

Symposium on Machine Learning for anomaly Detection
Cordura Hall, Stanford University

Detecting Unique Documents via Visualization

May 22, 2004

Nippon Telephone and Telegraph Corporation
NTT Communication Science Laboratories

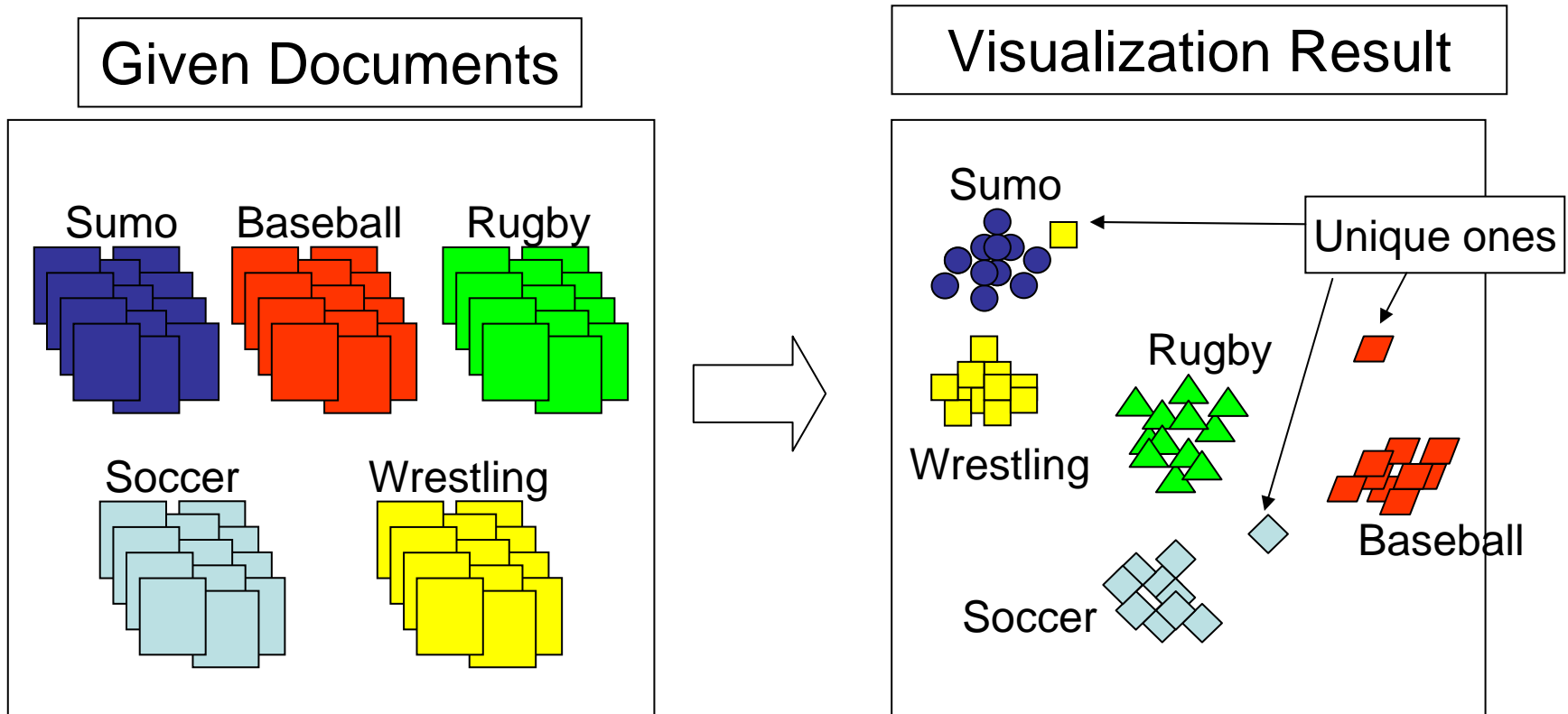
T. Iwata, K. Saito, and N. Ueda

Research Motivation

- A huge number of documents have been stored in the Web.
- One might want to find unique documents;
 - unusual combinations of words:
 - ice cream made of salt, ...
 - very specific description of some topics:
 - a complete survey of ice cream types, ...
- The uniqueness (anomaly) in this talk is defined by unusual usage of words in a document compared with the other ones.

