Geometric Representations for Multiple Documents

Jangwon Seo
jangwon@cs.umass.edu

W. Bruce Croft
croft@cs.umass.edu

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA 01003

ABSTRACT

Combining multiple documents to represent an information object is well-known as an effective approach for many Information Retrieval tasks. For example, passages can be combined to represent a document for retrieval, document clusters are represented using combinations of the documents they contain, and feedback documents can be combined to represent a query model. Various techniques for combination have been introduced, and among them, representation techniques based on concatenation and the arithmetic mean are frequently used. Some recent work has shown the potential of a new representation technique using the geometric mean. However, these studies lack a theoretical foundation explaining why the geometric mean should have advantages for representing multiple documents. In this paper, we show that the arithmetic mean and the geometric mean are approximations to the center of mass in certain geometries, and show empirically that the geometric mean is closer to the center. Through experiments with two IR tasks, we show the potential benefits for geometric representations, including a geometry-based pseudo-relevance feedback method that outperforms state-of-the-art techniques.

Categories and Subject Descriptors

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

General Terms

Algorithms, Measurement, Experimentation

Keywords

multiple documents, information geometry, geometric mean

1. INTRODUCTION

A typical goal in Information Retrieval (IR) is to find relevant documents, where we rank the documents using a representation for a single document. Often, however, a representation for multiple documents is needed. For example, tasks such as relevance feedback, passage retrieval and resource selection in distributed information retrieval or in aggregated search, use representations for sets of multiple documents.

One of standard approaches for relevance feedback is to estimate an underlying relevance model from given feedback documents and sample likely terms from the model for query expansion. That is, the estimated underlying model can be considered as a representation of the feedback documents. In passage retrieval, representations of text passages can be used to rank passages or documents. In the latter case, we represent a document using a combination of some or all of its passages. In resource selection tasks, the resource or collection is represented using the documents in the collection.

As many tasks require representations for multiple documents, various approaches have been introduced. Among them, representation techniques based on the arithmetic mean and concatenation are frequently used. Representation techniques based on the arithmetic mean literally compute the arithmetic mean of multiple language models or vector representations. Representation techniques based on concatenation make a large document by concatenating multiple documents and use a language model or vector to represent the large document.

In addition to traditional group representation techniques, some recent studies show the potential of a new representation technique, the geometric mean representation of language models [26, 30, 11, 31]. Liu and Croft [26] compared various representation techniques for cluster retrieval and demonstrated that representations using the geometric mean outperformed others via empirical evaluation. Seo and Croft [30] applied a resource selection technique based on the geometric mean to blog site search. Moreover, Elsas and Carbonell [11] and Seo et al. [31] showed that a thread representation using the geometric mean of postings in the thread can be a good choice for online forum search.

The previous work which uses the geometric mean to represent a group of documents, however, did not theoretically analyze the geometric mean in the language modeling framework. In other words, although they have demonstrated the performance of representation techniques based on the geometric mean empirically, theoretical evidence or the assumptions behind the geometric mean have not been sufficiently addressed to justify its use in IR.

Therefore, in this paper, we give a theoretically grounded explanation for geometric mean-based techniques for representing multiple documents. To do this, we consider Information Geometry as a tool and discuss how the arithmetic mean as well as the geometric mean can be inter-

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

preted in certain geometries. More specifically, we show that the arithmetic mean and the geometric mean relate to the Fréchet sample mean which minimizes the Fréchet sample function. Furthermore, we empirically show that the geometric mean is closer to the Fréchet mean.

In addition, we address two applications considering the geometric interpretation: cluster retrieval and pseudo-relevance feedback. Particularly, for pseudo-relevance feedback, we introduce a variation of the relevance model [21], the geometric relevance model, and show that this new approach performs better than the relevance model.

The remainder of this paper is organized as follows. Section 2 reviews previous work. In Section 3, we introduce the Fréchet mean and geometric representations correspond to the Fréchet mean in two different metric spaces using Information Geometry. In Section 4, we provide empirical evidence for the geometric representations through experiments for two IR tasks. Section 5 discusses other evidence for the geometric representations. Section 6 concludes this paper.

2. PREVIOUS WORK

Combining multiple evidence is one of the most frequently addressed topics in Information Retrieval. Belkin et al. [2] showed that different representations of the same information object leads to different results and combinations of such representations can improve retrieval performance. Various combination heuristics suggested by Fox and Shaw [12] and analyzed by Lee [23] are still used in many IR tasks such as passage retrieval and resource selection. Using passage-level evidence [7, 25, 3] for document retrieval necessarily requires combination techniques. Resource selection where a collection is represented by its own documents [6, 32] actively uses combination techniques as well.

Relevance feedback (and pseudo-relevance feedback) is another task using combination-based representation techniques. To estimate a query model for query expansion, the top ranked documents are combined. Rocchio [29] introduced a feedback technique to combine positive or negative feedback documents in vector spaces. Lavrenko and Croft [21] introduced a technique that estimates a underlying relevance model in the language modeling framework. In fact, these standard relevance feedback approaches implicitly use the arithmetic mean. Recently, Collins-Thompson and Callan [9] used a parametric approach using re-sampling to estimate a posterior Dirichlet distribution for the documents. That is, they use the mean and the variance of the Dirichlet distribution to get a feedback model.

The geometric mean-based representation technique was relatively recently introduced. Liu and Croft [26] demonstrated that representation by the geometric mean works well for cluster retrieval via comparisons with vairous representation techniques. Seo and Croft [30] suggested a resource selection technique by the geometric mean for blog site retrieval. Furthermore, the technique was shown to work well for thread search in online forums [11, 31]. The geometric mean is often used in other fields. For example, Kogan et al. [18] used the geometric mean for k-means clustering. Veldhuis [34] showed that a centroid of the symmetrical Kullback-Leibler divergence is related to the arithmetic mean and the normalized geometric mean.

