Pairwise Similarity of TopSig Document Signatures

Christopher M. De Vries
Electrical Engineering and Computer Science
Queensland University of Technology
Brisbane, Australia
chris@de-vries.id.au

Shlomo Geva
Electrical Engineering and Computer Science
Queensland University of Technology
Brisbane, Australia
s.geva@qut.edu.au

ABSTRACT

This paper analyses the pairwise distances of signatures produced by the TopSig retrieval model on two document collections. The distribution of the distances are compared to purely random signatures. It explains why TopSig is only competitive with state of the art retrieval models at early precision. Only the local neighbourhood of the signatures is interpretable. We suggest this is a common property of vector space models.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval Models*

Keywords

Signature Files, Topology, Vector Space IR, Random Indexing, Document Signatures, Search Engines, Document Clustering, Near Duplicate Detection, Relevance Feedback

1. INTRODUCTION

This paper investigates the properties of the pairwise similarities of document signatures produced by TopSig. TopSig is a retrieval model where documents are represented by d-bit binary strings that lie on a d-dimensional collection hypercube. The signatures are produced by a random process called random indexing [10] or random projection [1] which compresses the standard term-by-document matrix.

Pairwise similarity plays an important role in many information retrieval related tasks such as ad hoc retrieval, clustering, classification, filtering, near duplicate detection and relevance feedback.

The paper proceeds as follows. In Section 2, the TopSig retrieval model is introduced. Section 3 describes the document collections used in the experiments. The experimental setup is introduced in Section 4 and the results are presented in Section 5. The paper is concluded by a discussion of the implications of the results in Section 6.

2. TOPSIG

TopSig [7] offers a radically different approach to the construction of file signatures. Traditional file signatures [6] have been shown to be inferior to approaches using inverted indexes, both in terms of the time and space required to process and store the index [12, 13]. However, TopSig overcomes previous criticisms aimed at file signatures by taking a principled approach using the vector space model, dimensionality reduction and numeric quantisation. Previous approaches to file signatures were constructed in an ad hoc fashion by combining random binary signatures using a bitwise XOR which is a Bloom filter [2] for the terms contained in documents. In contrast, TopSig randomly indexes a weighted term-by-document matrix and then quantises it. TopSig is competitive with state of the art probabilistic and language retrieval models at early precision, and clustering approaches [7].

Let $D = \{d_1, d_2, ..., d_n\}$ be a document collection of n documents signatures, $D \subset \{+1, -1\}^d$, $|D| = n$. Let $F = \{f_1, f_2, ..., f_n\}$ be the same document collection as D where each document is represented by a v-dimensional real valued vector, $F \subset \mathbb{R}^v$, $|F| = n$, where v is the size of the vocabulary of the document collection. F is the term-by-document matrix in the full space of the collection vocabulary which underlies most modern retrieval systems.

TopSig indexes documents using a mapping function, $m : \mathbb{R}^v \rightarrow \{+1, -1\}^d$, that maps a document from the original v-dimensional continuous real valued term space, to a d-dimensional discrete binary valued space. The index is constructed using a mapping function, $D = \{f \in F : m(f)\}$. The mapping function creates a sparse random ternary index vector of d-dimensions for each term in the document with +1 and -1 values in random positions and the majority of positions containing 0 values. These randomly generated codes are almost orthogonal to each other and have been shown to provide comparable quality to orthogonal approaches such as principle component analysis [1]. The index vector is multiplied by the term weight and added to a d-dimensional real valued vector that represents the document. Once all the terms in a document have been processed, this reduced dimensionality document vector is then quantised to a d-dimensional binary vector by thresholding each value in each dimension to 1 if greater than 0 and 0 otherwise. The 1 and 0 values in the binary vector represent +1 and -1 values. This mapping function can be applied to each document independently, meaning that new documents can be indexed in isolation without having to update the existing index. This is a key advantage to random indexing [10] over other dimensionality reduction techniques such as latent semantic analysis [5] which requires global analysis of the term-by-document matrix using the singular value decomposition [8].

The indexing process of TopSig is similar to that of SimHash [3]. However, TopSig uses signatures an order of magnitude longer than SimHash and it uses much sparser random

codes. The search process for ad hoc retrieval also differs, where TopSig searches in the subspace of the query and applies relevance feedback.

