Human Assessments of Document Similarity

[1]Westerman, S.J., [2]Cribbin, T., & [3]Collins, J.

[1]Institute of Psychological Sciences, University of Leeds, UK

[2]Department of Information Systems and Computing, Brunel University, UK

[3]Department of Natural and Social Sciences, University of Gloucestershire, UK

Corresponding author:

Dr Steve Westerman,

Institute of Psychological Sciences

University of Leeds

Leeds LS2 9JT UK

Email: s.j.westerman@leeds.ac.uk

Notes: Inter-assessor reliabilities and data for 5-grams from Experiment 2 were reported at the INTERACT '99 conference. This research was partly supported by a grant from the British Engineering and Physical Sciences Research Council. The data for these studies were gathered at Aston University, Birmingham, UK.

Abstract

Two studies are reported that examined the reliability of human assessments of document similarity and the association between human ratings and the results of n-gram automatic text analysis (ATA). Human inter-assessor reliability (IAR) was moderate to poor. However, correlations between average human ratings and n-gram solutions were strong. The average correlation between ATA and individual human solutions was greater than IAR. N-gram length influenced the strength of association, but optimum string length depended on the nature of the text (technical versus non-technical). We conclude that the methodology applied in previous studies may have led to over-optimistic views on human reliability, but that an optimal n-gram solution can provide a good approximation of the average human assessment of document similarity, a result that has important implications for future development of document visualization systems.

Introduction

Automatic text analysis (ATA) methods are fundamental to many search and analytics applications. To fulfil their intended function it is critical that ATA-generated document similarity metrics provide a good approximation of human assessments of semantic relatedness. In this paper we examine the performance of the vector space model, a common approach to modelling document similarity where relatedness is described in terms of shared features. Human assessments of document similarity have been cited as the 'gold standard' against which ATA models of document similarity should be judged (Lee, Pincombe, & Welsh, 2005). However, previous work has not investigated fully the influence of individual differences in human assessments. This is mainly due to difficulties in obtaining full sets of ratings, given that the number of unique similarities increases quadratically with the number of documents. In two reported experiments we limit document set size (n=8), making it possible to obtain full sets of ratings from a substantial number of assessors on multiple document sets.

Inter-assessor reliability is a controversial issue when it comes to the evaluation of automatically generated document similarity models. Some (e.g., Harman & Vorhees, 2006) point to strong consistency while others (e.g., Saracevic, 2008; Morris, 2010) point to substantial individual differences. Research evidence on which conclusions can be drawn is rather limited (Saracevic, 2008). In this paper we report two experiments that contribute to this issue. These used relatively demanding (but reasonable) testing conditions for human judgments of semantic similarity. Comparisons were made with ATA solutions for the document sets—using the Vector Space Model with different length n-grams as terms. In this introductory section we first consider the role of human assessment in the evaluation of information retrieval algorithms/systems; we then describe how ATA can be achieved using the Vector Space Model with n-grams as terms; and finally we present the experimental rationale and aims in more detail.

The reliability of human assessments

A distinction has been made between: i) system-oriented; and, ii) user-oriented, or cognitive information retrieval (IR) research perspectives (see e.g., Järvelin, 2007; although see Hjørland, 2010, for an alternative view). Each perspective has different priorities and somewhat conflicting demands. The former focuses on the form and effectiveness of the computing algorithms that support information retrieval—whereas the primary concerns of the latter are the ways in which information retrieval outcomes are influenced by the characteristics of the human users of information retrieval systems and the context of use. From a user-oriented perspective rather poor levels of inter-assessor agreement have been suggested. Saracevic (2008, p. 773) argues that "people differ, sometimes considerably, in decisions related to a variety of information processes, such as indexing, classification, searching, and yes, relevance as well". From a systems-oriented perspective a more optimistic picture is put forward. Concern, here, is typically with judgments of item relevance with regard to specific search criteria. For example, Harman & Voorhees (2006) report strong inter-assessor agreement (using two additional assessors) for TREC-4 and conclude that the use of different assessors would not have produced differences in rankings of the ATA systems under evaluation. Given that Saracevic (2008) agrees that inter-assessor variability has rather limited effects on system assessments it might be argued that the consequences of inter-assessor variability are unimportant. However, this is not a sound conclusion for three reasons. First, the use of relevance judgments as a basis for assessment makes extrapolation difficult. Semantic similarity is a key feature of relevance judgments—obviously they are based on assessments of similarity between retrieved items and specified search criteria but, more importantly in this context, the implication is that retrieved items will be semantically similar to one another. However, there are problems with applying the construct of relevance to scenarios involving context-free or exploratory information seeking (a particular focus of the user-oriented approach). Second, sampling is an issue insofar as there is tight control over the selection of topic descriptions and the selection of assessors. Third, this approach evaluates information systems but takes only limited account of the efficiency of systems in meeting the needs of individual users.

