Query Expansion with Random Embeddings

Lucas Bernardi

January 2014

Query Expansion: Problem statement

- In the context of Information Retrieval, Query Expansion is a method to improve recall and precision by modifying a user query
- A specific case is the expansion of each term to *related* terms, which mainly improves recall but also precision as a side effect
- Example: User query: `soccer`, expanded query: `soccer football maradona`
- The Problem: **Learn to expand terms**

A General Approach

The distributional hypothesis

Words with similar distributional properties have similar meanings.

The geometric metaphor of meaning

Meanings are locations in a semantic space, and semantic similarity is proximity between the locations.

The Vector Space Model Approach

A solution

- An algebraic model for information retrieval and NLP
- Defines a vector space to represent terms
- Defines a dimension for each unique term
- Each term is represented by a semantic vector
- Each component of a semantic vector is the weight of the represented term in the direction of the dimension term
- Weights are a design decision, TF-IDF is widely used
- Compute the related terms of a given term as its k-nearest neighbors
- The vector distance/similarity measure is a design decision, cosine similarity is widely adopted

	play	soccer	week	favorite	sport	forget	ball	Football
play	0	1	1	0	0	0	0	1
soccer	1	0	1	1	1	1	1	0
week	1	1	0	0	0	0	0	0
favorite	0	1	0	0	1	0	0	0
sport	1	1	0	1	0	0	0	1
forget	0	1	0	0	0	0	1	0
ball	0	1	0	0	0	1	0	0
Football	1	0	0	0	1	0	0	0

$$s(soccer) = \begin{pmatrix} 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}$$

$$s(Football) = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$$similarity(soccer, Football) = cos(s(soccer), s(football)) = \frac{1}{\sqrt{3}} = 0{,}57$$

The Vector Space Model Approach

Some drawbacks

- Curse of dimensionality
- Sparse vectors
- Dynamic vector space

Alternatives

- Use documents as dimensions: Less dimensions, but still sparse vectors and still a dynamic space
- Use topics as dimensions: Less dimensions, dense vectors and fixed space, but how can we define topics and assign weights? Latent Semantic Analysis

The Random Projections Approach

Overcoming drawbacks

- Use a random low dimensional space
- Less dimensions, dense vectors, fixed space
- No semantic assigned to dimensions
- But, how are weights assigned?

The Random Projections Approach

A simple algorithm

- Assign a random k-dimensional vector to each term (index vector)
- Allocate a null k-dimensional vector to each term (semantic vector)
- For each sentence in the corpus compute bigrams (*left right*)
- For each bigram
  - $semantic(left) \mathrel{+}= index(right)$
  - $semantic(right) \mathrel{+}= index(left)$

The Random Projections Approach

Illustration

Document 1:
We play soccer every week. Soccer is my favorite sport. Don't forget the soccer ball

Context 1: play soccer week.
Context 2: Soccer favorite sport.
Context 3: forget soccer ball.

Document 2:
Football is a popular sport. Do you play football?

Context 1: Football popular sport
Context 2: play football

$s(\texttt{soccer}) =$
$r(\texttt{play}) + r(\texttt{week}) + r(\texttt{favorite}) + r(\texttt{sport}) + r(\texttt{forget}) + r(\texttt{ball}) =$
$\langle 10001 \rangle + \langle 10100 \rangle + \langle 00011 \rangle + \langle 00110 \rangle + \langle 11000 \rangle + \langle 10010 \rangle = \langle 41232 \rangle$

$s(\texttt{football}) =$
$r(\texttt{popular}) + r(\texttt{sport}) + r(\texttt{play}) = \langle 10010 \rangle + \langle 00110 \rangle + \langle 10001 \rangle = \langle 20121 \rangle$

$cos(s(\texttt{soccer}), s(\texttt{football})) = cos(\langle 41232 \rangle, \langle 20121 \rangle) = 0{,}97$