In this paper, to justify the use of the geometric mean in IR, we find evidence from Information Geometry. Rao [28] and Jeffreys [14] are the first people who considered the Fisher information metric as a Riemannian metric. Later, Efron [10] focused on differential geometry in statistics considering the curvature of statistical models. Recently, Lebanon [22] applied the theory to many machine learning tasks. See Amari and Nagaoka [1] and Kass and Vos [16] for comprehensive introduction to Information Geometry.

3. GEOMETRY OF MULTIPLE DOCUMENTS

We introduce the Fréchet mean and derive the mean in two different metric spaces, i.e., the Euclidean metric space and the Riemannian manifold defined by the Fisher information metric.

3.1 Fréchet Mean

Let us consider a Riemannian manifold $\mathcal{M}$ with a distance measure $dist(\mathbf{x}, \mathbf{y})$ where $\mathbf{x}$ and $\mathbf{y}$ are points on the manifold. Assume that we have a distribution Q on a convex set $\mathcal{U} \subset \mathcal{M}$. Now we define a function $F : \mathcal{M} \to \mathbb{R}$ as follows:

$$\Phi(\mathbf{c}) = \int_{\mathbf{p} \in \mathcal{U}} dist^2(\mathbf{c}, \mathbf{p}) Q(d\mathbf{p})$$

This function is known as the Fréchet function. A set of points which minimize the function is called the Fréchet mean set of Q. If there is only a point in the set, the point is called the Fréchet mean. This general notation for a center or centroid associated with a probability distribution was introduced by Fréchet [13] and Karcher [15]. This mean is called by various names, e.g., the center of mass, barycenter, Karcher mean and Fréchet mean. In this work, we refer to this mean as the Fréchet mean[1]. The concept of the Fréchet mean is general and not limited to any specific metric; accordingly, this can be applied to any metric space. Indeed, as we will see soon, it also generalizes the ordinary Euclidean mean.

Kendall [17] proved that if the support of Q is in a geodesic ball of sufficiently small radius r, then one Fréchet mean uniquely exists. As we see later, we consider a statistical manifold for multinomial distributions, and the distributions are mapped onto a simplex or a positive sphere. Since the mapped area is sufficiently small, a unique Fréchet mean exists. For example, in case of a sphere, the radius of the geodesic ball is $\pi/4$ and the positive sphere is contained in the ball.

If we have n unique points $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n$ in m i.i.d. samples from distribution Q, then we consider the sample Fréchet mean which minimizes the Fréchet sample function given by

$$\bar{\Phi}(\mathbf{c}) = \sum_{i=1}^{n} dist^2(\mathbf{c}, \mathbf{p}_i) \hat{Q}(\mathbf{p}_i) \quad (1)$$

where $\hat{Q}$ is an empirical distribution estimated from the samples.

Bhattacharya and Patrangenaru [5] showed that every measurable choice from the Fréchet sample mean set of $\hat{Q}$ is a strongly consistent estimator of the Fréchet mean of Q. In this paper, we consider multiple documents to represent as samples and the Fréchet sample mean as a representation.

[1]Strictly speaking, this is the intrinsic Fréchet mean in that we use a geodesic distance. However, since we address only the intrinsic Fréchet means in this paper, we omit term "intrinsic".

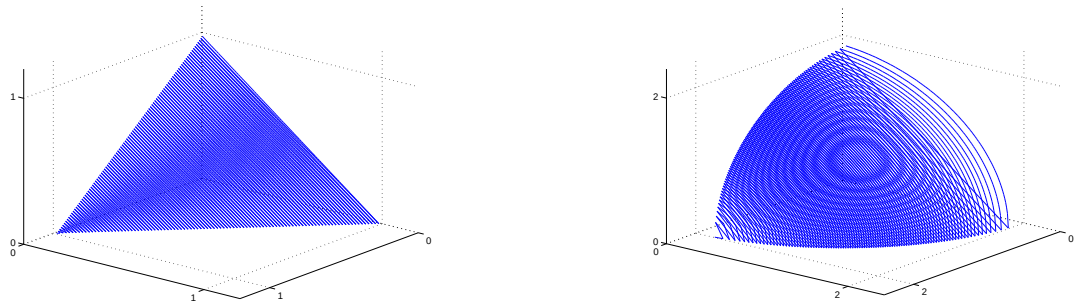

Figure 1: Assuming the Euclidean metric space, a $n+1$ dimensional multinomial distribution is mapped to a point in the n-simplex in Euclidean space (left). Assuming the Riemannian manifold defined by the Fisher information metric, the same point is mapped to a point in the positive n-sphere of radius 2 (right).

Therefore, we address how to compute the sample Fréchet mean from the multiple documents in the following sections.

3.2 Euclidean Metric space

Let's begin with the Euclidean metric space. We assume that terms observed in a document are samples from a multinomial distribution and each document has a distinct distribution. Assuming a conjugate Dirichlet prior, we estimate the multinomial distribution, i.e. a language model, using Dirichlet smoothing [35] as follows:

$$\Pr(w|D) = \frac{tf_{w,D} + \mu \cdot cf_w/|C|}{|D| + \mu} \tag{2}$$

where $tf_{w,D}$ is the occurrence of term w in document D, cf_w is the occurrence of w in a set of observations C considered for the prior distribution (typically, a corpus), $|D|$ is the number of observations, i.e. the length of D, $|C|$ is the length of C, and μ is the Dirichlet smoothing parameter. Note that $P(w|D)$ is a parameter which corresponds to outcome w in the multinomial distribution.