The binary vectors in D provide a faithful representation of the original document vectors in F. The topological relationships in the original space are preserved in the reduced dimensionality space. This is supported by the Johnson-Lindenstrauss lemma [9] that states if points in a high-dimensional space are projected into a randomly chosen subspace, of sufficiently high-dimensionality, then the distances between the points are approximately preserved. It also states that the number of dimensions required to reproduce the topology is asymptotically logarithmic in the number of points.

3. DOCUMENT COLLECTIONS

We have used the INEX Wikipedia 2009 collection and the TREC Wall Street Journal (WSJ) Collection to evaluate pairwise distances of TopSig signatures. The INEX Wikipedia collection contains 2,666,190 documents with a vocabulary of 2,132,352 terms. We have used 2 subsets of this collection during evaluation. The first is a 144,265 document subset used for the INEX 2010 XML Mining track [4]. This is the reference run for the ad hoc track in 2010 produced by an implementation of Okapi BM25 in the ATIRE search engine [11]. It is denoted by $\text{INEX}_{reference}$. The second is a randomly selected 144,265 document subset chosen to match the size of the XML Mining subset. It is denoted by INEX_{random}. Subsets of the INEX Wikipedia 2009 collection were used for this experiment because calculating pairwise distances has a time complexity of $O(n^2)$ and becomes intractable for millions of documents. The mean document length in the Wikipedia has 360 terms, the shortest has 1 term and the longest has 38,740 terms. The Wall Street Journal Collection consists of 173,252 documents and a vocabulary of 113,288 terms. The mean WSJ document length is 475 terms, the shortest has 3 terms, and the longest has 12,811 terms.

The INEX Wikipedia 2009 collection consists of 12GB of uncompressed text or 50GB of uncompressed XML which includes semantic markup. The 2,666,190 documents are split into 3,617,380 passages. 1024-bit TopSig signatures use a total of 441MB to index the collection. The TREC Wall Street Journal consists of 518MB of uncompressed text. The 173,252 documents are split into 222,238 passages. 1024-bit TopSig signatures use a total of 27MB to index the collection.

4. EXPERIMENTAL SETUP

Pairwise similarities define the topology of a set of documents. Each document is compared to every other document. These similarities define the relationships between all documents in a collection. If two documents are nearby each other they share the same semantic context. TopSig uses the Hamming distance to measure similarity between two documents. It produces values in the range $[0, d]$ where 0 indicates the documents are indentical and values from 1 to d indicate decreasing similarity between documents where d is the most dissimilar two documents can be.

The TopSig indexing process uses random codes to compress document vectors. These random codes are also called index vectors in the random indexing process. The codes are influenced by the original document vectors. Similar documents are placed close together in the reduced binary vector space that are close together in the original vector space. Therefore, it is expected that the pairwise relationships between documents will be biased by this process. If the indexing process has no effect then the document signatures would appear no different to purely random signatures. The pairwise distances between randomly generated random signatures can be described by the Binomial distribution. The distribution of pairwise distances produced by the TopSig indexing process can be estimated by creating a histogram of similarity counts at all Hamming distances.

All $\frac{n^2-n}{2}$ pairwise distances between document signatures in D are calculated. This is all the similarities contained in the upper triangular form of the pairwise distance matrix without the entries along main diagonal. The lower half of the pairwise distance matrix does not need to be calculated as the Hamming distance is symmetric. Measuring a pair of signatures both ways around does not add any extra information. The Hamming distance similarity function, $s : \{+1, -1\}^d \times \{+1, -1\}^d \rightarrow \mathbb{N}$, is symmetric such that two documents compared in either order produce the same result, $d_x, d_y \in D : s(d_x, d_y) = s(d_y, d_x)$. The estimated probability of finding a signature at Hamming distance, h, is the fraction of similarities at that distance over the total number of distance comparisons. The probability mass function, $pmf_e : \mathbb{P}\{+1, -1\}^d \times \mathbb{N} \rightarrow \mathbb{R}$, produces the estimated probability from the pairwise distances in D where n is the number of signatures in the collection D, $|D| = n$,

$$pmf_e(D, h) = \frac{\left|\left\{(d_x, d_y) : \begin{array}{c} d_x, d_y \in D \wedge d_x \neq d_y \\ \wedge s(d_x, d_y) = h \end{array}\right\}\right|}{\frac{n^2-n}{2}}. \quad (1)$$

Note that pmf_e is the estimated probability for finding a signature at distance, h, when averaged across all documents in the collection, D.