The Random Projections Approach

A simple algorithm: a formal description

Given a text documents corpus D we define
$\mathbb{T} = \{\text{All terms in } D\}$, $t = |\mathbb{T}|$
$\mathbb{C} = \{\text{All contexts in } D\}$, where a context is a set of terms
t random k-dimensional vectors $r_i, 1 \le i \le t$, $\mathbb{R} = \{r_i\}$
t semantic k-dimensional vectors $s_i, 1 \le i \le t$, $\mathbb{S} = \{s_i\}$
A mapping $S : w \in \mathbb{T} \to \mathbb{S}$, which maps a term to its semantic vector
A mapping $R : w \in \mathbb{T} \to \mathbb{R}$, which maps a term to its random vector
A mapping $C : w \in \mathbb{T} \to \{\mathbb{C}\}$, which maps a term to a set containing all the contexts it appears
Then we can express the semantic vector assigned to the term w as:

$$s_w = \sum_{c \in C(w)} \sum_{p \in c} R(p) \tag{1}$$

The Random Projections Approach

Algebraic Interpretation

Defining a co-occurrence matrix $M \in \mathbb{Z}^{t \times t}$, $M_{ij} = \sum_{c \in C(t_i)} \mathbb{1}_c(t_j)$

Each cell i, j in M counts the amount of contexts in D containing the ith and jth terms

Introducing a random matrix $R \in \mathbb{Z}^{t \times k}$, $R_i = R(t_i)$, that is, all random vectors as rows, we can express a semantic vector as:

$$s_i = \sum_{j=1}^{t} R(t_j) M_{ij} = \sum_{j=1}^{t} R_j M_{ij} \tag{2}$$

If we also define the semantic matrix $S \in \mathbb{Z}^{t \times k}$ with $S_i = s_i$, that is, all semantic vectors as rows, we can finally write:

$$S = MR \tag{3}$$

The Random Projections Approach

Algebraic Interpretation

$$S = MR \tag{4}$$

- This equation is a mathematical expression of our initial algorithm
- Semantic vectors are just the matrix product of the VSM vectors and a random matrix
- The algorithm is simply reducing the dimensionality of the VSM vectors
- For certain distributions of R, R is a nearly orthogonal
- Then we can interpret this product as the projection of the VSM vectors onto a random lower dimensional space

The Random Projections Approach

Theoretical support

The Johnson and Lindenstrauss Lemma

Given $\epsilon > 0$ and an integer n, let k be a positive integer such that $k \geq k_0 = O\big(\epsilon^{-2} \log n\big)$. For every set $\mathbb{P}$ of n points in $\mathbb{R}^d$ there exists $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $u, v \in \mathbb{P}$

$$(1 - \epsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \epsilon)\|u - v\|^2$$

The Random Projections Approach

Theoretical support

Achlioptas Theorem

Let $\mathbb{P}$ be an arbitrary set of n points in R^d represented as an $n \times d$ matrix A. Given $\epsilon, \beta > 0$ let $k_0 = \frac{4+2\beta}{\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}} \log n$. For integer $k \geq_0$, let $\mathbb{R}$ be a $d \times k$ random matrix with $R(i,j) = r_{ij}$, where $\{r_{ij}\}$ are independent random variables from either one of the following two probability distributions:

$$r_{i,j} = \begin{cases} +1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases} \quad r_{i,j} = \sqrt{3} \times \begin{cases} +1 & \text{with probability } \frac{1}{6} \\ 0 & \text{with probability } \frac{2}{3} \\ -1 & \text{with probability } \frac{1}{6} \end{cases}$$

Let $E = \frac{1}{\sqrt{k}} AR$. Let $f : \mathbb{R}^d \to \mathbb{R}^k$ map the i^{th} row of A to the i^{th} row of E. With probability at least $1 - n^{-\beta}, forall u, v \in \mathbb{P}$

$$(1-\epsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1+\epsilon)\|u - v\|^2$$