The size of vocabulary of a language model is defined as the number of terms observed in C, which also determines the number of dimensions of the Euclidean metric space for a multinomial distributions. When the number of dimensions is $n+1$, a multinomial distribution corresponds to a point in n-simplex $\mathcal{P}_n$ which is defined as follows:

$$\mathcal{P}_n = \left\{ \mathbf{x} \in \mathbb{R}^{n+1} : \forall i, x^{(i)} > 0, \sum_{i=1}^{n+1} x^{(i)} = 1 \right\}$$

An example of 2-simplex embedded in 3-dimensional Euclidean space is shown in Figure 1.

Since a geodesic linking two points in n-simplex is a straight line, the distance between two multinomial distributions is calculated by the Euclidean distance as follows:

$$dist(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n+1} (x^{(i)} - y^{(i)})^2}$$

Consider multinomial distributions of k given documents, $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_k$ as samples from distribution Q over the n-simplex. Then, the Fréchet sample function is given by

$$\bar{\Phi}(\mathbf{c}) = \sum_{i=1}^{k} \hat{Q}(\mathbf{p}_i) \sum_{j=1}^{n+1} (c^{(j)} - p_i^{(j)})^2$$

Therefore, we have the following optimization problem to obtain the Fréchet sample mean.

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{k} \hat{Q}(\mathbf{p}_i) \sum_{j=1}^{n+1} (c^{(j)} - p_i^{(j)})^2 \\ \text{subject to} \quad & \sum_{j=1}^{n+1} c^{(j)} = 1, \qquad \forall j, c^{(j)} > 0 \end{aligned} \tag{3}$$

It is trivial to solve this problem using the method of Lagrange multipliers. Finally, we have a solution as follows:

$$c^{(j)} = \sum_{i=1}^{k} p_i^{(j)} \hat{Q}(\mathbf{p}_i) \tag{4}$$

This is the Fréchet sample mean in the Euclidean metric space. Indeed, if $\hat{Q}(\mathbf{p}_i)$ is uniform, i.e, $1/k$, then this is the same as the ordinary Euclidean mean or the arithmetic mean. Therefore, the Fréchet sample mean in the Euclidean metric space generalizes the arithmetic mean.

We use the Fréchet sample mean as a representative multinomial distribution for the given group of multiple documents.

3.3 Riemannian manifold defined by the Fisher information metric

Many IR approaches assume that data is embedded in the Euclidean geometry. However, assumptions of non-Euclidean geometries may lead to a better understanding of data. We here consider a Riemannian space where a Riemannian metric is the Fisher information metric. This metric space is used for investigating the geometric structures of statistical models in most of the Information Geometry literature [28, 1, 16]. Furthermore, a number of approaches assume this metric space for statistical inference and machine learning [20, 22, 1]. Particularly, for text classification, Lafferty and Lebanon [20] showed that techniques based on this metric space perform better than techniques based on the Euclidean metric.

The Fisher information metric is defined as follows:

$$\begin{aligned} g_{i,j}(\boldsymbol{\theta}) &= \int \frac{\partial \log p(x;\boldsymbol{\theta})}{\partial \theta^{(i)}} \frac{\partial \log p(x;\boldsymbol{\theta})}{\partial \theta^{(j)}} p(x;\boldsymbol{\theta}) dx \\ &= E_{\boldsymbol{\theta}} \left[\frac{\partial \log p(x;\boldsymbol{\theta})}{\partial \theta^{(i)}} \frac{\partial \log p(x;\boldsymbol{\theta})}{\partial \theta^{(j)}} \right] \end{aligned}$$

where $\boldsymbol{\theta}$ is a point in a differential manifold and corresponds to a statistical model in a parametric famiilty $p(x;\boldsymbol{\theta})$, i and j are indices for a coordinate system. In this work, it is easy

to think that $\boldsymbol{\theta}$ is a multinomial model for a document while i and j are indices for unique terms in vocabulary.

This metric has some nice properties. By Cramér-Rao inequality [28], the variance of unbiased estimators is bounded by the inverse of the metric. Particularly, an unbiased estimator achieving the bound is called an efficient estimator which is the best unbiased estimator because it minimizes the variance. Furthermore, by Chentsov's theorem [8], the Fisher information metric is the only Riemannian metric which is invariant under basic probabilistic transformations.

We now look into the Riemannian geometry with the Fisher information metric as a Riemannian metric. First of all, let us consider the positive n-sphere of radius 2, $\tilde{\mathcal{S}}_n^+$ instead of n-simplex P_n.

$$\tilde{\mathcal{S}}_n^+ = \left\{ \mathbf{x} \in \mathbb{R}^{n+1} : \forall i, x^{(i)} > 0, \sum_{i=1}^{n+1} (x^{(i)})^2 = 2^2 \right\}$$

Figure 1 shows an example of the positive 2-sphere of radius 2.

We can define transformation $\phi : \mathcal{P}_n \to \tilde{\mathcal{S}}_n^+$ by

$$z^{(j)} = \phi(\mathbf{x})^{(j)} = 2\sqrt{x^{(j)}}$$

The inverse transformation ϕ^{-1} is well known to pull back the Fisher information metric on $\mathcal{P}_n$ to the Euclidean metric on $\tilde{\mathcal{S}}_n^+$ [16, 22]. Therefore, the transformation is an isometry, and we can compute the distance between two statistical models by the Fisher information metric using the geodesic distance between two corresponding points on the sphere. In other words, the distance is the length of the shortest curve linking two corresponding points on the sphere and is given by

$$dist(\mathbf{x}, \mathbf{y}) = 2 \arccos \left(\sum_{j=1}^{n+1} \sqrt{x^{(j)} y^{(j)}} \right)$$

This is called the information distance.