An alternative approach to the evaluation of inter-assessor reliability, that avoids the difficulties associated with relevance judgments, is to obtain judgments of inter-item similarity. Strong inter-assessor reliabilities have been reported using this method for relatively simple semantic stimuli. For example, Resnik (1999) examined participants' assessments of the semantic similarity of word pairs. Correlations of individual participant's ratings with the average ratings obtained in an earlier study (Miller & Charles, 1991) were generally strong, averaging r=0.88. Of course we might expect agreement to decline as semantic complexity of the stimuli increases. Lee, Pincombe, & Welsh (2005) report inter-rater reliability of 0.61, based on inter-document similarity ratings obtained from a sample of 83 participants for a set of 50 documents (news stories of 51-126 words). A limitation with this study (understandable given the number of possible pairwise ratings) was that participants only rated subsets of pairs, such that each pair received between 8 and 12 ratings, so inter-rater reliability was calculated, somewhat unusually, by selecting one rating for each document pair at random, and examining correlations with the mean across the set of document pairs. Belz & Reiter (2006) used a similar method (correlating an individual score with an average) to investigate reliability of assessments of text quality (for a set of brief weather forecasts) and obtained a similar outcome for non-expert assessors (r=0.61). Our results (see below) indicate that there are problems with making inferences based on this analytic approach. The correlation between a score and the average of a series of other scores can be very different to the average of the individual correlations, and the former can lead to over-estimation of reliability. To overcome this problem, in the reported studies we limit document set size to make it possible to obtain full sets of ratings from each assessor.

Automatic text analysis: The Vector Space Model with n-grams as terms

Following Lee et al's (2005) protocol, we compare human models of inter-document similarity to similarity models automatically generated from n-gram term vector space models. In the vector space model each document is represented as a vector of dimensionality *t*, where *t* is equal to the number of unique terms occurring within the

corpus. The cell values in the resulting term-document matrix reflect the relative importance or weight of each term within each document. For a given document, i ,and term, j, the weight, w_{ij} can be calculated as a function of local and often global term frequency. In the experiments reported here, following Lee et al. (2005), we used a simple term frequency (*tf*) count measure, with no adjustment made for global frequency. For each term-document matrix, a symmetric all-pairs document similarity matrix was computed using the normalised dot product or cosine measure. Although other similarity measures exist (e.g. Jaccard, Correlation), cosine is probably the most widely accepted similarity measure in information retrieval. Moreover, Lee et al.'s (2005) results demonstrated the performance of cosine to be marginally superior to the other measures tested.

N-grams are consecutive letter strings, *n* characters in length (see e.g. Cavnar, 1995; Damashek, 1995a). N-gram analysis involves moving a 'window', *n* characters wide, through each document, one character at a time. Each unique term (n-gram) is entered into a hash table where its frequency of occurrence in each document is recorded. So, for example, using a 5-gram to analyze the sentence "The cat sat on the mat", would produce the terms: 'The c', 'he ca', 'e cat', ' cat ', 'cat s', and so on. There are a number of advantages associated with the use of n-grams, as opposed to words, as terms. First, this process requires no implicit knowledge of the material under analysis and is language independent. Second, it reduces difficulties caused by words of similar meaning having different prefixes or suffixes (e.g. 'computer' and 'computers'). Similarly, effects of differences in spelling (e.g. English vs. American) or misspelling are reduced.

The n-gram approach has been paired with the Vector Space Model as a means of automatic text analysis. For up to 5-gram length it has been shown that it conforms to a Zipfian distribution (Cavnar & Trenkle, 1994). Performance—relative to other information retrieval methods— has been assessed using the 'ad hoc' and 'routing' tasks developed for the Text REtrieval Conferences (TREC: see Harman, 1995). The former requires "document retrieval based on brief stylized descriptions of the desired document", and the latter requires "document retrieval based on the full text of exemplars

certified to be of interest" (4, pp. 845-846). Cavnar (1995) reported the results of an n-gram system that performed at approximately the average level when compared with other information retrieval systems participating in TREC-3. With reference to the same criterion, Damashek (1995b) reported relatively poor performance of a similar n-gram system on the first of these tasks, but suggested that performance of the latter task "… compared favorably to state-of-the-art retrieval systems" (p. 847). This appraisal was subsequently disputed by Harman et al. (1995), and Salton (1995), who felt it was over-optimistic. Nevertheless, these results indicate that n-gram systems may be relatively more effective when the task involves comparison of 'full' texts rather than brief descriptors.