Overview

- A tool for detecting unique documents via visualization



Proposing Framework & Method

- A new framework for detecting unique documents via visualization
- A new method for efficiently visualizing document categorization structures
- Procedure:
 1. Constructing a document categorization model
 2. Estimating posterior probabilities
 3. Visualizing documents in a low dim. space

Document Representation

- Naïve Bayes topic model:

$$p(\mathbf{x} | k) \propto \prod_{i=1}^V \theta_{ki}^{x_i}, \quad \theta_{ki} > 0, \quad \sum_{i=1}^V \theta_{ki} = 1,$$

- Word frequency vector: $\mathbf{x} = (x_1, x_2, \dots, x_V)$
- Vocabulary set: $\mathbf{W} = \{w_1, w_2, \dots, w_V\}$
- Topic: $k \in \{1, 2, \dots, K\}$

Parameter Estimation

- MAP objective function:

$$L_k = \sum_{n \in \mathbf{D}_k}^N \sum_{j=1}^V x_{nj} \log \theta_{kj} + \lambda_k \sum_{j=1}^V \log \theta_{kj}$$

Hyper parameter

- Optimally estimated value:

$$\hat{\theta}_{kj} = \frac{\sum_{n=1}^{N_k} x_{nj} + \lambda_k}{N_k + \lambda_k V}$$

- Hyper parameters can be efficiently estimated by using LOO (Leave-One-Out) cross-validation

Posterior Estimation

- ME (Maximum Entropy) estimation:

$$\frac{1}{N} \sum_{n=1}^N f(d_n, c_n) = \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K P(k | d_n) f(d_n, k)$$

- Posterior:
$$P(k | d_n) = \frac{\exp(\beta f(d_n, k))}{\sum_{k=1}^K \exp(\beta f(d_n, k))}$$

- Feature:

$$f(\mathbf{x}_n, k) = \sum_{j=1}^V x_{nj} \log \theta_{kj} = \log(p(\mathbf{x}_n | k))$$

Augmented Posterior Vector

- Random variable X : $P\left(X(w_j) = \log \hat{\theta}_j\right) = \hat{\theta}_j$
- Central Limit Theorem:

$$\frac{X_1 + \dots + X_M}{M} \sim N\left(E(X), \frac{V(X)}{\sqrt{M}}\right)$$

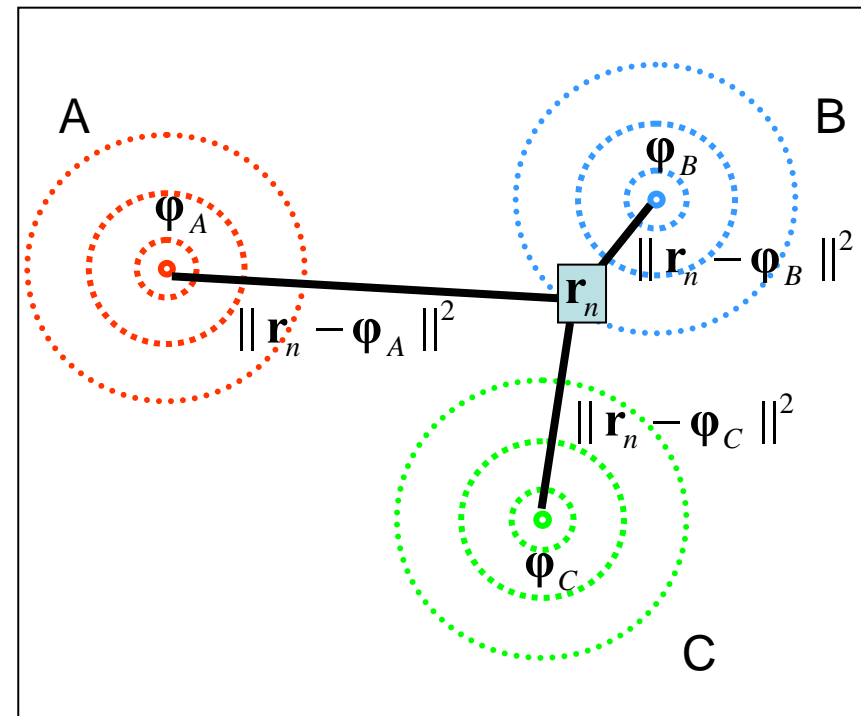
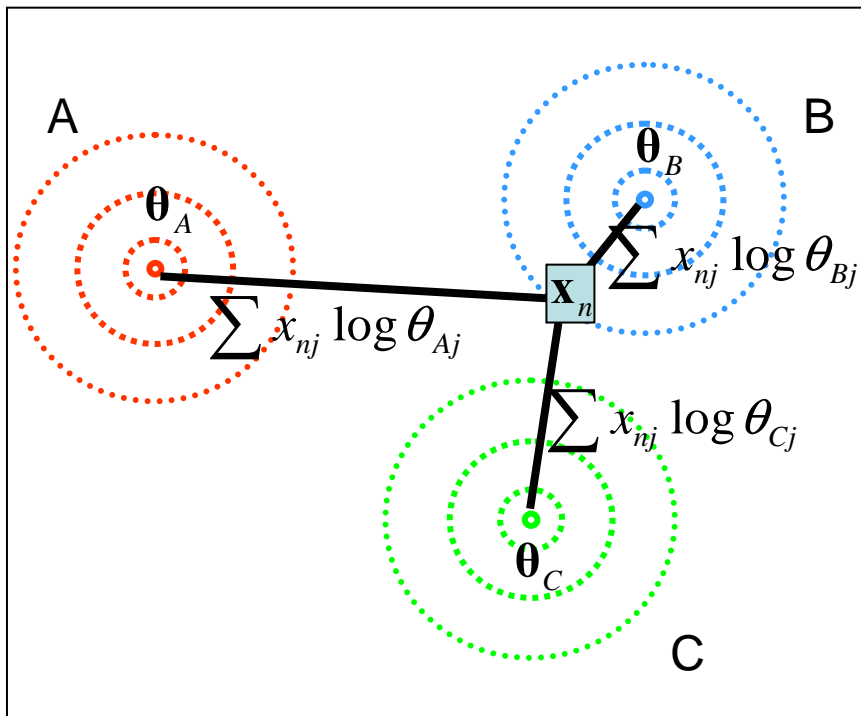
$$E(X) = \sum_{j=1}^V \hat{\theta}_j \log \hat{\theta}_j$$

- A new feature for uniqueness detection:

$$f(d_n, K + 1) = E(X) - 3\sqrt{V(X)/M_n}$$

Basic Idea of Visualization

- Transformation of distribution by focusing on posterior probabilities
 - from mixtures of high (V-) dimensional Multinomials
 - to mixtures of low (2- or 3-) dimensional Gaussians



Posterior Probabilities

- Multinomial mixture:

$$P(k | \mathbf{x}_n) = \frac{P(k) \exp\left(\sum x_{nj} \log \theta_{kj}\right)}{\sum_{l=1}^{K+1} P(l) \exp\left(\sum x_{nj} \log \theta_{lj}\right)}$$

- Gaussian mixture:

$$P(k | \mathbf{r}_n) = \frac{P(k) \exp\left(-\|\mathbf{r}_n - \boldsymbol{\phi}_k\|^2 / 2\right)}{\sum_{l=1}^{K+1} P(l) \exp\left(-\|\mathbf{r}_n - \boldsymbol{\phi}_l\|^2 / 2\right)}$$

Visualization Method

- Minimization KL-divergence

$$L_v = - \sum_{n=1}^N \sum_{k=1}^{K+1} P(k | \mathbf{x}_n) \log P(k | \mathbf{r}_n; \Phi)$$

- Coordinate descent approach:
 - Optimize \mathbf{r}_n while fixing Φ_k (global optimal)
 - Optimize Φ_k while fixing \mathbf{r}_n

Evaluation using Web Data

ウェブ イメージ グループ **ディレクトリ**

Google 検索 [表示設定](#)

カテゴリー別Google!