With this distance, we have the following Fréchet sample function.

$$\bar{\Phi}(\mathbf{c}) = 4 \sum_{i=1}^{k} \arccos^2 \left(\sum_{j=1}^{n+1} \sqrt{x^{(j)} y^{(j)}} \right) \hat{Q}(\mathbf{p}_i)$$

Unfortunately, there is no closed form solution for the Fréchet sample mean which minimizes this function. Although we can use some convex optimization techniques, such approaches may be impractical in case that n is large. Indeed, in many IR tasks, $n + 1$ is the size of vocabulary and can be very large.

Therefore, to find the Fréchet sample mean, we try an approximation approach using the Kullback-Leibler (KL) divergence which is defined as follows:

$$D(\mathbf{x}||\mathbf{y}) = \sum_{j=1}^{n+1} x^{(j)} \log \frac{x^{(j)}}{y^{(j)}}$$

As $y \to x$, approximately by the Taylor expansion,

$$\begin{aligned} &\log x^{(j)} - \log y^{(j)} \\ &= -\frac{(y^{(j)} - x^{(j)})}{x^{(j)}} + \frac{(y^{(j)} - x^{(j)})^2}{2(x^{(j)})^2} + O((y^{(j)} - x^{(j)})^3) \end{aligned}$$

From this,

$$\begin{aligned} &D(\mathbf{x}||\mathbf{y}) + D(\mathbf{y}||\mathbf{x}) \\ &= \sum_{j=1}^{n+1} \left[x^{(j)} \left(\log x^{(j)} - \log y^{(j)} \right) + y^{(j)} \left(\log y^{(j)} - \log x^{(j)} \right) \right] \\ &= \frac{1}{2} \sum_{j=1}^{n+1} \frac{(y^{(j)} - x^{(j)})^2}{x^{(j)}} + \frac{1}{2} \sum_{j=1}^{n+1} \frac{(x^{(j)} - y^{(j)})^2}{y^{(j)}} + O(||\mathbf{y} - \mathbf{x}||^3) \end{aligned} \tag{5}$$

Since $\mathbf{y}$ approaches $\mathbf{x}$ along geodesic c linking them, we can parameterize the path by arclength s so that $c(s_0) = \mathbf{x}$, $c(s_1) = \mathbf{y}$ and $s_1 - s_0 = dist(\mathbf{x}, \mathbf{y})$. The difference between two points is expressed by a product of the geodesic length and the tangent vector to the curve as follows:

$$y^{(j)} - x^{(j)} = (s_1 - s_0) \frac{\partial c^{(j)}}{\partial s} = dist(\mathbf{x}, \mathbf{y}) \frac{\partial c^{(j)}}{\partial s}$$

Then, the first term in Equation (5) can be rewritten as follows:

$$\begin{aligned} &\frac{1}{2} \sum_{j=1}^{n+1} \frac{1}{x^{(j)}} \left(dist(\mathbf{x}, \mathbf{y}) \frac{\partial c^{(j)}}{\partial s} \right)^2 = \frac{1}{2} dist^2(\mathbf{x}, \mathbf{y}) \sum_{j=1}^{n+1} \frac{1}{c^{(j)}(s)} \left(\frac{\partial c^{(j)}}{\partial s} \right)^2 \\ &= \frac{1}{2} dist^2(\mathbf{x}, \mathbf{y}) \sum_{j=1}^{n+1} c^{(j)}(s) \left(\frac{\partial \log c^{(j)}}{\partial s} \right)^2 = \frac{1}{2} dist^2(\mathbf{x}, \mathbf{y}) I(s) \end{aligned}$$

where $I(s)$ is the Fisher information for s. By definition of the length of the curve,

$$\int_{s_0}^{s_1} I(s) ds = dist(\mathbf{x}, \mathbf{y}) = s_1 - s_0$$

Hence, $I(s) = 1$, and we finally have the following:

$$\frac{1}{2} \sum_{j=1}^{n+1} \frac{(y^{(j)} - x^{(j)})^2}{x^{(j)}} = \frac{1}{2} dist^2(\mathbf{x}, \mathbf{y}) \tag{6}$$

Similarly, the second term in Equation (5) can be also written as Equation (6). Therefore, we have an approximation of Equation (5) as follows:

$$\begin{aligned} D(\mathbf{x}||\mathbf{y}) + D(\mathbf{y}||\mathbf{x}) &= dist^2(\mathbf{x}, \mathbf{y}) + O(||\mathbf{y} - \mathbf{x}||^3) \\ &\approx dist^2(\mathbf{x}, \mathbf{y}) \end{aligned}$$

Similar relationships between divergences and distances can be founded in various texts [1, 16].