Of course, shorter n-grams have the advantage of relatively reduced computational demands as they create a smaller number of unique terms. For example, in their study of document search, Fox, Frieder, Knepper, and Snowberg (1999) used a 3-gram, as part of an initial filtering mechanism, for this reason. However, it can be argued that longer n-grams will be better at determining context. Using longer n-grams will tend to emphasize the frequencies with which words co-occur and this may provide richer semantic information (cf. Landauer & Dumais, 1997; Lund & Burgess, 1996). Cavnar (1995) restricted his analysis to a 4-gram. Damashek (1995a, p. 843) suggests that n-gram length is 'arbitrary' and that "consistent results are … obtained for a range of n-gram lengths". He mentions using "5-grams for English language examples and 6-grams for … Japanese". However, Damashek (1995b) states that a 7-gram performs better than a 6-gram, that performs better than a 5-gram, when reconstructing English documents from the term list by "… concatenating those n-grams that can be uniquely paired with a predecessor or successor" (p. 1419). This is broadly consistent with the results of Lee et al. (2005) who found that optimal performance occurred with 6-gram, with no further benefit for 7- to 9-gram, and a weak suggestion of a tendency to tail off from 10-gram onwards. They concluded that when compared directly with word terms, n-grams result in superior similarity models, although optimal performance depends on selecting correct n (Lee et al., 2005). Whilst Lee et al. limited their ATA to 10-grams, in the experiments reported here we extend the analysis to 25-grams. This seemed worthwhile given the

relatively 'technical' (and therefore more potentially complex) content of the document sets used in Experiment 2.

Experimental aims

The experiments reported here examined human variability in assessments of semantic similarity and the implications for ATA. Existing evidence suggests that inter-assessor agreement on context-free assessments of document similarity rating is relatively good (Lee et al., 2005). Here were test this further giving particular consideration to the importance of examining the average inter-assessor correlation based on the full set of document comparisons (rather than using partial sets and correlating individual scores with averages). This was achieved through the expedient of using small document sets (n=8). The general methodology applied in these experiments was thought to be challenging for inter-assessor agreement. As was the case in Lee et al. (2005), participants were asked to make inter-document assessments of similarity in a context-free manner (i.e., with no information retrieval task goal). This is a useful approach on the basis that: i) for many information retrieval tasks information needs are only poorly specified (Belkin, 1982); and, ii) system users will inevitably bring their own idiosyncratic mental model to any information set (Johnson-Laird, 1983; Haenggi, Gernsbacher, & Bolliger, 1994)—and will tend to describe the same concepts using different language (see Furnas, Landauer, Gomez, & Dumais, 1987; Saracevic, 2008). However, inter-assessor variability is likely to be greater in this context-free circumstance. Variability is also likely to be greater when the textual material under consideration is relatively more complex—in this instance comparing document pairs rather than term/concept similarity. This bears on the fundamental issue of the extent to which algorithmic representation of information can provide a model to facilitate information retrieval for all users. Given that individual differences may be influenced by the nature of the material under consideration (Morris, 2010) the experiments also examined the effect of document complexity - in Experiment 1 non-technical documents were used as materials—in Experiment 2 technical documents were used as materials.

A further question of interest was the extent to which semantic similarity influences inter-assessor agreement. This was addressed by examining the association between: i) mean ratings for document pairs; and, ii) inter-assessor variability for those pairs. Three possible outcomes were considered. First, it may be that, consistent with Harman & Voorhees (2006), agreement is stronger for document pairs that are relatively dissimilar—producing a positive correlation. Second, it may be that agreement is stronger at either end of the similarity rating scale—i.e., some document pairs are identified by all as being highly (dis)similar—producing an inverted-u relationship. Third, we consider the possibility that there is no systematic association between semantic ratings and inter-assessor variability.

Finally, following Lee et al. (2005) we examined the association between ATA solutions and human assessments. Human-ATA correlations provide the basis for consideration of: i) differences in the magnitude of agreement based on individual versus group level (averaged) data; ii) the extent to which a common solution can be identified and context-free information retrieval can be supported; and related to this, iii) the extent to which ATA solutions support idiosyncratic views of an information space. Even if ATA provides a good approximation of the average human judgment, the greater the inter-individual variability the important it is to consider how well a system serves individual users. As part of this process we examined whether this changes with the type of document (technical versus non technical). We tested the quality of solutions using different length n-grams. This allowed us to verify Lee et al.'s (2005) findings, with respect to optimal-n, using a range of different document sets.

Experiment 1: Non-technical documents

Two document sets were selected from the TREC 6 database (see Vorhees & Harman, 1998), each comprising eight newspaper articles from the LA Times. This material was regarded as 'non-technical' (NT) and contrasts with the 'technical' (T) material used for Experiment 2. One document set (NT1) related to 'risks taken by

journalists' (as assessed by TREC). The second document set (NT2) related to 'acts of piracy' (commercial rather than maritime).