The probability of finding a random binary code of length, d, at Hamming distance, h, is described by the Binomial probability mass function, $pmf_b : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}$,

$$pmf_b(d, h) = \left(\begin{array}{c} d \\ h \end{array} \right) p^h (1-p)^{d-h}, \quad (2)$$

where p is the probability of a bit being set, $p = 0.5$.

The cumulative distribution function for either the estimated, $cdf_e : \mathbb{P}\{+1, -1\}^d \times \mathbb{N} \rightarrow \mathbb{R}$, or Binomial, $cdf_b : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}$, probability distributions are the sum of the probability mass function from 0 to h,

$$cdf_e(D, h) = \sum_0^h pmf_e(D, h), \quad (3)$$

$$cdf_b(d, h) = \sum_0^h pmf_b(d, h). \quad (4)$$

An implementation of the TopSig [1] search engine was used to index the document collections. It splits documents into passages on a sentence boundary between a minimum and maximum number of word tokens. If the maximum word token limit is reached before the end of a sentence, it is split at that point. Therefore, documents have multiple signatures. This has been found to be effective for retrieval of

[1] `http://topsig.googlecode.com`

Collection	Documents	Signatures (Passages)
$\text{INEX}_{reference}$	144,265	328,207
INEX_{random}	144,265	195,369
WSJ	173,252	222,238

Table 1: Number of Signatures Generated by TopSig

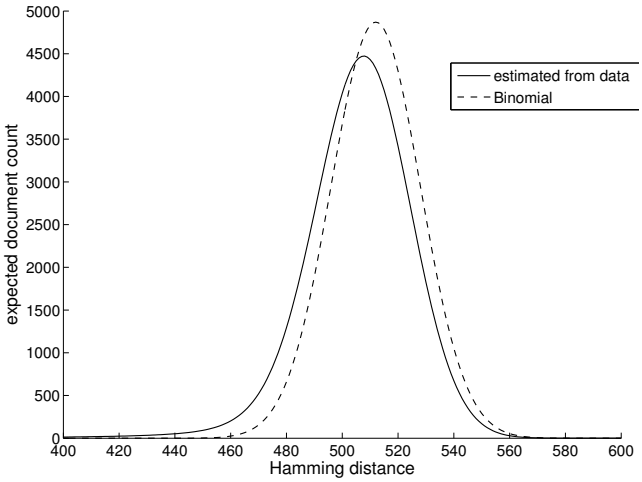

Figure 1: INEX_{random} pmf

documents of varying length. The INEX collection was split on a minimum of 256 and maximum of 280 word tokens. The WSJ collection was split on a minimum of 256 and a maximum of 384 word tokens. All indexes use 1024-bit signatures, resulting in the number of signatures as listed in Table 1.

The resulting probability distributions have been multiplied by the number of signatures in a collection to produce the expected number of signatures at a given Hamming distance. In this case, the *pmf* gives the average number of signatures expected at a particular Hamming distance when comparing a signature to the entire collection. The *cdf* gives the average number of signatures expected to lie within a given Hamming distance when comparing a signature to the entire collection, i.e. the number of nearest neighbours to expect within a particular Hamming distance.

5. EXPERIMENTAL RESULTS

Figures 1 to 12 highlight the difference between the distributions estimated from the pairwise distances and the distributions expected from random binary signatures from the Binomial distribution. It can be seen that all the estimated distributions are left skewed towards a Hamming distance of 0. This indicates that the signatures produced by TopSig are biased in such a way that documents are more similar to each other. There are more documents expected at a more similar, lower Hamming distance, than expected at random.

The probability mass functions in Figures 1, 2 and 3 represent the expected number of signatures to be seen at a particular Hamming distance. The graphs have been centred around the middle of the distributions to allow better visualisation of the separation between the distributions. The