From this approximation, we can express the Fréchet sample mean with the KL divergence as follows:

$$\bar{\Phi}(\mathbf{c}) \approx \sum_{i=1}^{k} \left(D(\mathbf{p}_i||\mathbf{c}) + D(\mathbf{c}||\mathbf{p}_i) \right) \hat{Q}(\mathbf{p}_i) \tag{7}$$

This means that finding the Fréchet sample mean is reduced to finding the symmetrized Bregman centroid $\mathbf{c}^F$ [27] which is defined as follows:

$$\mathbf{c}^F = \arg\min_c \sum_{i=1}^{k} \frac{1}{2} \left(D_F(\mathbf{p}_i||\mathbf{c}) + D_F(\mathbf{c}||\mathbf{p}_i) \right) \hat{Q}(\mathbf{p}_i)$$

where $D_F(\mathbf{x}||\mathbf{y})$ is the Bregman divergence defined by $F(\mathbf{x}) - F(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla F(\mathbf{y}) \rangle$ and F is a generator function. For example, if F is the negative Shannon entropy, i.e. $\sum_j x^{(j)} \log x^{(j)}$,

then the Bregman divergence is the same as the KL divergence. That is, the Bregman divergence is a generalized divergence. In addition, right-sided centroid $\mathbf{c}_R^F$ and left-sided centroid $\mathbf{c}_L^F$ are defined as follows:

$$\mathbf{c}_R^F = \arg\min_c \sum_{i=1}^{k} D_F(\mathbf{p}_i||\mathbf{c})\hat{Q}(\mathbf{p}_i)$$

$$\mathbf{c}_L^F = \arg\min_c \sum_{i=1}^{k} D_F(\mathbf{c}||\mathbf{p}_i)\hat{Q}(\mathbf{p}_i)$$

Nielsen and Nock [27] show that symmetrized Bregman centroid $\mathbf{c}^F$ lies on a geodesic linking $\mathbf{c}_R^F$ and $\mathbf{c}_L^F$ via the Bregman Pythagoras' theorem. We can apply the result to the KL divergence.

We can easily compute $\mathbf{c}_R^F$ using the method of Lagrange multipliers with the same constraints as Equation (3), and the solution coincides with the arithmetic mean as follows:

$$c_R^{F(j)} = \sum_{i=1}^{k} \hat{Q}(\mathbf{p}_i)p_i^{(j)}$$

Similarly, using the method of Lagrange multipliers, we compute $\mathbf{c}_L^F$ as follows:

$$c_L^{F(j)} = \prod_{i=1}^{k} \left(p_i^{(j)^{\hat{Q}(\mathbf{p}_i)}} \right) / \sum_{j=1}^{n+1} \prod_{i=1}^{k} \left(p_i^{(j)^{\hat{Q}(\mathbf{p}_i)}} \right)$$

If $\hat{Q} = 1/k$, then this is the ordinary normalized geometric mean.

Therefore, the symmetrized Bregman centroid when F is the negative Shannon entropy, or the approximated Fréchet sample mean lies on the geodesic linking the arithmetic mean and the normalized geometric mean.

We consider the two means as approximations to the Fréchet sample mean and take the following approach to decide a representation among them:

1. Compute the arithmetic mean $\mathbf{c}_A$ and the normalized geometric mean $\mathbf{c}_G$ from multinomial models of multiple documents.
2. Compute $\bar{\Phi}(\mathbf{c}_A)$ and $\bar{\Phi}(\mathbf{c}_G)$ by Equation (1)
3. As a representation, choose $\mathbf{c}_G$ if $\bar{\Phi}(\mathbf{c}_A) > \bar{\Phi}(\mathbf{c}_G)$, $\mathbf{c}_A$ otherwise.

That is, we choose a point which is closer to the Fréchet sample mean as a representation. We call this approach "geometric selection".

4. EXPERIMENTS

To evaluate representation techniques derived in the previous section, we conduct experiments for two different tasks: cluster retrieval and pseudo-relevance feedback.

For the experiments, we use 3 standard collections from TREC. Table 1 shows the statistics of the collections. To estimate a language model from each document, we use the Dirichlet smoothing. For each task, the initial results are obtained by query-likelihood scores which are computed under an independence assumption as follows:

$$P(Q|D) = \prod_{q \in Q} P(q|D)$$

where $P(q|D)$ is estimated by Equation (2).

	AP	WSJ	GOV2
TREC topics	51-200	51-200	701-800
#docs	242,918	173,252	25,205,179

Table 1: Test collections.

For index building, we used the Indri system [33]. Each document was stemmed by the Krovetz stemmer and stopped by a standard stopword set. To test the significance of results, we performed a randomization test.

4.1 Cluster Retrieval

Cluster retrieval involves finding the best document cluster [24, 26]. We first retrieve the top 100 documents for each query according to query-likelihood scores. Next, we perform kNN clustering [19]. That is, assuming that each returned document is a cluster centroid, a cluster is formed by its $k-1$ nearest neighbors (k is set to 5). We use cosine similarity as a similarity measure. In fact, since cosine similarity assumes the Euclidean metric space, other similarity measures may perform better for our representation technique which assumes a different metric. However, since arbitrary clusters are assumed in cluster retrieval, we use the same similarity measure as used in previous work [26].

Once we have clusters, we represent each cluster by the arithmetic mean of language models of documents in a cluster assuming the Euclidean metric. On the other hand, assuming the Fisher information metric, we can determine a representation via geometric selection between the arithmetic mean and the normalized geometric mean of the documents.

Evaluation of various representation techniques such as concatenation or CombMax [12] for cluster retrieval has been already done by Liu and Croft [26]. They concluded that the geometric mean representation outperforms other techniques. Therefore, we do not intend to repeat the same work. Instead, we focus on geometric interpretations for experimental results.

For a fair comparison, the same clusters are given to each representation technique. The only parameter to be tuned is the smoothing parameter for the initial results. We set the parameter so that Mean Average Precision (MAP) for the initial results by the query-likelihood $P(Q|D)$ is maximized. Evaluation is performed using all topics. Since our goal is to find the best cluster, we use Precision at 5 (P@5) in order to evaluate the cluster first ranked by each representation technique, i.e. how many relevant documents the cluster has. Table 2 shows the results. In addition to the arithmetic mean and geometric selection, we present results using the geometric mean as well.