Table 1. Basic document statistics for Non-Technical (NT) document sets

	NT1		NT2	
	Mean	SD	Mean	SD
Average word length	5.14	0.26	5.08	0.37
Average number of words per document	290.88	112.43	297.63	112.76
Average reading ease	51.84	8.31	56.84	10.04

Basic statistics for the non-technical document sets—word lengths, number of words per document, and reading ease (Flesch, 1949)—are presented in Table 1. None of these differences between document sets were significant, $t(14)=0.39$, $t(14)=-.12$, and $t(14)=-.16$, respectively.

Four men and 20 women (mean age = 19.96 years, sd=2.65) were recruited from the student population of Aston University. They were allocated randomly in equal numbers to one of the two document set conditions (as described above). Using purpose-written software, they were presented with all possible pairings of documents (28 pairs) in a random sequence, and required to indicate the degree of perceived similarity between document pairs using a visual analogue scale. This task was presented in a context-free manner (i.e. participants were not instructed to rate pairs according to any particular criteria). Each document set also was analyzed using n-grams of varying length (3-25 characters) to produce a further 23 document similarity matrices. Punctuation was retained for this analysis. All matrices were converted to vectors, comprising 28 unique cells in each.

To examine the importance of measuring reliability using full sets of ratings, inter-assessor agreement was measured in two ways. First by taking the average of all assessor-pair correlations; and second, in a similar manner to Lee et al. (2005) by calculating correlations between each assessor and the mean of all other assessors and then computing the average of these.

Results

Reliability coefficients are presented with the full pair-wise average first, followed (in parentheses) by the average of the individual correlations with the mean of the remaining assessors.

Inter-assessor reliability for the 'journalists' risks' document set was +0.52 (+0.70), and for the 'piracy' document set was +0.38 (+0.58). Correlations were calculated between the average human document similarity vectors and the document similarity vectors produced by each length of n-gram (MEANCORR). The average correlation between n-gram solutions and individual human assessors' vectors was also calculated (INDCORR).

An 8-gram produced the strongest correlation between ATA solution and MEANCORR for the 'journalists' risks' document set, r(26)=+0.76 (see Figure 1). INDCORR peaked at +0.57 (n=12, sd=0.10), and this also occurred with a 8-gram.

The peak correlation between n-gram solution and MEANCORR, for the 'piracy' document set, was r(26)=+0.68, and occurred with a 5-gram (see Figure 2). The peak for INDCORR was +0.43 (n=12, sd=0.23), and also occurred with a 5-gram. These results are discussed following a description of Experiment 2.

The correlations between the mean rating for each document pair (n=28) and the relevant standard deviation were r(26)=0.25, $p>0.05$, for the Journalists' risks document set and r(26)=0.64, $p<.001$ for the piracy document set.

FIG. 1. Correlations between ATA solution and average human rating, and average correlation between ATA solution and individual assessors, for n-grams between 3 and 25 characters in length, for 'journalists' risks' document set.

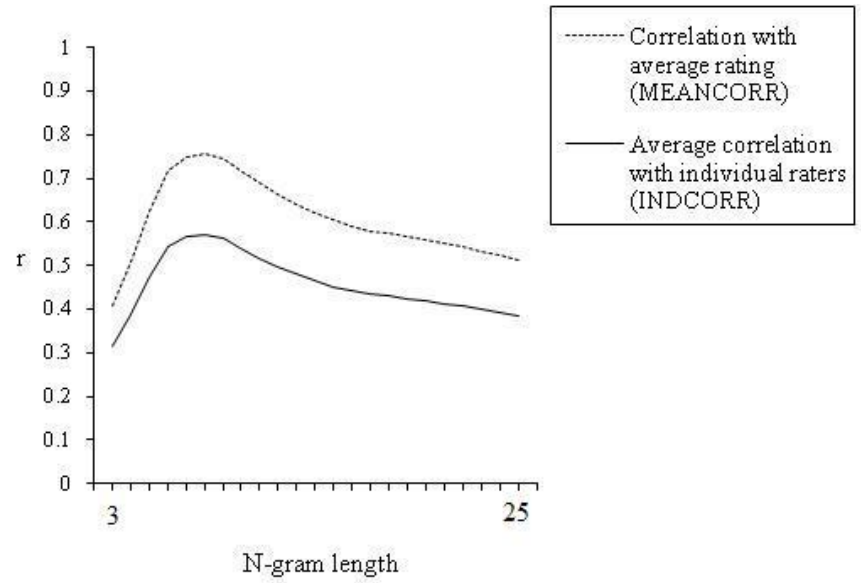
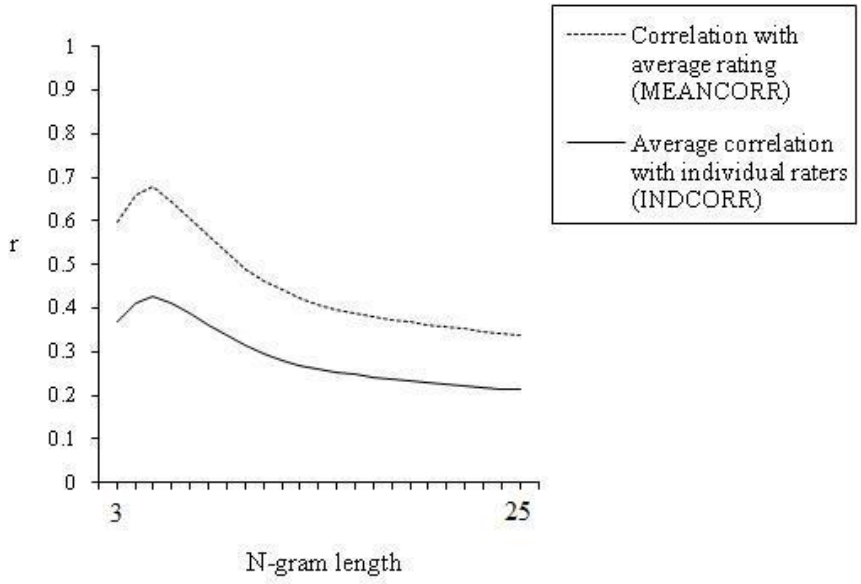