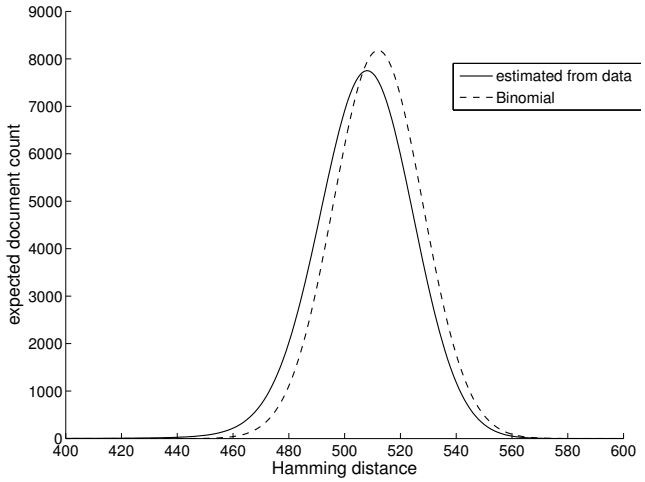

Figure 2: $\text{INEX}_{reference}$ pmf

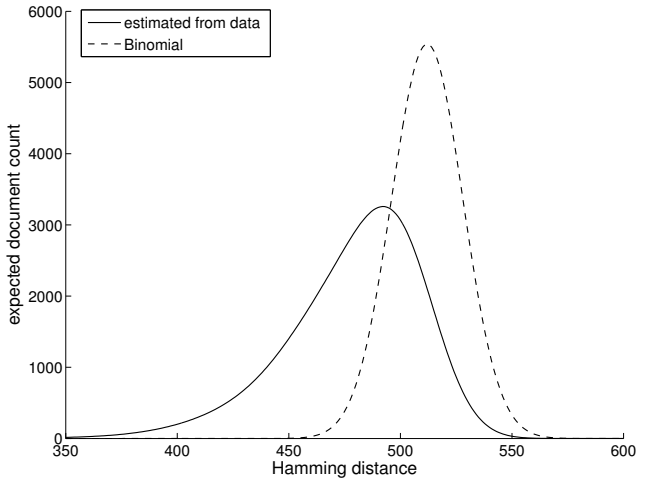

Figure 3: WSJ pmf

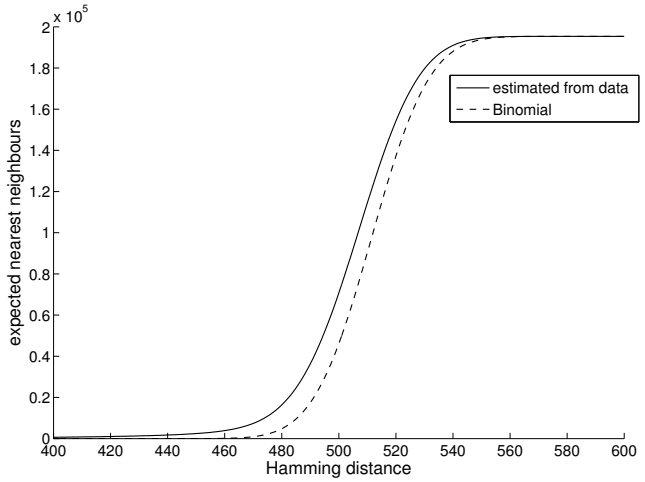

Figure 4: INEX$_{random}$ *cdf*

tails of the distributions tend towards 0 as expected. For example, the graph in Figure 3 has a y value for the estimated distribition of 1033.39 at a Hamming distance of 441. When comparing a signature to the entire collection, it would be expected on average to encounter 1033.39 signatures that are exactly at a Hamming distance of 441. However, the expected number of signatures at a Hamming distance of 441 for purely random signatures is only 0.29. This suggests that the signatures produced by TopSig are not uniformly distributed throughout the feature space. The number of nearest neighbours at a given Hamming distance, as described by the pmf, quickly increases when starting from a Hamming distance of 0 and proceeding to a Hamming distance of d. This is the same order that TopSig ranks signatures in the ranked list, or, any other task that compares relative orderings of documents such as clustering. This is true for both the estimated and Binomial distributions. As the neighbourhood of analysis is increased, more and more documents become equidistant; i.e. they share the same Hamming distance. This is a property of vector space models known as the "curse of dimensionality". However, the left skewness of estimated distributions indicates that the pairwise distances of the document collections allow better differentiation between documents than expected at random. It is this left skewness of the distributions that allows TopSig to compete with state of the art retrieval models at early precision. Documents are topically clustered and are not random bags of words. Neither of the document collections have signatures further apart than a Hamming distance of 617, meaning that the indexing process has moved the random signatures from the right side of the distribution to the left. This again indicates that similar signatures are being placed closer together and are therefore more topically related and clustered.