<u>アート</u> 写真 , 文学 , 映画 , ...	<u>ニュース</u> 新聞 , テレビ , ...	<u>各種資料</u> 辞書 ・ 事典 , ...
<u>ゲーム</u> ビデオゲーム , オンライン , ...	<u>ビジネス</u> 金融 , ...	<u>家庭</u> ガーデニング , 料理 , 暮らし , ...
<u>コンピュータ</u> インターネット , ソフトウェア , ...	<u>レクリエーション</u> アウトドア , 旅行 , 車 ・ バイク , ...	<u>社会</u> 政治 , 教育 , 時事 , ...
<u>スポーツ</u> サッカー , ゴルフ , 野球 , ...	<u>健康</u> 体調 ・ 症例 , 美容 , ...	<u>科学</u> 天文 , 社会科学 , ...

関連カテゴリー:
[Regional > Asia > Japan](#) (7287)

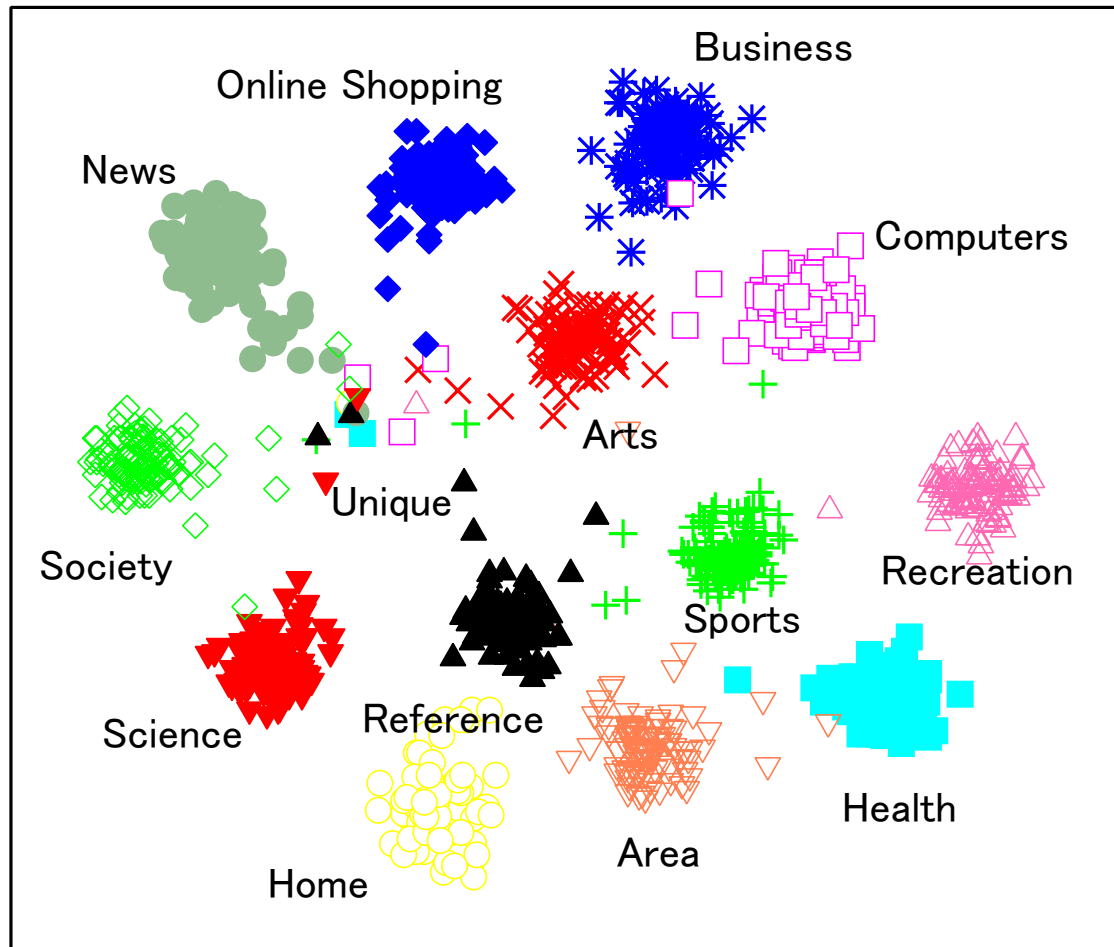
[広告掲載について](#) - [人材募集](#) - [Googleについて](#)