FIG. 2. Correlations between ATA solution and average human rating, and average correlation between ATA solution and individual assessors, for n-grams between 3 and 25 characters in length, for 'piracy' document set.



Experiment 2: Technical documents

Two documents sets were prepared, each comprising eight psychology journal paper abstracts, drawn from the social-science section of the Bath Information and Data Services Social Sciences on-line database using term-specific search queries. The first set (T1) comprised abstracts retrieved using the query terms "Working" and "Memory". The second set (T2) resulted from the query term "Schizophrenia". Documents for each set were selected from the chronologically most recent 32 retrieved items. Basic statistics for the technical document sets—word lengths, number of words per document, and reading ease (Flesch, 1949)—are presented in Table 2. None of these differences between document sets were significant, $t(14)=1.42$, $t(14)=.48$, and $t(14)=1.44$, respectively.

Table 2. Basic document statistics for Technical (T) document sets

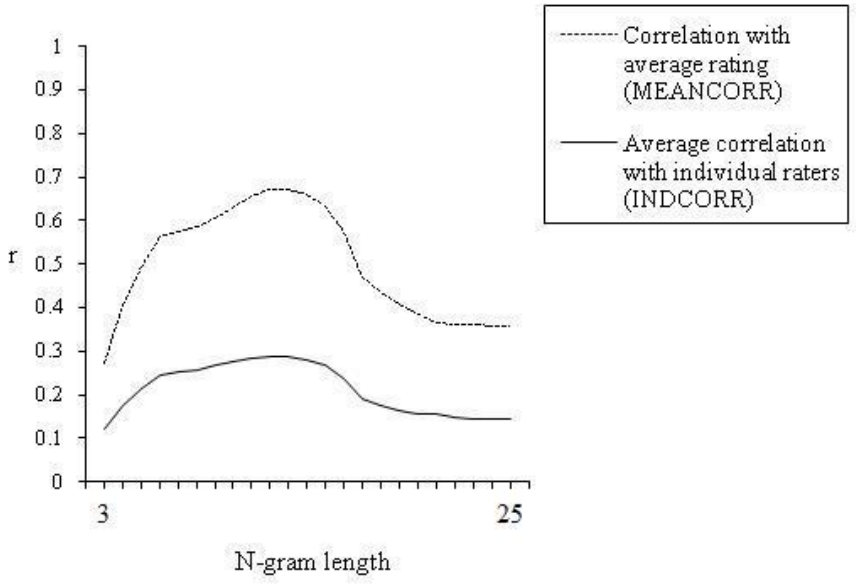
	T1		T2	
	Mean	SD	Mean	SD
Average word length	7.04	0.30	7.29	0.40
Average number of words per document	181.75	27.23	193.00	60.17
Average reading ease	18.95	11.82	11.03	10.19

Thirteen men and 23 women were recruited from the Psychology staff and Psychology students (2nd Year undergraduate) of Aston University (mean age=23.67 years, sd=8.34) and allocated in equal numbers to one of the document sets. Data gathering and analyses proceeded as for Experiment 1.

Results

Inter-rater reliability for the 'working memory' document set was +0.14 (+0.33), and for the 'schizophrenia' set was +0.34 (+0.55).

FIG. 3. Correlations between ATA solution and average human rating, and average correlation between ATA solution and individual assessors, for n-grams between 3 and 25 characters in length, for 'working memory' document set.



For the 'working memory' document set the peak MEANCORR was r(26)=+0.67, and occurred with a 13-gram (see Figure 3). The strongest INDCORR was +0.29 (n=18, sd=0.18) and occurred with a 12-gram.