The cumulative distribution functions in Figures 4, 5 and 6 represent the area under the curve for each of the probability mass functions. The y value at a given Hamming distance indicates the average number of nearest neighbours expected within a given Hamming distance when comparing a signature to the entire collection. For example, the graph in Figure 6 has a y value for the estimated distribition of 25975.78 at a Hamming distance of 441. When comparing a signature to the entire collection, it would be expected on

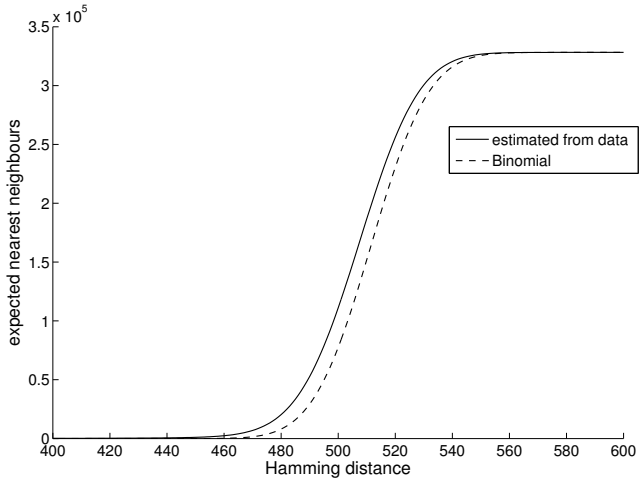

Figure 5: INEX$_{reference}$ *cdf*

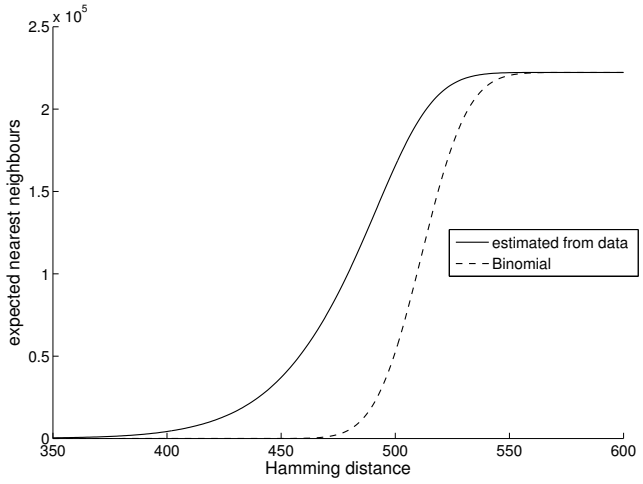

Figure 6: WSJ *cdf*

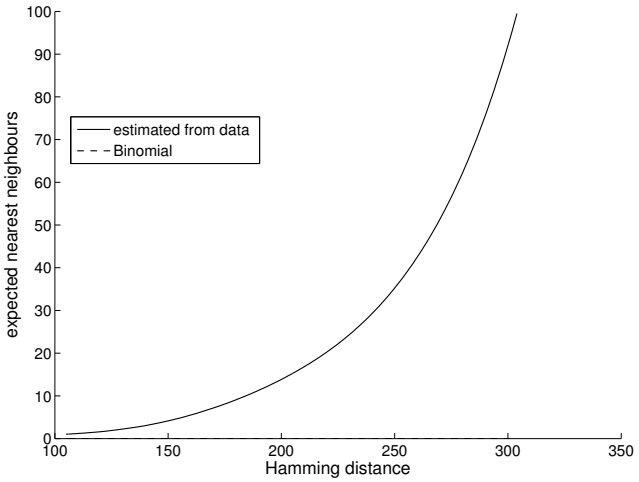

Figure 7: INEX$_{random}$ *cdf*
First 100 signatures from estimated distribution

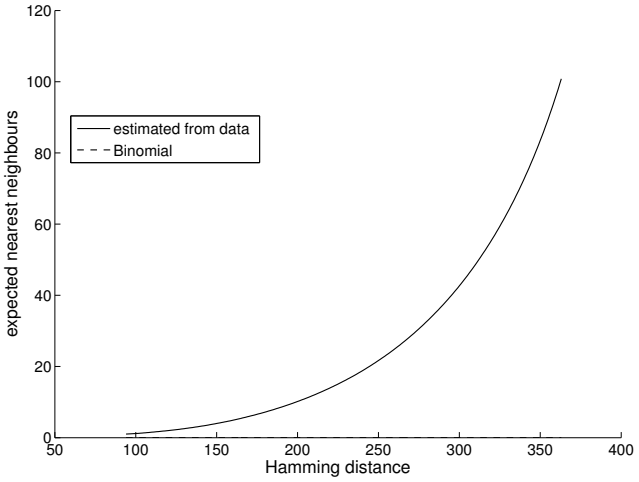

Figure 8: INEX$_{reference}$ *cdf*
First 100 signatures from estimated distribution

average to encounter 25975.78 signatures that are nearest neighbours at a Hamming distance of 441. However, the expected number of signatures at a Hamming distance of 441 for purely random signatures is only 1.13. Again, the separation between the curves indicates that TopSig is placing semantically related documents close together and preserving the topological relationships of the original document vectors.