For all collections, representations by the geometric mean and geometric selection show better performance than representations by the arithmetic mean. Except for GOV2, The improvements are statistically significant. These experiments indicate some interesting points. First, in geometric selection, the normalized geometric means were selected as representations which minimize the Fréchet sample function for all queries across all collections. In other words, the normalized geometric means are better approximations to the Fréchet sample mean. Second, since the normalized geometric means selected by geometric selection lead to consistently better retrieval results, we may say that the goodness of a representation for this task is related to how close the rep-

	AP	WSJ	GOV2
A-MEAN	0.3053	0.4747	0.5374
G-MEAN	0.3347*	0.5040*	0.5576
SELECT	0.3347*	0.5027*	0.5556

Table 2: Results for cluster retrieval. A-MEAN, G-MEAN and SELECT mean representations by the arithmetic mean, by the geometric mean, and by geometric selection, respectively. The numbers are P@5 scores. A * indicates a statistically significant improvement over A-MEAN ($p < 0.05$).

resentation is to the center of mass, i.e. the Fréchet sample mean. Moreover, this justifies the assumption of the geometry defined by the Fisher information metric. Lastly, since geometric selection does not consider the geometric mean but the normalized geometric mean, the results in the 'SELECT' row are exactly the same as those by the normalized geometric means. Therefore, the differences between the 'G-MEAN' row and the 'SELECT' row are caused by the normalization. As you see, since the differences are small, we suggest that the geometric mean without normalization can be a better choice in practice.

4.2 Pseudo-Relevance Feedback

Lavrenko and Croft's relevance model [21] is one of the standard language modeling approaches for pseudo-relevance feedback. The model assumes that the top k retrieved documents for query q are sampled from an underlying relevance model for q. That is, a hidden multinomial model relevant to a user information need exists, and we estimate the model from the top k documents. Then, we sample terms which describe the information need better than the original query and use the terms for query expansion.

Estimation of the relevance model is done by the following formula:

$$P(w|q) = \frac{\sum_{i=1}^{k} p(w|D_i)P(q|D_i)P(D_i)}{p(q)} \quad (8)$$

where q is a user query, w is a candidate for expansion terms, and D_i is a document in the top k initial results, respectively.

Although this is derived from a Bayesian model, we can see this as a representation for the top k documents by the arithmetic mean rewriting Equation (8) as follows:

$$\sum_{i=1}^{k} p(w|D_i)\frac{P(q|D_i)P(D_i)}{p(q)} = \sum_{i=1}^{k} p(w|D_i)P(D_i|q)$$

This has the same form as the weighted arithmetic mean of Equation (4). In other words, $P(w|D_i)$ is a multinomial parameter and $P(D_i|q)$ represents a distribution over a sample space limited by q, i.e, $\hat{Q}$. In the standard implementation of the relevance model by the Indri system [33], $P(D)$ is assumed to be uniform. Hence,

$$P(D_i|q) = \frac{P(q|D_i)P(D)}{\sum_{i=1}^{k} P(q|D_i)P(D)} = \frac{P(q|D_i)}{\sum_{i=1}^{k} P(q|D_i)}$$

That is, the weight $\hat{Q} = P(D_i|q)$ is the normalized query-likelihood scores obtained in the initial retrieval phase. Therefore, we can say that the relevance model represents a group of the top k documents combining the language models by the arithmetic mean weighted by the initial search results.

	AP	WSJ	GOV2
RM	0.2541	0.3531	0.3204
GRM	0.2769*	0.3851*	0.3300*

Table 3: Results for pseudo-relevance feedback. RM and GRM mean the relevance model and the geometric relevance model, respectively. The numbers are MAP scores. A * indicates a statistically significant improvement over RM ($p < 0.01$).

In this sense, we can say that the relevance model implicitly assumes the Euclidean metric space.

We can replace the arithmetic mean by the normalized geometric mean to develop a new representation as follows:

$$P(w|q) = \prod_{i=1}^{k} p(w|D_i)^{P(D_i|q)} / \sum_{w \in V} \prod_{i=1}^{k} p(w|D_i)^{P(D_i|q)} \quad (9)$$

We can consider the original relevance model and this model as two approximated representations in the Riemannian manifold defined by the Fisher information metric. To determine a representation, we use geometric selection and call the selected model the "geometric relevance model".

We compare the geometric relevance model with the relevance model. For each query, we first retrieve the top k documents by query-likelihood scores and build a relevance model or geometric relevance model for the documents. Then, we choose the top M terms according to probabilities of the terms in the models. Finally, we expand the original query combining the expansion terms using an interpolation weight λ in the Indri query language. The paremeters k, M and λ are tuned so that MAP scores by the relevance model are maximized. The same parameters are used for the geometric relevance model. Topic 51-150 for AP and WSJ and topic 701-750 for GOV2 are used as training topics to learn the parameters. Topic 151-200 for AP and WSJ and topic 751-800 for GOV2 are used as test topics. We retrieve up to 1000 results for each expanded query and use MAP as the evaluation metric.

Table 3 shows the results. The geometric relevance model significantly outperforms the relevance model for all three collections. Similar to cluster retrieval, geometric selection selected models by Equation (9) rather than the original relevance model as representations for all queries except for three queries of GOV2. That is, the geometric mean is a better approximation to the center of mass for this task. This provides more empirical evidence that the geometric mean can be an appropriate choice for representation.

5. DISCUSSIONS

5.1 Visualization of geometries

To show how multiple documents, the arithmetic mean and the normalized geometric mean are distributed in each geometry, we use the following visualization. First, we construct a weighted complete graph, where each node is a document or the mean and a weight is determined by a kernel reflecting each geometry.