For the 'schizophrenia' document set the peak MEANCORR was r(26)=+0.85, was obtained with a 16-gram (see Figure 4). The strongest INDCORR was +0.52 (n=18, sd=0.12) and also occurred with a 16-gram.

FIG. 4. Correlations between ATA solution and average human rating, and average correlation between ATA solution and individual assessors, for n-grams between 3 and 25 characters in length, for 'schizophrenia' document set.

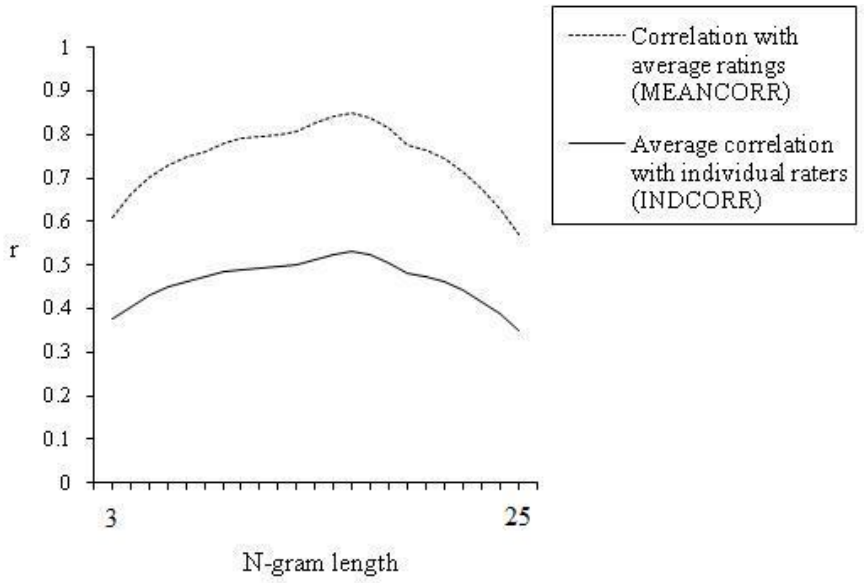
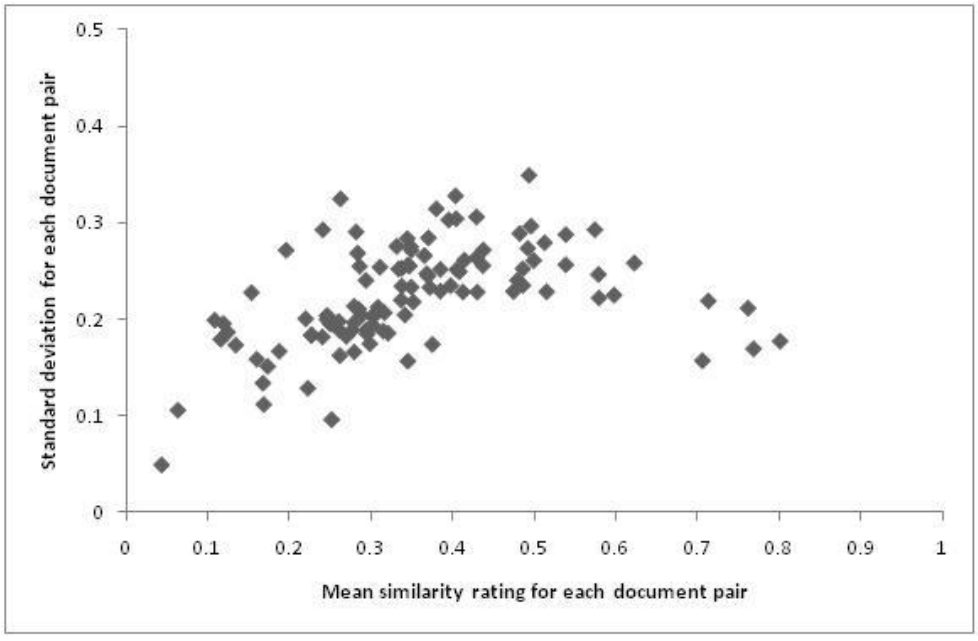


FIG. 5. Scatterplot of mean similarity ratings against the standard deviation of ratings for each document pair. Data from each document set have been combined.



Discussion

This paper reports two experiments that examined the reliability of human assessments of inter-document similarity for four small sets (two with non-technical content and two with technical content). The extent to which inter-assessor agreement depended on document (dis)similarity was also tested; and the strengths of associations between human assessments and ATA solutions—generated using the Vector Space Model with varying length n-grams as terms—were analysed.