Figures 7, 8 and 9 zoom in on the *cdf* where the first 100 nearest neighbours are expected for the distribution estimated from the pairwise distances of the collections. In all cases almost zero signatures are expected at random where there are TopSig signatures expected in the range $[1, 100]$. This indicates that the signatures produced by TopSig return nearest neighbours at a Hamming distance much earlier than expected by purely random signatures. The start and end points of these curves are listed in Table 2.

Figures 10, 11 and 12 zoom in on the *cdf* where the first 100 nearest neighbours are expected for random binary signatures as described by the Binomial distribution. The start

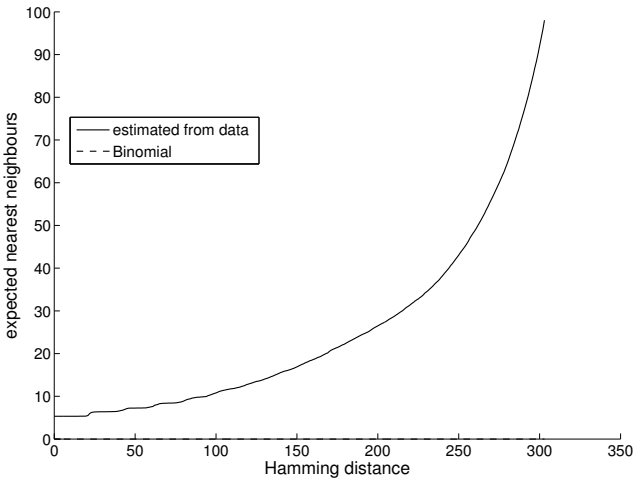

Figure 9: WSJ *cdf*
First 100 signatures from estimated distribution

Collection	cdf_b @ $cdf_e = 1$	cdf_b @ $cdf_e = 100$
INEX$_{reference}$	1.66×10^{-168}	1.104×10^{-15}
INEX$_{random}$	1.80×10^{-158}	2.06×10^{-34}
WSJ	0	1.59×10^{-34}

Table 2: Nearest Neighbours Expected from cdf_b

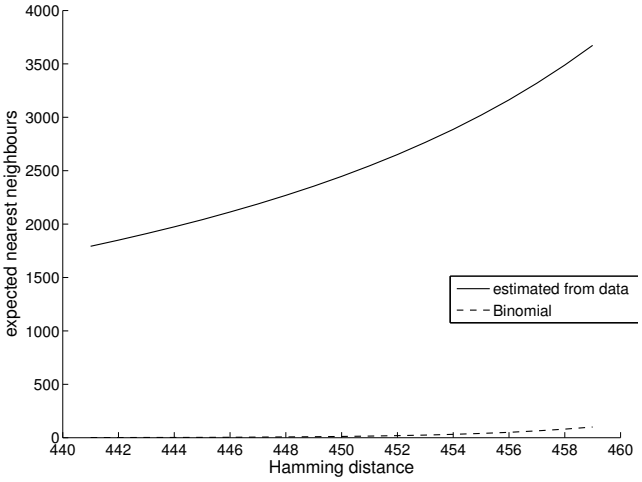

Figure 10: INEX$_{random}$ *cdf*
First 100 signatures from Binomial distribution

Collection	cdf_e @ $cdf_b = 1$	cdf_e @ $cdf_b = 100$
INEX$_{reference}$	567.86	2129.02
INEX$_{random}$	1793.26	3673.78
WSJ	25495.67	50709.94

Table 3: Nearest Neighbours Expected from cdf_e

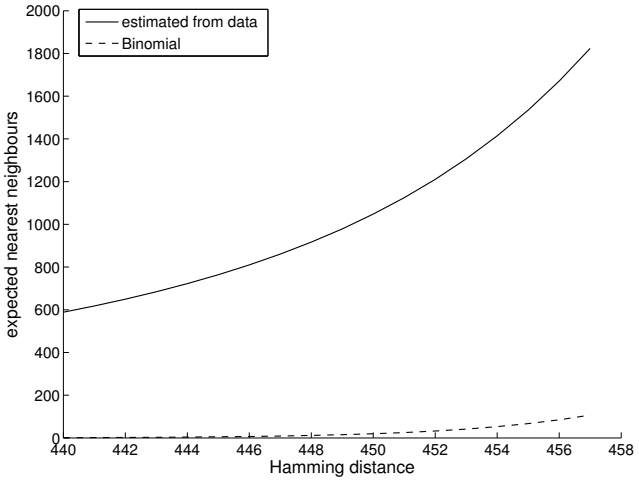

Figure 11: INEX$_{reference}$ *cdf*
First 100 signatures from Binomial distribution