For the Euclidean metric, we use the following heat kernel:

$$K(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(\left(-\sum_{j=1}^{n+1}\left(x_1^{(j)} - x_2^{(j)}\right)^2\right)/4t\right)$$

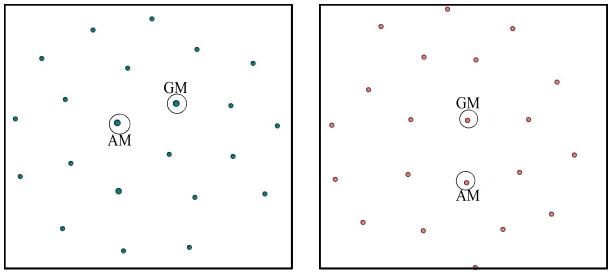

Figure 2: Geometric visualization of the top 20 documents for Topic 770 (GOV2), the arithmetic mean (AM) and the normalized geometric mean (GM) for different metrics, i.e. the Euclidean metric (left) and the Fisher information metric (right).

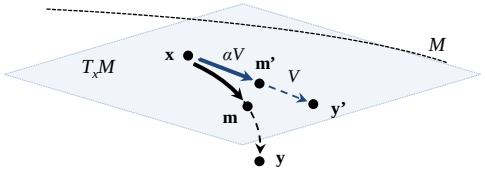

Figure 3: Determinination of a middle point m on a geodesic linking x and y

where t is a time parameter.

For the Fisher information metric, we use the following information diffusion kernel [20]:

$$K(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\arccos^2\left(\sum_{j=1}^{n+1}\sqrt{x_1^{(j)}x_2^{(j)}}\right)/4t\right)$$

We visualize each geometry using CCVisu [4] which is a tool implementing energy models so that the higher weight between two points results in the smaller Euclidean distance between them. A visualization example is shown in Figure 2. As you see, the arithmetic mean appears closer to the center in the Euclidean metric space while the normalized geometric mean appears closer in the Riemannian manifold defined by the Fisher information metric. Since the visualization tool uses random seeds to initialize the layout, the results vary every time. However, the trend for the locations of the means was consistent.

5.2 More accurate estimation

Geometric selection is a somewhat simple approach to determine the approximated Fréchet sample mean. That is, we choose one among only two options: the normalized geometric mean and the arithmetic mean. We now consider a more accurate estimation technique for the Fréchet sample mean.

A point which minimizes the approximated Fréchet sample function of Equation (7) lies on a geodesic linking the arithmetic mean and the normalized geometric mean. Let M, $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{c}$ be the statistical manifold defined by the Fisher information metric, the arithmetic mean, the normalized geometric mean and a geodesic linking the two points, respectively. First, we get vector V on tangent space $T_{\mathbf{x}}M$ via log map $\log_{\mathbf{x}} : M \rightarrow T_{\mathbf{x}}M$. In case of a sphere, the log map is given by:

$$V^{(j)} = \log_{\mathbf{x}}(\mathbf{y})^{(j)} = \frac{\arccos(\langle \mathbf{x}, \mathbf{y}\rangle)}{\sqrt{1-\langle \mathbf{x}, \mathbf{y}\rangle^2}}\left(y^{(j)} - \langle \mathbf{x}, \mathbf{y}\rangle x^{(j)}\right)$$

Then, V links $\mathbf{x}$ to $\mathbf{y}'$ on $T_{\mathbf{x}}M$ corresponding to $\mathbf{y}$ on M.

$\mathbf{m}'$ denotes a middle point between $\mathbf{x}$ and $\mathbf{y}'$ on $T_{\mathbf{x}}M$, reached by αV $(0 \leq \alpha \leq 1)$. We now get a middle point $\mathbf{m}$ on $\mathbf{c}$ via exponential map $\exp_{\mathbf{x}} : T_{\mathbf{x}}M \rightarrow M$. The exponential map of a sphare is:

$$m^{(j)} = \exp_{\mathbf{x}}(\alpha V)^{(j)} = \cos(\alpha||V||) + \frac{\sin(\alpha||V||)}{||V||}V^{(j)}$$

Figure 3 illustrates this procedure. Note that the arithmetic mean $\mathbf{x}$ and the geometric mean $\mathbf{y}$ are interchangeable in the above formulation because a sphere is symmetric.

We apply this result to pseudo-relevance feedback experiments. We perform grid search on the geodesic varying α in [0,1] by step-size 0.1, and a point which minimizes the Fréchet sample function of Equation (1) is selected as a representation. Figure 4 shows α's selected for test queries for each collection. For all test topics except for three topics of GOV2, the selected α's are equal to or greater than 0.5. That is, the more accurately estimated Fréchet sample means are also closer to the normalized geometric mean than the arithmetic mean. Table 4 shows the results when the representations are used for pseudo-relevance feedback. All results are equal to or a little bit better than the results of the GRM in the Table 3, but not significantly. Therefore, we can say that the geometric relevance model is a reasonable approximation to the Fréchet sample mean for this task.

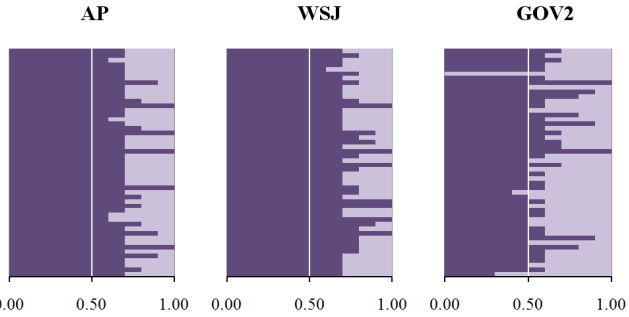

Figure 4: Relative locations of the more accurately estimated Fréchet sample means. The x-axis corresponds to the relative locations, and the y-axis corresponds to queries for each collection. As a relative location is closer to 1.0, the estimated mean for the topic is located near the normalized geometric mean.