Our results show inter-assessor reliability lower than found in previous studies. We attribute this to having gathered full sets of ratings from each participant, whereas previous studies produced estimates of inter-assessor reliability based on partial data sets. An important difference was identified between results obtained by correlating ATA solutions with an average rating for the sample (MEANCORR) versus averaging the individual correlations with ATA solutions (INDCORR) (see Figures 1-4). The former were consistently stronger. It seems that studies based on partial rating sets in which correlations involve averages (e.g., Lee et al., 2005) run the risk of overestimating reliability and supporting an overly optimistic view of the need for human testing when assessing information retrieval systems. As expected, in the reported studies human agreement was weaker for the 'technical' document sets. Overall, this is an important demonstration that variability of human assessments of document similarity can be substantial—depending, at least in part, on document content—and indicates that inter-individual variability and text content must be given prominent consideration as part of ATA system evaluations—particularly those relating to the browsing of document sets where context is weakly defined.

Of course, it might be argued that low inter-assessor reliability in these experiments is the result of 'noisy' data—due to participants experiencing ratings fatigue or lack of motivation—and that this would not apply beyond the experimental scenario. Further analyses of the data tend not to support this conclusion. Moderate to strong correlations with ATA solutions would not occur with random data. Moreover,

correlations between the mean ratings of document pairs and the variability for the respective pair indicate that for some document pairs—predominantly those that were identified as being relatively dissimilar to one another—agreements were relatively good. This would not be expected if participants' judgments were random. There were strong positive correlations between average ratings of semantic similarity and variability of ratings for three of the four document sets. When data were aggregated across the four document sets (see Figure 5) a tendency was also apparent for agreement to increase at higher levels of similarity—although there were insufficient 'high similarity' data points to have confidence in this trend. In summary, it would seem that there are some document pairs that are generally agreed to be strongly dissimilar; some document pairs for which there may be relatively good agreement that they are strongly similar; and many documents that are not considered strongly similar or strongly dissimilar, and for which there is generally weak agreement.

It might then be argued that a strong negative correlation between average ratings and rating variability indicates that participants' analyses of document contents in these experiments were not sufficiently detailed to identify subtle differences between documents, so only extreme differences were reliably detected. Level of analysis and degree of understanding will be a factor of human interaction in any setting. However, again, the data indicate that this is not sufficient explanation, as the non-technical document set with the stronger inter-assessor reliability had a relatively weak correlation between average ratings and variability of ratings. This suggests that inter-assessor reliability was not dependent on the identification of strong dissimilarity and leads us to believe that inter-document similarity judgments are generally also influenced by the effects of individual differences in schema that relate to the material contained in the document sets; schema that each participant brings to the task. This effect will be exacerbated in situations where document sets contain more complex material and where no task context is provided—as was the case for these experiments. However, this is not an unrealistic situation, as many 'real world' information retrieval tasks start with poorly defined task goals (Belkin, 1982).

Even when inter-assessor agreement was poor, ATA using an optimal length n-gram was able to produce strong correlations with average human ratings. The smooth, graded transitions between correlations for n-grams of different lengths suggests that lack of human agreement is not problematic in circumstances where human assessments are being used as the basis for evaluation of different ATA algorithms—as long as there are sufficient human assessors to provide a stable average. Moreover, once optimal n-gram length has been determined ATA can provide a good approximation of average human response. This is valuable information in the context of the design of information spaces (e.g., Skupkin & Fabrikant, 2003; Lin, 1997). The data from these experiments suggest it may be advantageous to use ATA (with n-grams) for such purposes even if document sets are of a size and stability where they could be catalogued by a human assessor.

There was no consistent difference in the size of the human-ATA correlation based on whether the document set contained non-technical (Experiment 1) or technical (Experiment 2) material. However, optimal n-gram length did discriminate between these two document types. For nontechnical material the optimal n-gram length (5- to 8-gram) was similar to that previously reported by Lee et al. (2005). However, for technical material (not previously investigated with a variable n parameter), the optimal n-gram length was substantially longer (13- to 16-gram). It may be that longer n-grams provide a more intricate assessment of the conceptual structure of documents, by emphasizing the frequency of word co-occurrence. In contrast, shorter n-grams emphasize more simplistic 'word level' information. If the most effective length of n-gram varies in a predictable manner with the properties of the document, a dynamic system of ATA that takes into account document characteristics such as average word length or reading ease could produce optimised performance (cf. Morris, 2010). Related to this, it is important to note that in these experiments the nature of the effects of n-gram length on associations with human assessments were different to those reported by Lee at al. (2005). In these experiments an optimal region existed for each document set. Lee et al. (2005) found an asymptotic pattern of associations such that there was no (or little) cost for longer n-grams (other than increased computing costs). On the basis of Lee et al.'s (2005) data a

cautious strategy might be to select an overly long n-gram. However, the data reported here suggest this may result in poorer ATA performance.