©2004 Google

あなたも「ウェブ最大のディレクトリ作り」に参加しませんか
[URLを登録する](#) - [Open Directory Project](#) - [編集者募集](#)

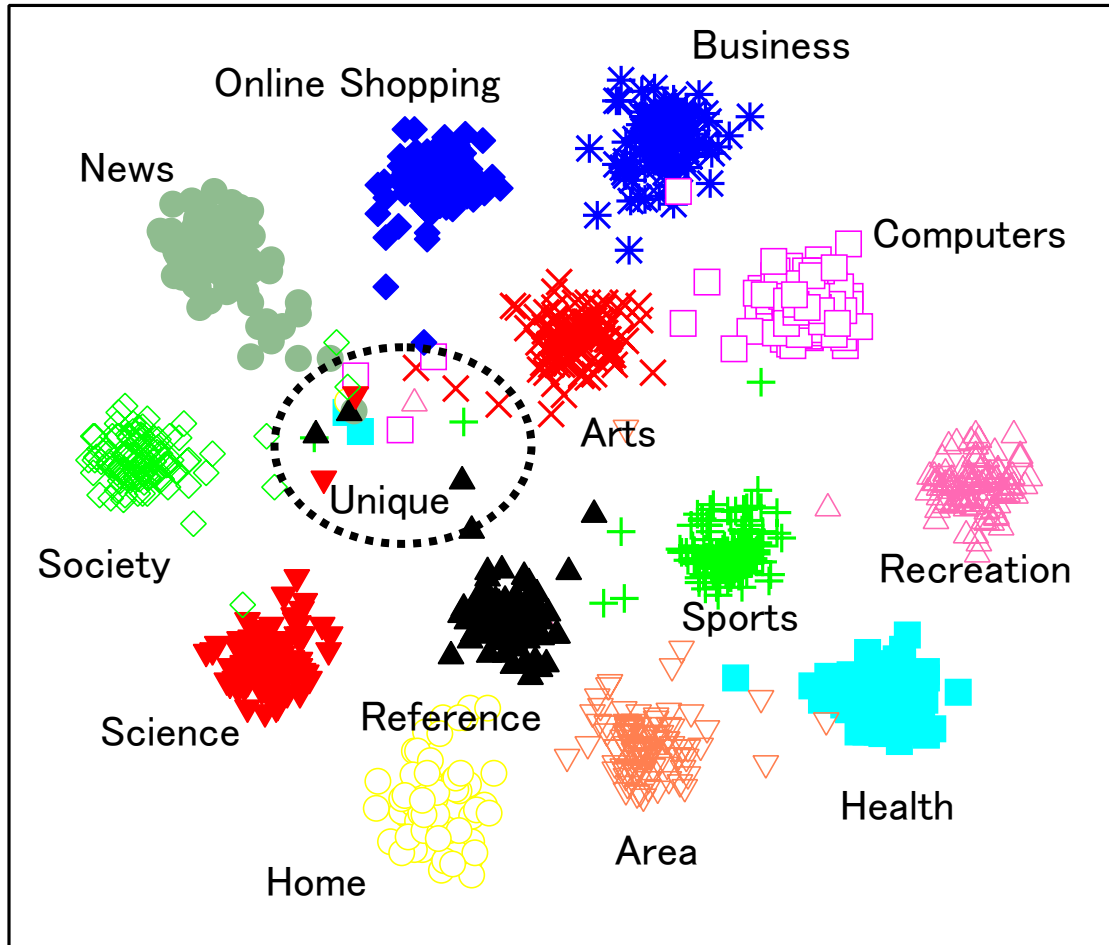
A Visualization result

- Related topics were nearly placed (Business & Online Shopping, Sports & Health, ...)



Unique Documents

- A new cluster for unique documents appeared, which is placed near news, online shopping, and society.