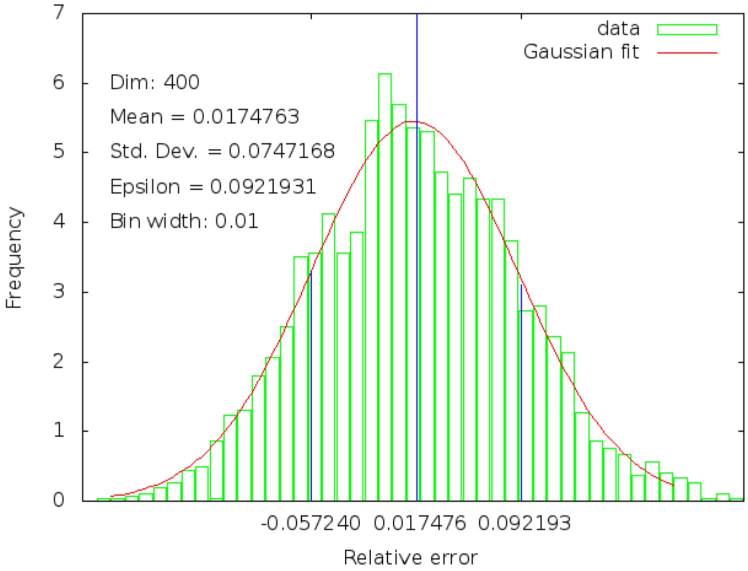
Empirical Study

Set up

- Natural language corpus with 140000 unique terms
- Filtered out low frequency terms, $n = 42905$
- Plain VSM vectors (original space is represented by the co-occurrence matrix)
- Evaluated random space dimension (k_0) ranging from 100 to 2000
- Random space generated with the dense Achlioptas distribution
- Pick two random vectors u, v, compute squared euclidean distance $\|u - v\|^2$ and $\|f(u) - f(v)\|^2$
- Compute relative error as $\frac{\|f(u)-f(v)\|^2-\|u-v\|^2}{\|u-v\|^2}$
- Sample size: 3000

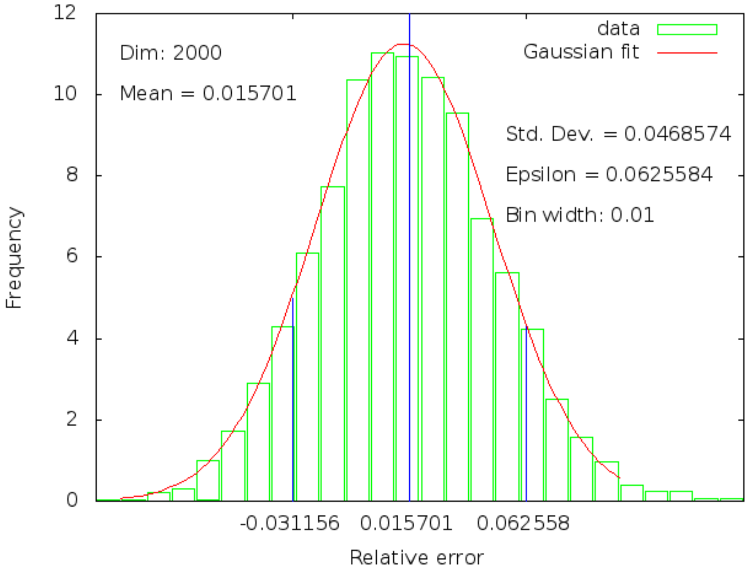
Empirical Study

Results



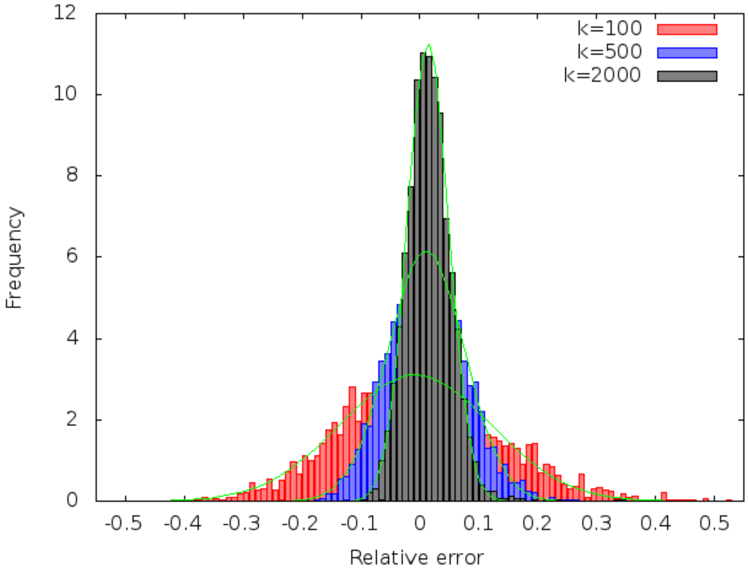
Empirical Study

Results



Empirical Study

Results



Empirical Study

Measuring ϵ

How can we compute ϵ from these histograms?