In summary, in these experiments the reliability of human assessments of inter-document semantic similarity was moderate to poor—and substantially weaker than previous estimates. Generally people were more consistent in their assessments of documents that were very dissimilar (although there were indications that the same may apply for documents that were very similar) and for non-technical document sets. ATA using n-grams as terms provided a good approximation of the average human assessment (MEANCORR). The somewhat weaker correlations between ATA and individual ratings (INDCORR) may reflect the extent to which common ground exists in the schemas that people bring to an information set. This may be inherent in all conceptual spaces due to people's need to communicate certain fundamental constructs (see e.g., Gardenfors, 2000). Consistent with this position there is some evidence that users of information spaces are able to use an 'averaged' space to positive effect (Westerman & Cribbin, 2000).

References:

Belkin, N. J., Oddy, R., & Brooks, H. (1982). ASK for Information Retrieval: Part 1, Background and Theory. *Journal of Documentation, 38*, 2, 61-71.

Belz, A. & Reiter, E. (2006). Comparing automatic and human evaluation of NLG systems. In Proceedings of EACL.

Cavnar, W.B. (1995). Using an n-gram based document representation with a vector processing retrieval model. *Proceedings of the Third Text REtrieval Conference. NIST Special Publication 500-226*. Gaithersburg, MD: National Institute of Standards and Technology.

Cavnar, W.B. & Trenkle, J.M. (1994). N-gram based text categorization. *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval.*

Damashek, M. (1995a). Gauging similarity with n-grams: Language-independent categorization of text. *Science, 267,* 843-848.

Damashek, M. (1995b). Performance of text retrieval systems. Science, 268, 1417-1418.

Flesch, R. (1949). *The Art of Readable Writing.* New York: Harper and Evanston, New York.

Fox, K.L., Frieder, O., Knepper, M.M., Snowberg, E.J. (1999). SENTINEL: A multiple engine information retrieval and visualization system. *Journal of the American Society for Information Science, 50,* 616-625.

Furnas, G., Landauer, T., Gomez, L., & Dumais, S. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30,* 11, 964-971.

Gardenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought.* Cambridge, MA: MIT Press.

Haenggi, D., Gernsbacher, M.A., & Bolliger, C.M. (1994). Individual differences in situation-based inferencing during narrative text comprehension. In H. Van Oostendorp & R.A. Zwaan (Eds.), *Naturalistic Text Comprehension* (pp. 76-97). Norwood, NJ: Ablex.

Harman, D. (1995). Overview of the second text retrieval conference (TREC-2). *Information Processing and Management, 31,* 3, 271-288.

Harman, D. *et al.*, (1995). Performance of text retrieval systems. *Science, 268,* 1417-1418.

Harman, D.K. & Voorhees, E.M. (2006). TREC: An overview. *Annual Review of Information Science and Technology, 40,* 113-155.

Hjørland, B. (2010). The foundation of the concept of relevance. *Journal of the American Society for Information Science and Technology, 61,* 217-237.

Järvelin, K. (2007). An analysis of two approaches in information retrieval: From frameworks to study designs. *Journal of the American Society for Information Science and Technology, 58,* 7, 971-986.

Johnson-Laird, P. (1983). *Mental Models.* Cambridge: Cambridge University Press.

Landauer, T.K. & Dumais, S.T. (1997). *Psychological Review, 104,* 211-240.

Lee, M.D., Pincombe, B., & Welsh, M. (2005). An empirical evaluation of models of text document similarity. In *Proceedings of 27th Annual Conference of the Cognitive Society (pp. 1254-1259).* Mahawah, NJ: Erlbaum.

Lin, X. (1997). Map displays for information retrieval. *Journal for the American Society for Information Science, 48,* 1, 40-54.

Lund, K. & Burgess, C. (1996). *Behavior Research Methods, Instruments, & Computers, 28,* 203-208.

Miller, G.A. & Charles, W.G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes, 6,* 1, 1-28.

Morris, J. (2010). Individual differences in the interpretation of text: Implications for information science. *Journal of the American Society of Information Science and Technology, 61,* 1, 141-149.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research, 11,* 95-130.

Salton, G. (1995) Performance of text retrieval systems. *Science, 268,* 5216, 1418-1419.

Saracevic, T. (2008). Effects of inconsistent relevance judgments on information retrieval test results: A historical perspective. *Library Trends, 56,* 4, 763-783.

Skupin, A., & Fabrikant, S. (2003). Spatialization Methods: A Cartographic Research Agenda for Non-Geographic Information Visualization. *Cartography and Geographic Information Science, Transitions in U.S. Cartography and Geographic Information Science, 30,* 2, 95-119.

Voorhees, E. & Harman, D. (Eds.) (1998). *The Sixth Text Retrieval Conference (TREC-6).* National Institute of Standards and Technology, Gaithersburg, MD.

Westerman, S.J. & Cribbin, T. (2000). Mapping semantic information in virtual space: Dimensions, variance, and individual differences. *International Journal of Human-Computer Studies, 53,* 765-787.