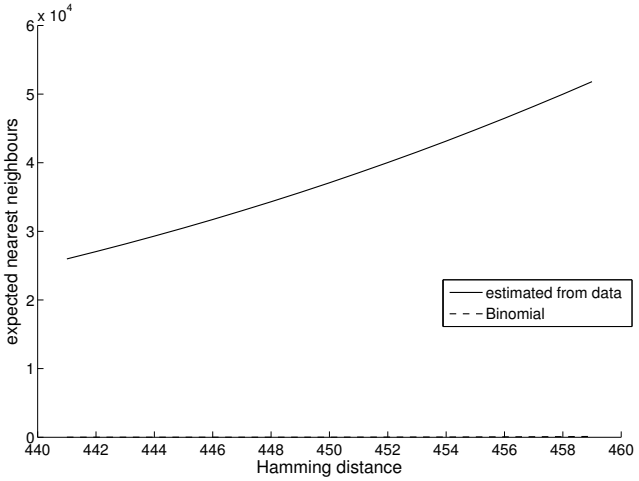

Figure 12: WSJ *cdf*
First 100 signatures from Binomial distribution

Collection	$\frac{cdf_b@h=512}{n}$	$\frac{cdf_e@h=512}{n}$
INEX$_{reference}$	0.51	0.61
INEX$_{random}$	0.51	0.63
WSJ	0.51	0.89

Table 4: Signatures Expected within $\frac{d}{2}$

and end points of these curves are listed in Table 3. There are many more signatures expected to be nearest neighbours when using TopSig signatures. However, both the estimated and Binomial distributions have many equidistant documents around the middle of their distributions. This suggests that only the local neighbourhood of the signatures has semantic meaning. This can also be seen in the pmf distributions where most of the signatures exist around the middle of the distribution. Another perspective is that are too many ties at these distances for the feature space to differentiate signatures.

The skewness of the estimated distributions suggests that the feature space is not uniform and is clustered. Some areas of the space are more dense than others. This is vital for any document representation because it is this non-uniformity that allows differentiation of meaning.

Table 3 lists the number of nearest neighbours expected from the distribution estimated from pairwise distances when the Binomial distribution expects 1 and 100 nearest neighbours, as listed in columns 2 and 3 respectively. For example, the INEX$_{random}$ collection expects on average 1793.26 signatures to be nearest neighbours to other signatures when purely random signatures would expect 1. When purely random signatures expect on average 100 nearest neighbours, the INEX$_{random}$ collection expects 3673.78 nearest neighbours. These values are linearly interpolated as they exist in between two Hamming distances under the *cdf*. These values are the start and end points for the curves in Figures 10, 11 and 12. Table 2 lists the opposite, i.e., the number of nearest neighbours expected from the Binomial distribution when the distribution estimated from pairwise distances expects at 1 and 100 nearest neighbours.

Table 4 lists the number of signatures expected within a Hamming distance of $\frac{d}{2}$. This summarises the distributions in a single number, where the difference between the distributions indicates the fraction of the signatures shifted from the left hand side of the Binomial distribution to the right by the indexing process. It is also the value under the pmf at $\frac{d}{2}$ which is also the y value of the *cdf* at $\frac{d}{2}$.

The difference in distributions between the INEX reference and random subsets indicates that the reference run is not suitable for estimating properties of the entire collection. This is to be expected as the reference run has been biased by the queries used for ad hoc retrieval. Table 3 shows that INEX$_{reference}$ expects 567.86 nearest neighbours where as INEX$_{random}$ expects 1793.26 nearest neighbours when purely random signatures expect 1 nearest neighbour. This indicates that the reference run is less clustered than than a random sample from the INEX Wikipedia collection. This can explained because the documents returned by the reference run are more diverse than a random sample from the collection. As the diverse topics are further apart, i.e. more dissimilar, there are more inter-topic distances than

intra-topic distances, leading to less signatures being located nearby. Note that the reference run is determined by only searching in a few dimensions determined by the keywords in the queries, where as the pairwise distances compare entire documents, using their entire vocabulary.