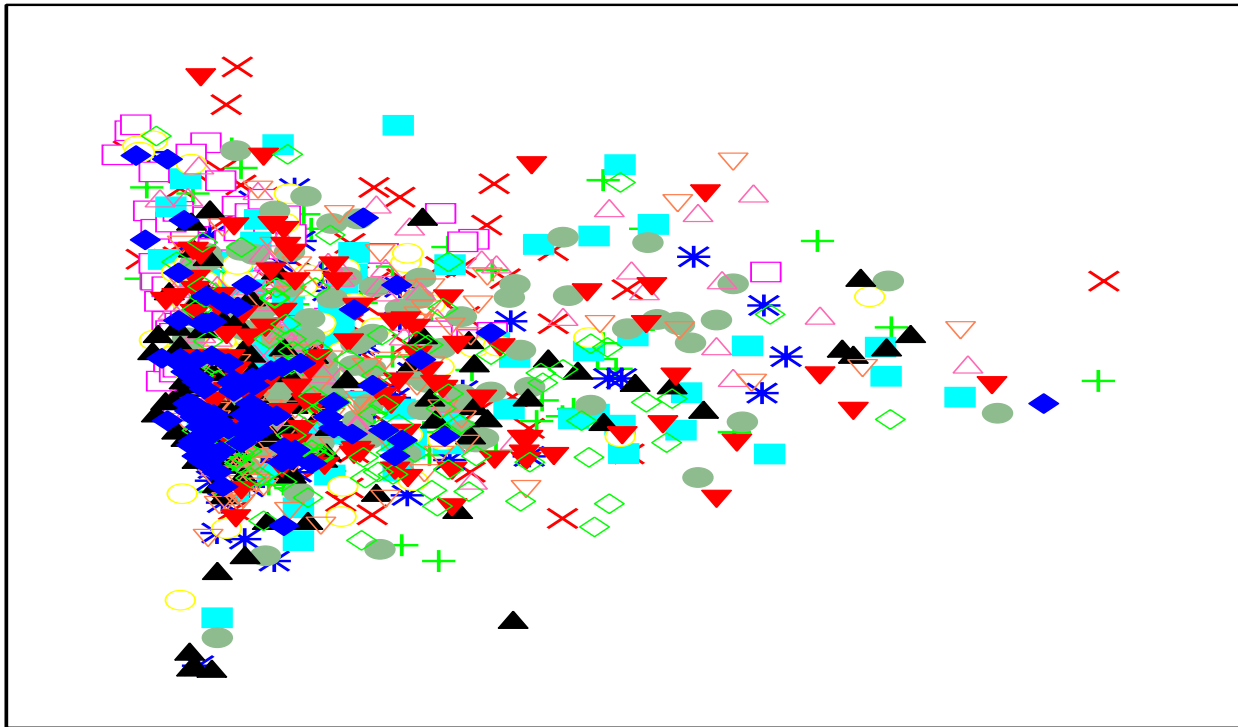
Demonstration

- A tool for intuitively understanding and easily picking up unique documents.

The screenshot displays the 'Experiment Tool' interface. The top menu bar includes 'ファイル' (File) and 'Address'. Below the menu, there are navigation buttons: '戻る' (Back), '進む' (Forward), '中止' (Stop), 'ホーム' (Home), '可視化' (Visualization), 'フィルタ' (Filter), '開く' (Open), and '保存する' (Save). A 'node size' section is visible with a checkbox for 'checked node is displayed'. The main content area shows a list of documents with checkboxes and titles in Japanese. A search bar and various filters are also present. On the right side, a network visualization shows a cluster of nodes connected by lines, with nodes colored in various colors like purple, yellow, and green. The bottom of the interface has a status bar with '新規作成' (New), '編集' (Edit), '削除' (Delete), and '更新' (Update) buttons, along with a 'URL' field.