$$(1-\epsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\epsilon)\|u-v\|^2$$

$$1-\epsilon \leq \frac{\|f(u)-f(v)\|^2}{\|u-v\|^2} \leq 1+\epsilon$$

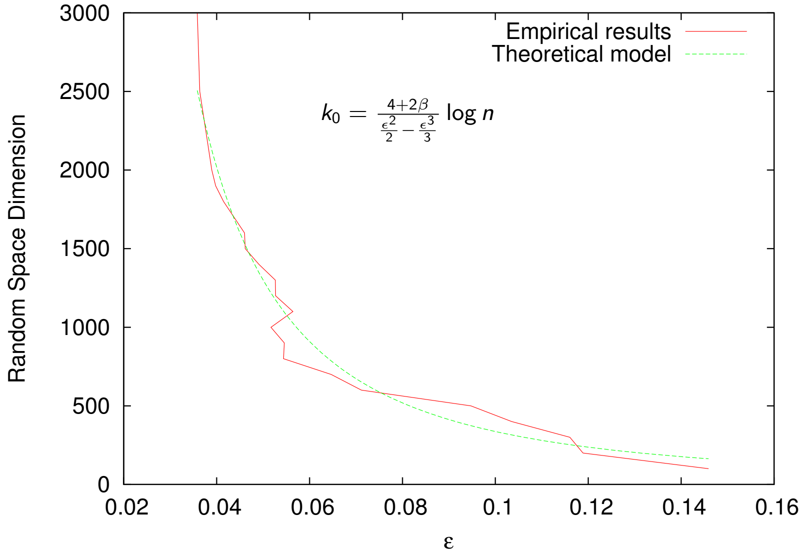
$$-\epsilon \leq \frac{\|f(u)-f(v)\|^2}{\|u-v\|^2} - 1 \leq \epsilon$$