6. DISCUSSION

The results presented indicate why TopSig is only competitive at early precision in comparison to probabilistic and language models for ad hoc retrieval. As the Hamming distance increases when proceeding down the ranked list more and more documents become equidistant. This can be seen in Figures 1, 2 and 3 containing plots of probability mass functions indicating the expected number of documents at a given Hamming distance. The curves quickly increase to the point where thousands of documents are equidistant. This is likely to be a property of any vector space model due to the "curse of dimensionality". Only the tails of the distribution of distances are useful for differentiation of relevant and non-relevant documents.

Approaches to near duplicate detection such as SimHash [3] use short signatures that are 64-bits in length. This only allows the few nearest neighbours to be differentiated which is adequate for near duplicate detection. This can be explained by the probability mass functions in Figures 1, 2 and 3. The x axis for 64-bit signatures will only contain 65 positions for the Hamming distances 0 to 64. As the number of equidistant documents is a function of the x value, or, Hamming distance, many documents will appear equidistant much sooner than with signatures of 1024-bits in length. The same curve has to be squeezed into 65 positions instead of 1025 positions. A duplicate is expected to be very similar to other documents it is a duplicate of, so these short signatures will suffice. In contrast, TopSig uses much longer signatures that allow for better separation for tasks such as ad hoc retrieval and clustering.

Document clustering places similar documents into groups of topically related documents. The results presented in this paper suggest that only document clusters that exist within the local neighbourhood of a vector space are interpretable. As the documents within a cluster become more dissimilar, the grouping of these documents loses its meaning for the same reason precision at higher recall suffers in ad hoc retrieval, there are many equidistant documents that are unable to be differentiated from one another. This suggests that only a large number of smaller document clusters are meaningful. The maximum interpretable radius for a document cluster can be esimated heuristically from the distributions of estimated from the pairwise data. This heuristic is to stop at the point where the distribution starts to sharply increase. In Figure 1 this would be approximately a Hamming distance of 450, or, the point before the elbow in the left hand side of the distribution occurs.

Furthermore, TopSig is likely to be useful for increased computational efficiency of document-to-document comparisons. Examples of this include clustering, classification, filtering, relevance feedback, near duplicate detection and explicit semantic analysis. All of these tasks can exploit the left tails of the probabilty mass function distributions depicted in Figures in 1, 2 and 3. In fact, TopSig has been shown to provide a 1 to 2 magnitude increase in processing speed for document clustering [7] over traditional sparse vector representations.

The analysis presented in this paper is expected to be useful for any vector space model. It would be expected that similar behaviour would be exhibited whether comparing entire documents in the full vocabulary space of the term-by-document matrix or comparing dimensionality reduced documents in a continuous space such as those produced by latent semantic analysis, principal component analysis or random indexing.

7. REFERENCES

[1] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *KDD 2001*, pages 245–250. ACM, 2001.

[2] B.H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970.

[3] M.S. Charikar. Similarity estimation techniques from rounding algorithms. In *STOC*, pages 380–388, New York, NY, USA, 2002. ACM.

[4] C.M. De Vries, R. Nayak, S. Kutty, S. Geva, and A. Tagarelli. Overview of the INEX 2010 XML mining track. *INEX 2010, LNCS*, 2011.

[5] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[6] C. Faloutsos and S. Christodoulakis. Signature files: an access method for documents and its analytical performance evaluation. *ACM Trans. Inf. Syst.*, 2:267–288, October 1984.

[7] S. Geva and C.M. De Vries. TopSig: topology preserving document signatures. In *CIKM 2011*, pages 333–338. ACM, 2011.

[8] G. Golub and W. Kahan. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis*, 2(2):205–224, 1965.

[9] W.B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189-206):1–1, 1984.

[10] M. Sahlgren. An introduction to random indexing. In *TKE 2005*, 2005.

[11] A. Trotman, X. Jia, and S. Geva. Fast and effective focused retrieval. In *Focused Retrieval and Evaluation*, LNCS, pages 229–241. 2010.

[12] I.H. Witten, A. Moffat, and T.C. Bell. *Managing gigabytes: compressing and indexing documents and images*. Morgan Kaufmann, 1999.

[13] J. Zobel, A. Moffat, and K. Ramamohanarao. Inverted files versus signature files for text indexing. *ACM Trans. Database Syst.*, 23:453–490, December 1998.