Impact of Similarity Measures on Web-page Clustering

Alexander Strehl
Joydeep Ghosh
Raymond Mooney
University of Texas at Austin

17th **AAAI** (2000)

presented by Eli Rosenberg
infs 795

Overview

- Web documents are sparse, high dimensional data as represented in bag of words format
- No studies comprehensively evaluating possible similarity metrics have been undertaken
- Evaluate the following similarity metrics: Euclidean, Cosine, Jaccard, and Pearson Correlation
- Utilize information theoretic means for evaluation (mutual information)

Dimensionality

• **Curse of Dimensionality** always a problem in vector space approaches, so we need to utilize algorithmic means to perform the clustering

• Possibilities analyzed: Graph Partitioning, Hyper-Graph Partitioning, Generalized k-means, and Kohonen Self Organizing Feature Map

• 11 total combinations were used in experimentation, 4 each for k-means and graph partitioning, 1 each for SOFM, hypergraph, and random clustering.

• (question: why exactly these combinations, does SOFM not allow for changing the metric?)

Metric Space Formulation

n web-pages, d features, $d \times n$ matrix
samples $\mathbf{x}_j$, $j \in \{1, \ldots, n\}$, thus $\mathbf{x}_j$ are the columns

• Want to label each $\mathbf{x}_j$ with a single label λ_j, such that similar columns (the $\mathbf{x}_j$-s) have the same label (hard clustering)

• We can view this as a transformation process:

1. take the $d \times n$ representation to a similarity space $\mathbf{S}$ of dimension $n \times n$ over the domain [0,1]

2. then transform to an n-vector by labeling each page with a label from the domain $\{1, \ldots, k\}$, where k is the number of clusters.

Algorithms, 1

Self Organizing Feature Map

• SOMs are well known body of methods by Kohonen

▷ ran for upto 10 minutes using MATLAB neural net toolbox implementation

Generalized k-means

• determine cluster centers and location of the nearest center, parameterized by the underlying distance metric.

▷ iterate until convergence of a fixed labeling in time complexity $O(n \cdot d \cdot k \cdot m)$, where m is the number of iterations, k the number of clusters, d terms, n pages

• (my comment: does not sound too fast or flexible, possibly runs into problems with computational optimization strategy)

Algorithms, 2

Weighted Graph Partitioning

• connect web-pages with non zero similarity to form $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with edges weighted by the degree of similarity. Now we can use body of combinatorial approaches for undirected graphs with positive flows.

▷ determine disjoint subgraphs by utilizing sparsity, the most expensive portion having complexity $O(n^2 \cdot d)$. This is done by a variation of min-cut partitioning called OPOSSUM

• note that Euclidean metrics do not introduce sufficient sparsity, because it is difficult to prune edges under some chosen threshold

Algorithms, 3

Hyper-graph partitioning

• hyper-graphs are graphs containing edges with connectivity to >2 vertices

• Again formulated as a min-cut partitioning problem, where edges correspond to words, with each edge connecting all vertices (web-pages) containing that word. Edges are weighted proportional to term frequency. The time complexity of $O(n \cdot d \cdot k)$ for computing the partitions is the same order as the first graph partitioning method, but clearly lower.

Measures, 1

Euclidean

$$\left(\sum_{i=1}^{d} |x_{i,a} - x_{i,b}|^p \right)^{(1/p)} \quad , \quad p = 2$$

• Change to [0,1] normalized $s^{(E)}(\mathbf{x}_a, \mathbf{x}_b) = e^{-\|\mathbf{x}_a - \mathbf{x}_b\|^2}$ which allows for a dualization of errors (distance) and probabilities (similarity). The proof relates the metrics via [convex] least squares minimization and a dual maximization problem.

Cosine

$$s^{(C)}(\mathbf{x}_a, \mathbf{x}_b) = \frac{\mathbf{x}_a^{\mathrm{T}} \mathbf{x}_b}{\|\mathbf{x}_a\|_2 \cdot \|\mathbf{x}_a\|_2}$$

• nice scale invariance property, very popular for text

Measures, 2

Pearson Correlation

• Frequency means of columns are substracted to yield the degree of linearity between two web-pages.

• (question: what are its invariance properties?)

$$\frac{1}{2}\left(\frac{(\mathbf{x}_a - \overline{\mathbf{x}}_a)^{\mathrm{T}}(\mathbf{x}_b - \overline{\mathbf{x}}_b)}{\|\mathbf{x}_a - \overline{\mathbf{x}}_a\| \cdot \|\mathbf{x}_b - \overline{\mathbf{x}}_b\|} + 1\right)$$

Jaccard Similarity

• measures shared attributes out of total attributes, as used in some types of correspondence matching, e.g. Market Basket

$$\frac{\mathbf{x}_a^{\mathrm{T}}\mathbf{x}_b}{\|\mathbf{x}_a\|^2 + \|\mathbf{x}_b\|^2 - \mathbf{x}_a^{\mathrm{T}}\mathbf{x}_b}$$

Evaluation Methodology

• Start with a cluster purity measure $\Lambda^{(P)}(C_\ell) = (1/n_\ell) \max_h(n_\ell^{(h)})$ where $n_\ell^{(h)}$ is the number of documents in class h as found in an external labeling

• Can be changed to an entropy across all g classes, but neither approach works that well, because a cluster with a single sample ends up being scored highest.

• use mutual information instead, which takes into account information dependencies across all clusters, small or large.

$$\frac{1}{n} \sum_{\ell=1}^{k} \sum_{h=1}^{g} n_\ell^{(h)} \frac{\log\left(\frac{n_\ell^{(h)} n}{\sum_{i=1}^{k} n_i^{(h)} \sum_{i=1}^{g} n_\ell^{(i)}}\right)}{\log(k \cdot g)}$$

Results

• YAHOO news (2340 docs) and YAHOO industry (966 docs) categories data are used as the two datasets, which are apparently far less clean than the REUTERS data. Under varying sample sizes, mutual information evaluator was calculated for the clusterings derived from the 11 combinations.

• About $2\times$ the number of classes was chosen as the fixed number of clusters, because preliminary investigations made this seem a reasonable estimate.

• best performer was weighted graph partitioning using cosine measure; other non-Euclidean graph partitioning methods followed closely, then k-means (cosine, correlation)

• Only unsuitable combinations were Euclidean (all) & SOFM.
$\triangleright$ See the large groups of plots for more quantitative view.

Conclusion

• Graph Partitioning algorithm captures more global information than k-means, making it a better approach for the clustering of text web-pages.

• Euclidean distance metrics perform far too poorly to be used in this type of clustering. (question: what about length normalized Euclidean vs. the [0,1] normalization tested here?)

• Summary: High dimensional, sparse domains require creative approaches to derive suitable metrics, as well as efficient algorithms for clustering based on these metrics.

• (I recommend looking at the 'confusion matrix' found in the paper, which was too large to reproduce here, and to also look at the corresponding mutual information values)