maosong Sun. 2011. PLDA+: Parallel latent Dirichlet allocation with data placement and pipeline processing. *ACM Transactions on Intelligent Systems Technology* 2, 3, 26, 1–26, 18. DOI:http://dx.doi.org/10.1145/1961189.1961198

Inderjeet Mani. 2001. Evaluation. In *Automatic Summarization.* Natural Language Processing. Philadelphia: John Benjamins Publishing, 221–259.

Michael Muller and Sacha Chua. 2012. Brainstorming for Japan: rapid distributed global collaboration for disaster response. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12).* New York, NY: ACM, 2727–2730. DOI:http://dx.doi.org/10.1145/2207676.2208668

Un Yong Nahm. 2004. *Text Mining with Information Extraction*. Austin, TX: University of Texas at Austin.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT'10).* Stroudsburg, PA: Association for Computational Linguistics, 100–108.

Rae R. Newton and Kjell Erik Rudestam. 1999. *Your Statistical Consultant: Answers to Your Data Analysis Questions*, Thousand Oaks, CA: Sage.

Leysia Palen et al. 2010. A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. In *Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference (ACM-BCS'10)*. Swinton, UK: British Computer Society, 8, 1–8, 12.

Brandon Pincombe. 2004. *Comparison of Human and Latent Semantic Analysis (LSA) Judgements of Pairwise Document Similarities for a News Corpus*, Edinburgh, South Australia: Australian Government Department of Defence: Defence Science and Technology Organisation.

Gregory J. Privitera. 2015. *Student Study Guide with IBM® SPSS® Workbook for Essential Statistics for the Behavioral Sciences*. Thousand Oaks, CA: SAGE Publications.

M. J. Salganik, P. S. Dodds, and D. J. Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311, 5762, 854.

Amit Singh, Deepak P, and Dinesh Raghu. 2012. Retrieving similar discussion forum threads: A structure based approach. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12).* New York, NY: ACM, 135–144. DOI:http://dx.doi.org/10.1145/2348283.2348305

Sergej Sizov. 2012. Latent geospatial semantics of social media. *ACM Transactions on Intelligent Systems Technology* 3, 4, 64, 1–64, 20. DOI:http://dx.doi.org/10.1145/2337542.2337549

Paul E. Spector. 1992. *Summated Rating Scale Construction: An Introduction*. Thousand Oaks, CA: SAGE.

Ellen Spertus, Mehran Sahami, and Orkut Buyukkokten. 2005. Evaluating similarity measures: A large-scale study in the Orkut social network. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD'05)*. New York, NY: ACM, 678–684. DOI:http://dx.doi.org/10.1145/1081870.1081956

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. In *Handbook of Latent Semantic Analysis*, Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch (Eds.). University of Colorado Institute of Cognitive Science Series. Mahwah, NJ: Lawrence Erlbaum Associates, 427–448.

Cass R. Sunstein. 2006. *Infotopia: How Many Minds Produce Knowledge*. New York; Oxford: Oxford University Press.

W. Ben Towne and James D. Herbsleb. 2012. Design considerations for online deliberation systems. *Journal of Information Technology & Politics* 9, 1, 97–115. DOI:http://dx.doi.org/10.1080/19331681.2011.637711

Jason Tsay, Laura Dabbish, and James Herbsleb. 2014. Let's talk about it: evaluating contributions through discussion in github. In *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE'14)*. New York, NY: ACM, 144–154. DOI:http://dx.doi.org/10.1145/2635868.2635882

Fernanda B. Viégas, Martin Wattenberg, Jesse Kriss, and Frank van Ham. 2007. Talk before you type: Coordination in Wikipedia. In *Proceedings of HICSS'07*. 78. DOI:http://dx.doi.org/10.1109/HICSS.2007.511

Ellen M. Voorhees. 2007. TREC: Continuing information retrieval's tradition of experimentation. *Communications of the ACM* 50, 11, 51–54. DOI:http://dx.doi.org/10.1145/1297797.1297822

Thomas P. Walter and Andrea Back. 2013. A text mining approach to evaluate submissions to crowdsourcing contests. In *46th Hawaii International Conference on System Sciences (HICSS'13)*. 3109–3118. DOI:http://dx.doi.org/10.1109/HICSS.2013.64

Gu Xu and Wei-Ying Ma. 2006. Building implicit links from content for forum search. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'06*). New York, NY: ACM, 300–307. DOI:http://dx.doi.org/10.1145/1148170.1148224

Zhijun Yin, Liangliang Cao, Quanquan Gu, and Jiawei Han. 2012. Latent community topic analysis: Integration of community discovery with topic modeling. *ACM Transactions on Intelligent Systems Technology* 3, 4, 63, 1–63, 21. DOI:http://dx.doi.org/10.1145/2337542.2337548

Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. 2011. Constrained LDA for grouping product features in opinion mining. In *Advances in Knowledge Discovery and Data Mining*, Joshua Zhexue Huang, Longbing Cao, and Jaideep Srivastava (Eds.). Lecture Notes in Computer Science. Springer, Berlin, 448–459.

Haizheng Zhang, Baojun Qiu, C. L. Giles, Henry C. Foley, and J. Yen. 2007. An LDA-based community structure discovery approach for large-scale social networks. In *IEEE Intelligence and Security Informatics*. 200–207. DOI:http://dx.doi.org/10.1109/ISI.2007.379553

Shiwan Zhao, Michelle X. Zhou, Xiatian Zhang, Quan Yuan, Wentao Zheng, and Rongyao Fu. 2011. Who is doing what and when: social map-based recommendation for content-centric social web sites. *ACM Transactions on Intelligent Systems Technology* 3, 1, 5, 1–5, 23. DOI:http://dx.doi.org/10.1145/2036264.2036269

Received December 2015; accepted February 2016