$$-\epsilon \leq \frac{\|f(u)-f(v)\|^2 - \|u-v\|^2}{\|u-v\|^2} \leq \epsilon$$

$$\epsilon = \text{máx}(|\mu-\sigma|, |\mu+\sigma|)$$

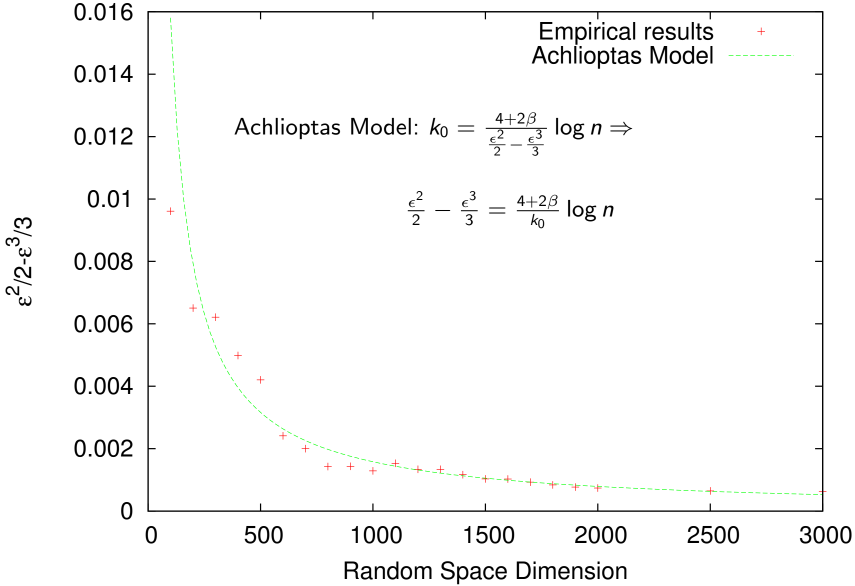
Empirical Study

Results



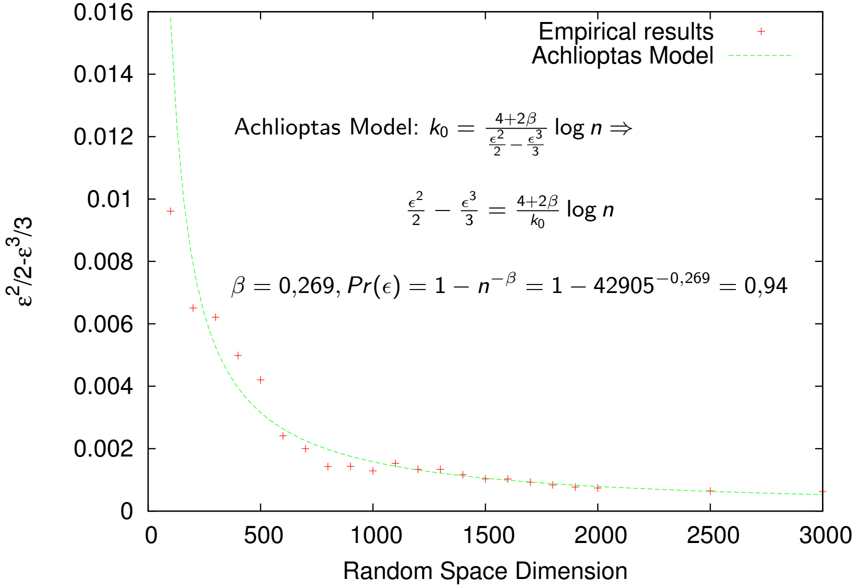
Empirical Study

Computing β



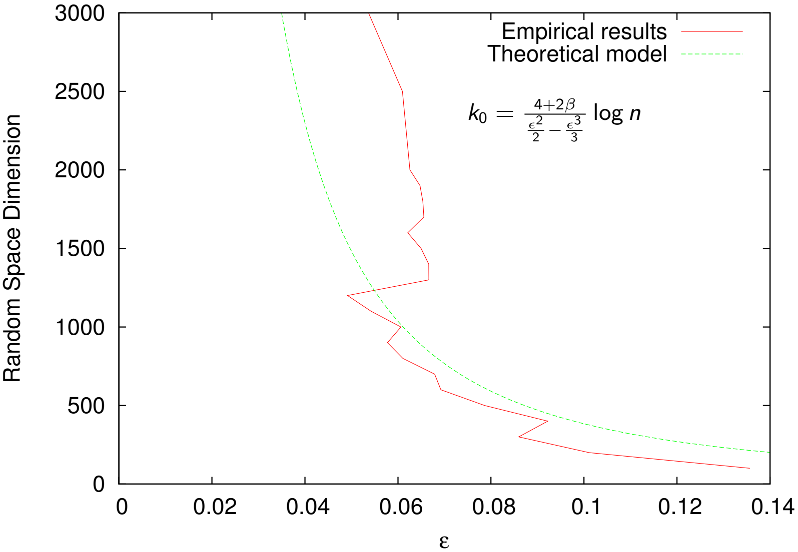
Empirical Study

Computing β



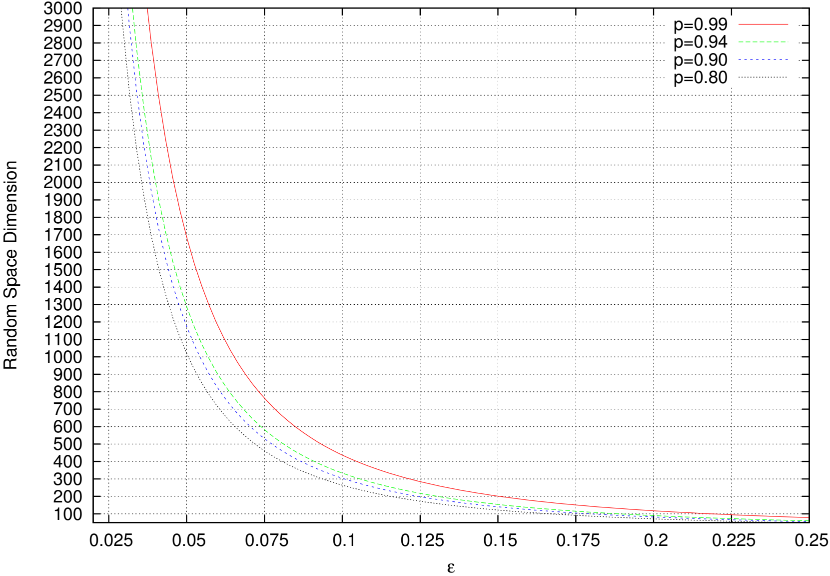
Empirical Study

Sparse projections



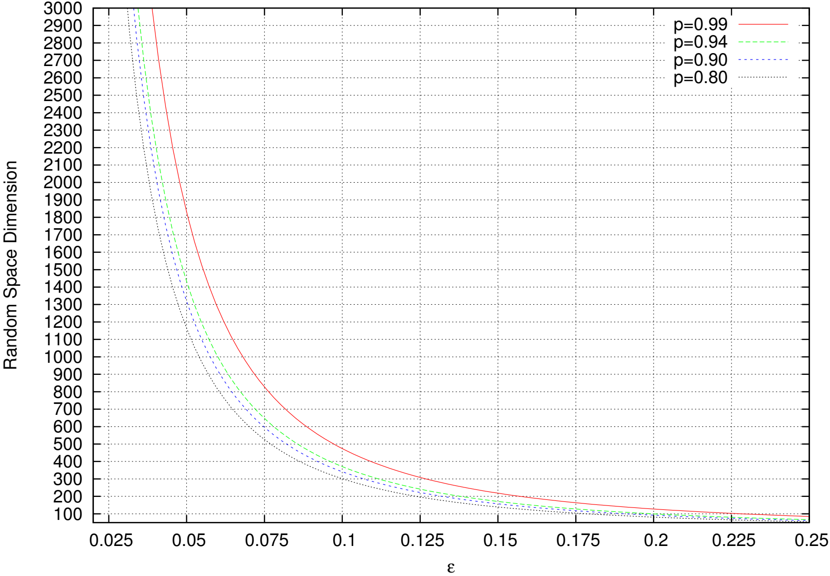
Trade offs

Dimensionality vs error vs certainty



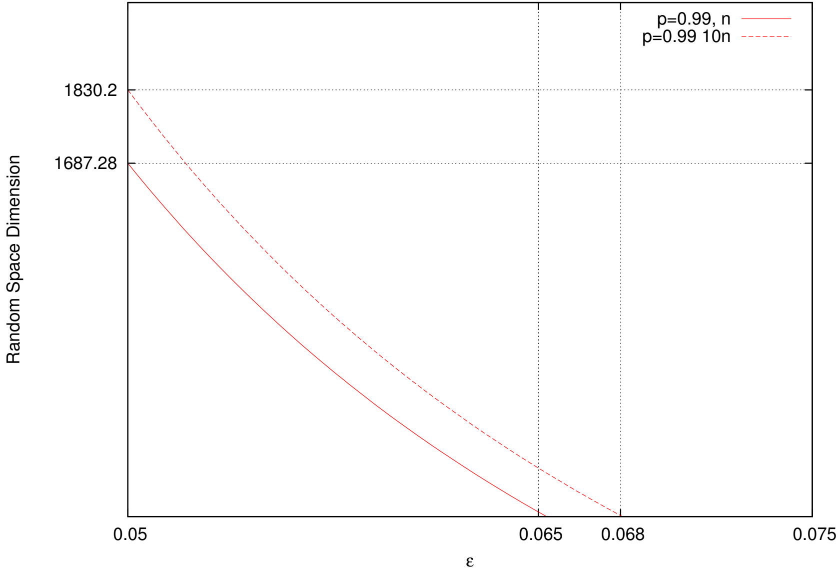
Trade offs

10 times more data



Trade offs

The effect of n



Random Projections Design

Selecting k_0

- Determine how much data is available: n
- Compute β for several certainty levels (99, 94, 90 typically)
- Compute and plot $k_0(\epsilon)$ for each p level
- Choose k_0 according to resources constraints and accepted ϵ
- More data? Start again. increase k_0, accept more error or more uncertainty

Random Projections Design

Selecting a random distribution

- Select k_0 using a dense random distribution
- No k_0 meets performance criteria?: Analyze sparsity.
- If data is indeed really sparse, consider matrix densification algorithms
- If k_0 meets performance criteria, evaluate a sparse random projection
- Achlioptas'distributions are not the only meeting JL property. Others could give better results