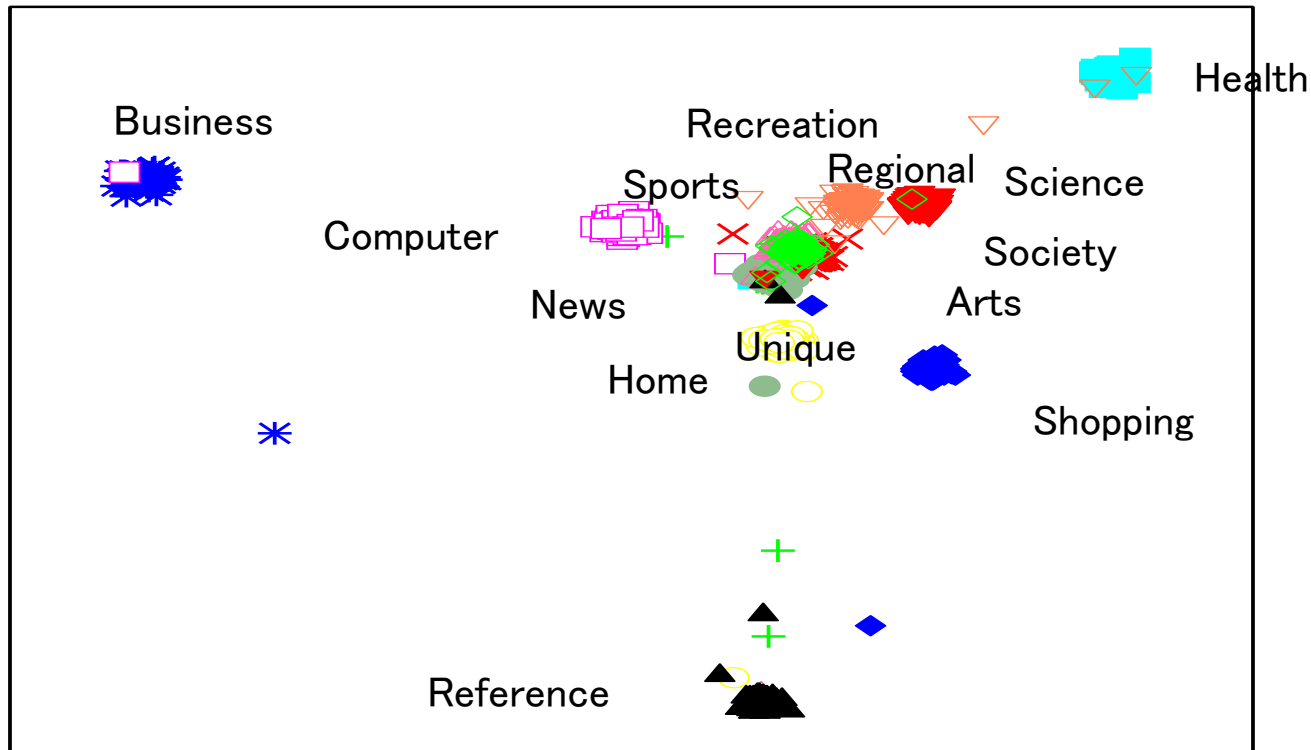
Linear methods (MDS, PCA, ...)

- embed documents into a low dimensional space, so as to preserve pair-wise document similarities;
- are limited to extract a linear structure;
- cannot use category information of documents.



Linear Methods (2)

- can be used to preserve pair-wise similarities of posterior probabilities;
- are limited to extract a linear structure;



Class Preserving Projection

- Fisher's linear Discriminant
 - Can use class information
 - Maximizes between-class scatters