	AP	WSJ	GOV2
GRM^+	0.2769	0.3852	0.3309

Table 4: Pseudo-relevance feedback results of the more accurately estimated Fréchet sample mean in the Riemannian manifold defined by the Fisher information metric.

5.3 Anoher reason for the geometric mean

We have addressed so far theoretical and empirical reasons explaining why the geometric mean should have advantages

for many IR tasks. There can be many other explanations. One of them is the log-linearity of the geometric mean. As more documents contain a specific term, the geometric mean for the term increases exponentially while the arithmetic mean increases linearly. Accordingly, the arithmetic mean can be sensitive to a few dominant terms in a small number of documents. On the other hand, the geometric mean favors the common terms across a whole set of documents and is relatively insensitive to such a few dominant terms. This shows the robustness of the geometric mean which can lead to a good representation for multiple documents.

6. CONCLUSIONS

Previous work which uses the geometric mean as a representation technique does not provide enough theoretical evidence explaining why the geometric mean should have advantages as a representation for IR. There are various explanations. In this work, we showed that using Information Geometry, the arithmetic mean and the normalized geometric mean are approximation points to the center of mass in the Euclidean space or in a statistical manifold. In particular, through empirical evidence, we demonstrated that the normalized geometric mean is closer to the center in the statistical manifold. In addition to this discovery, we introduced a new approach to pseudo-relevance feedback that outperformed the relevance model. For future work, we will investigate how geometric interpretations can be applied to other IR tasks. We expect that this effort will lead to not only the discovery of novel IR theories but also development of effective algorithms.

7. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant #IIS-0711348, and in part by NSF grant #IIS-0534383. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2000.

[2] N. J. Belkin, C. Cool, W. B. Croft, and J. P. Callan. The effect multiple query representations on information retrieval system performance. In *SIGIR '93*, 1993.

[3] M. Bendersky and O. Kurland. Utilizing passage-based language models for document retrieval. In *ECIR '08*, 2008.

[4] D. Beyer. CCVisu: Automatic visual software decomposition. In *Proc. Int'l Conf. on Software Engineering*, 2008.

[5] R. Bhattacharya and V. Patrangenaru. Nonparametic estimation of location and dispersion on riemannian manifolds. *Journal of Statistical Planning and Inference*, 108, 2002.

[6] J. Callan. Distributed information retrieval. In W. B. Croft, editor, *Advances in Information Retrieval*. Kluwer Academic Publishers, 2000.

[7] J. P. Callan. Passage-level evidence in document retrieval. In *SIGIR '94*, 1994.

[8] N. N. Chentsov. *Statistical Decision Rules and Optimal Inference*. American Mathematical Society, 1982.

[9] K. Collins-Thompson and J. Callan. Estimation and use of uncertainty in pseudo-relevance feedback. In *SIGIR '07*, 2007.

[10] B. Efron. Defining the curvature of a statistical problem. *The Annals of Statistics*, 3(6).

[11] J. L. Elsas and J. G. Carbonell. It pays to be picky: an evaluation of thread retrieval in online forums. In *SIGIR '09*, 2009.

[12] E. A. Fox and J. A. Shaw. Combination of multiple searches. In *TREC-2*, 1994.

[13] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré*, 10, 1948.

[14] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007), 1946.

[15] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5), 1977.

[16] R. E. Kass and P. W. Vos. *Geometrical Foundations of Asymptotic Inference*. Wiley-Interscience, 1997.

[17] W. Kendall. Probability, convexity, and harmonic maps with small image i: Uniqueness and fine existence. *Proc. London Math. Soc.*, 61, 1990.

[18] J. Kogan, M. Teboulle, and C. Nicholas. The entropic geometric means algorithm: An approach for building small clusters for large text datasets. In *the Workshop on Clustering Large Data Sets*, 2003.

[19] O. Kurland and L. Lee. Corpus structure, language models, and ad hoc information retrieval. In *SIGIR '04*, 2004.

[20] J. Lafferty and G. Lebanon. Diffusion kernels on statistical manifolds. *The Journal of Machine Learning Research*, 6, 2005.

[21] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR' 01*, 2001.

[22] G. Lebanon. *Riemannian Geometry and Statistical Machine Learning*. PhD thesis, 2005.

[23] J. H. Lee. Analyses of multiple evidence combination. In *SIGIR '97*, 1997.

[24] A. Leuski. Evaluating document clustering for interactive information retrieval. In *CIKM '01*, 2001.

[25] X. Liu and W. B. Croft. Passage retrieval based on language models. In *CIKM '02*, 2002.

[26] X. Liu and W. B. Croft. Evaluating text representations for retrieval of the best group of documents. In *ECIR '08*, 2008.

[27] F. Nielsen and R. Nock. Sided and symmetrized Bregman centroids. *IEEE Transactions on Information Theory*, 55(6), 2009.

[28] C. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37, 1945.

[29] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, 1971.

[30] J. Seo and W. B. Croft. Blog site search using resource selection. In *CIKM '08*, 2008.

[31] J. Seo, W. B. Croft, and D. A. Smith. Online community search using thread structure. In *CIKM '09*, 2009.

[32] L. Si and J. Callan. Unified utility maximization framework for resource selection. In *CIKM '04*, 2004.

[33] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proc. of the Intl. Conf. on Intelligence Analysis*, 2005.

[34] R. Veldhuis. The centroid of the symmetrical Kullback-Leibler distance. *IEEE Signal Processing Letters*, 9(3), 2002.

[35] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01*, 2001.