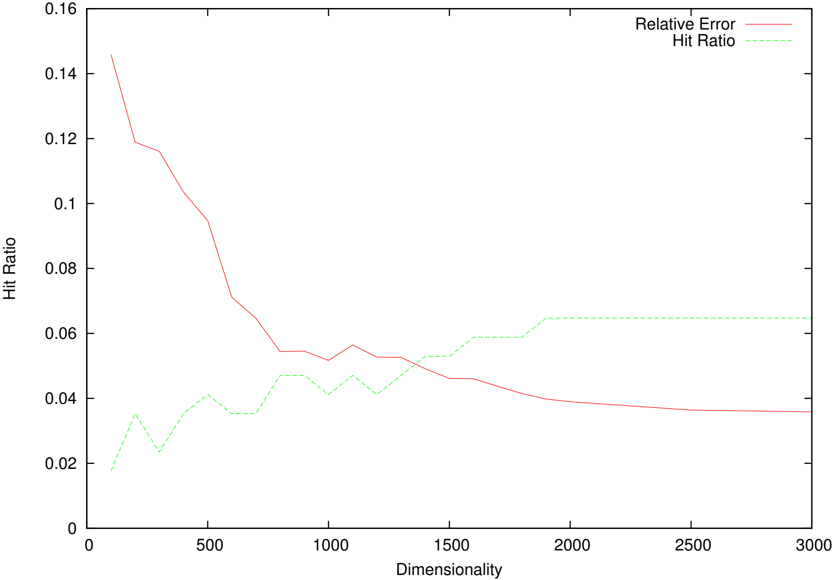
Random Projections Design

Weighting scheme

- Directional random vectors
- Distance permutations
- Reflective random projections