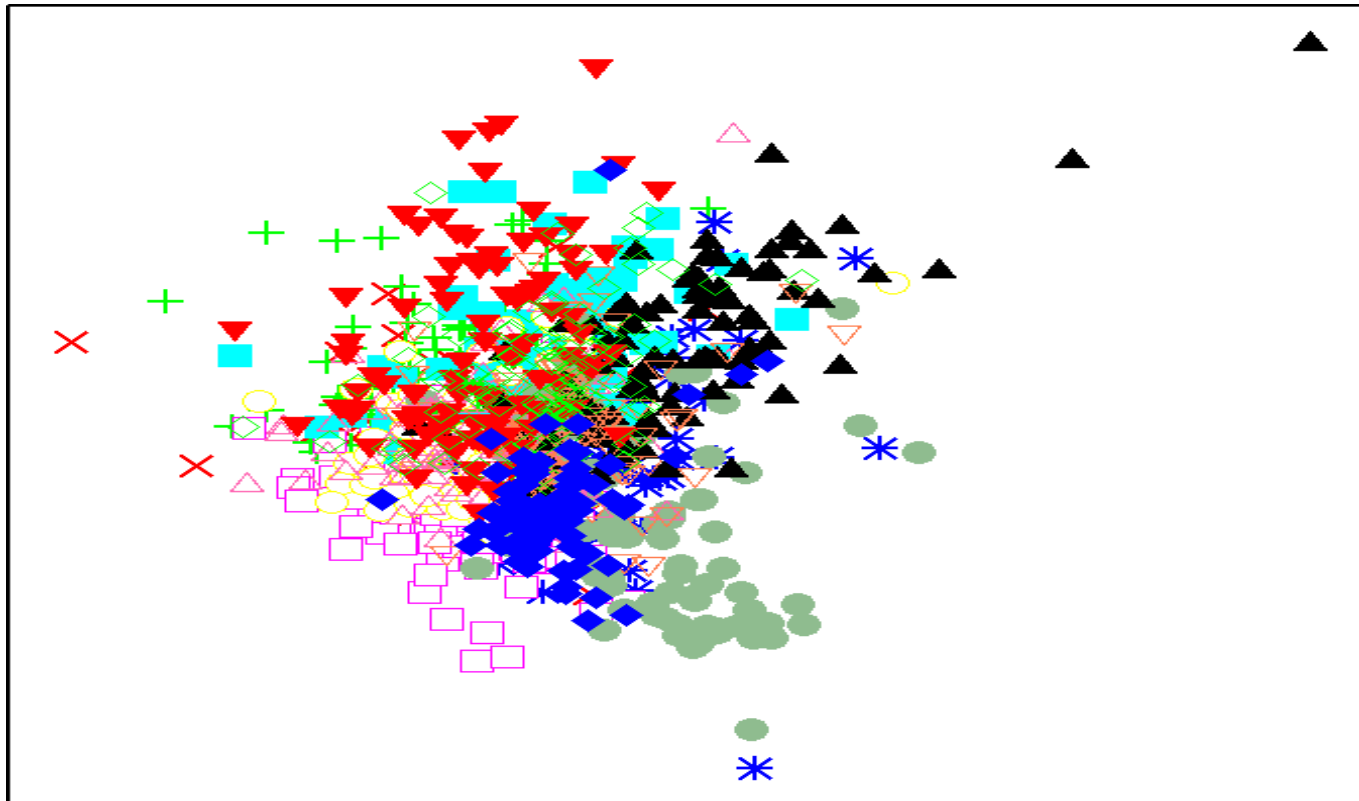
$$L_{CPP} = \text{trace}(\mathbf{W}^T \mathbf{S}_B \mathbf{W})$$

$$\mathbf{S}_B = \sum_{k=2}^K \sum_{j=1}^{k-1} N_k N_j \left(\frac{1}{N_k} \sum_{n \in D_k} \mathbf{x}_n - \frac{1}{N_j} \sum_{n \in D_j} \mathbf{x}_n \right) \left(\frac{1}{N_k} \sum_{n \in D_k} \mathbf{x}_n - \frac{1}{N_j} \sum_{n \in D_j} \mathbf{x}_n \right)^T$$

$$\mathbf{r}_n = \mathbf{W}^T \mathbf{x}_n$$

Class Preserving Projection (2)

- Placed documents within the same class nearly;
- But many documents overlapped



Nonlinear Methods

including Isomap, SNE (Stochastic Neighbor Embedding), Spring method, and so on

- cannot use category information;
- usually require a huge computational time.

- SNE

$$O(N^2) : L_{SNR} = \sum_{n=1}^N \sum_{k=1}^N p(d_k | d_n) \log q(r_k | r_n)$$

- Proposed

$$O(NK) : L_v = \sum_{n=1}^N \sum_{k=1}^K P(k | d_n) \log P(k | \mathbf{r}_n)$$

Summary

- We proposed a new framework for detecting unique documents via visualization.
 - representing each document as a posterior probability vector augmented with a score to detect its unusualness.
 - constructing a map of documents by embedding their vectors into 2- or 3-dimensional space, so as to preserve their probabilistic structures.
- Our experiments using real Web data showed
 - Our approach can be implemented as a promising tool
 - intuitively understanding categorization structures,
 - easily picking up unique documents.

Future Plans

- Evaluation using a wider range of documents and probabilistic models;
- TDT (Topic Detection and Tracking) from time series data of documents;
- Prediction of topic outbreaks
- ...