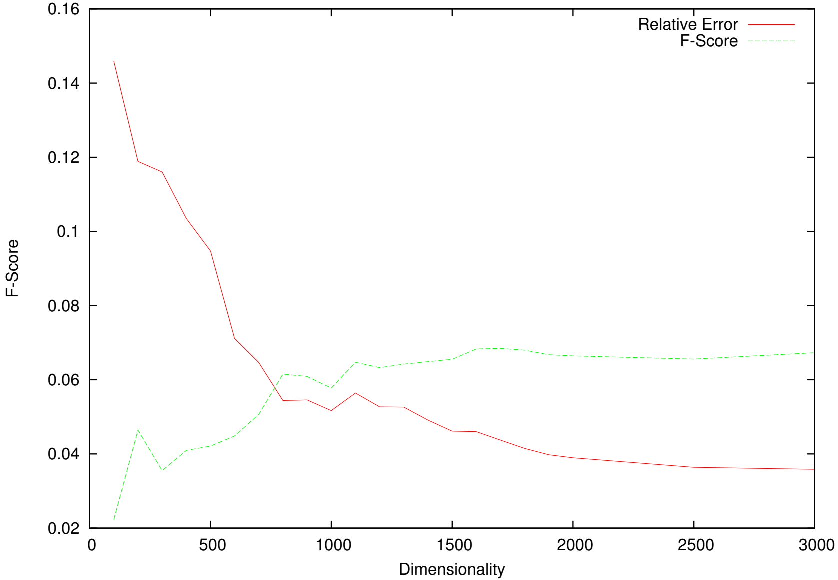
High Level Metrics Correlations

Hit Ratio



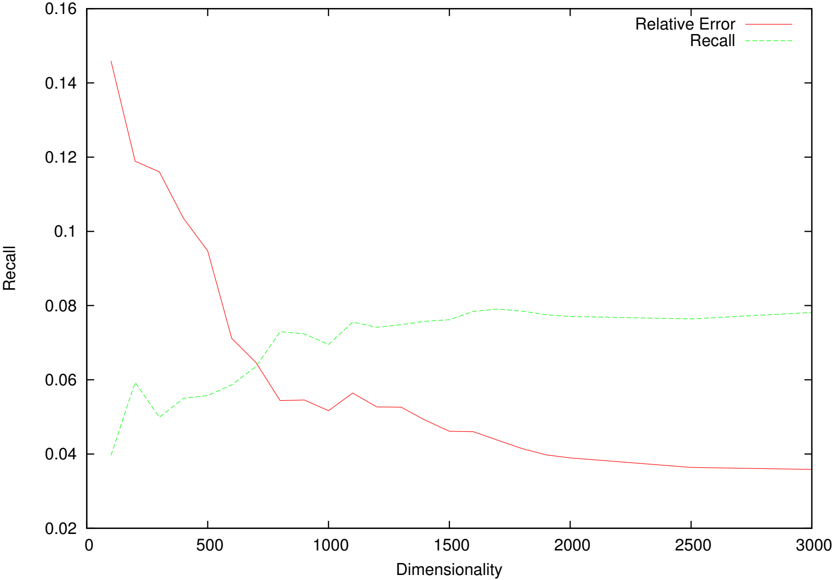
High Level Metrics Correlations

F-Score



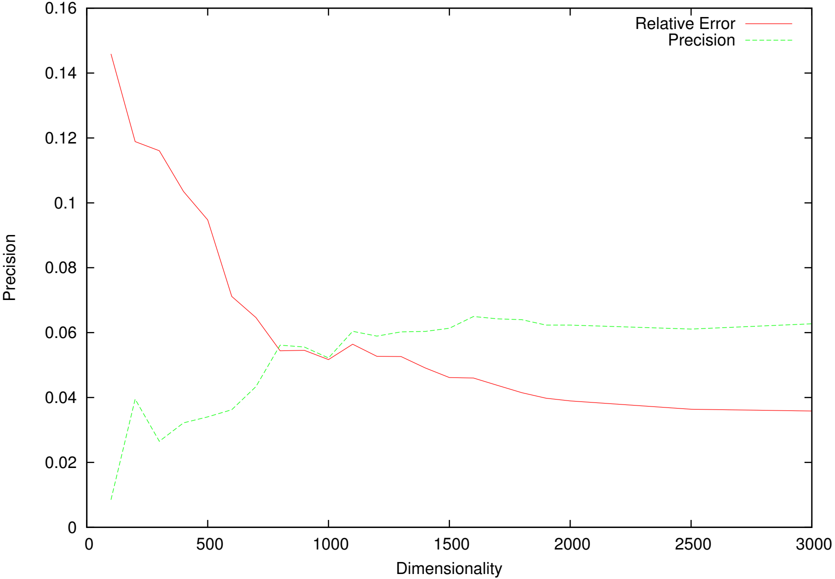
High Level Metrics Correlations

Recall



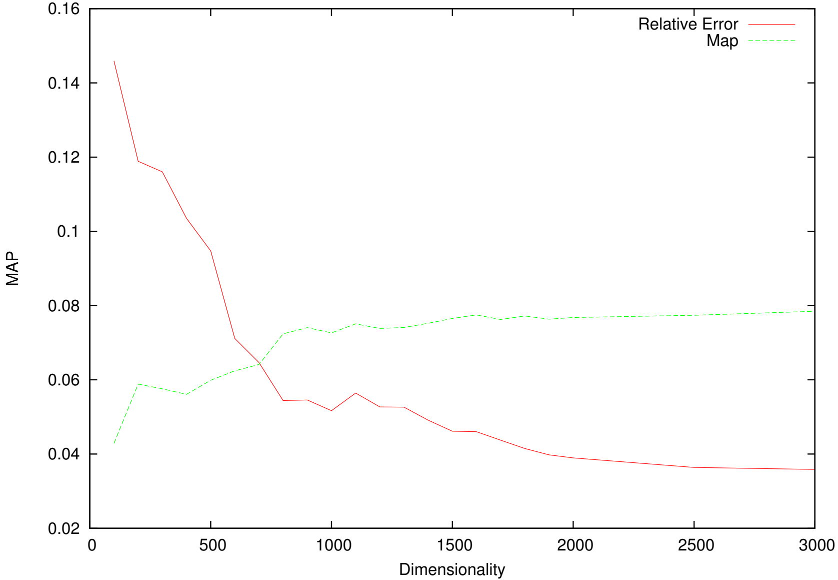
High Level Metrics Correlations

Recall



High Level Metrics Correlations

Recall



References

P. Kaneva, et al.
Random indexing of text samples for latent semantic analysis.
Proceedings of the 22nd annual conference of the cognitive science society. Vol. 1036. 2000.

D. Achlioptas.
Database-friendly random projections: Johnson-Lindenstrauss with binary coins.
Journal of computer and System Sciences 66.4 (2003): 671-687.