Chapter 4

Impact of Similarity Measures

I have had my results for a long time: but I do not yet
know how I am to arrive at them.
– Karl Friedrich Gauß[1]

In the last chapter, we explored the relationship-based approach to clustering in several domains. The work was initially motivated by retail data and extended naturally to other domains where high-dimensional representations are prevalent, such as text documents and web-logs. A particularly interesting application is clustering of text documents which enables unsupervised categorization and facilitates browsing and search. A critical step in adapting a relationship-based clustering to a specific domain is the choice of similarity measure. In this chapter, we investigate the impact of similarity measures on clustering quality. We will first introduce similarities and algorithms for text clustering, then develop a general comparative framework and, finally, conduct case studies on a variety of text corpora.

[1]Quoted in A. Arber, *The Mind and the Eye*, 1954

4.1 Motivation

Document clusters can provide a structure for organizing large bodies of text for efficient browsing and searching. For example, recent advances in Internet search engines (e.g., `http://vivisimo.com/`, `http://metacrawler.com/`) exploit document cluster analysis. For this purpose, a document is commonly represented as a vector consisting of the suitably normalized frequency counts of words or terms. Each document typically contains only a small percentage of all the words ever used. If we consider each document as a multi-dimensional vector and then try to cluster documents based on their word contents, the problem differs from classic clustering scenarios in several ways: Document data is high-dimensional[2], characterized by a very sparse term-document matrix with positive ordinal attribute values and a significant amount of outliers. In such situations, one is truly faced with the 'curse of dimensionality' issue [Fri94] since, even after feature reduction, one is left with hundreds of dimensions per object.

In the previous chapter, we developed the relationship-clustering framework to effectively side-step the 'curse of dimensionality'. In the relationship-based clustering process, key cluster analysis activities [JD88] can be associated with each step:

1. To obtain features $\mathbf{X} \in \mathcal{F}$ from the raw objects, a suitable *object representation* has to be found. We will not be concerned with representation in this chapter, since the significant amount of empirical studies on document clustering in the 80s and earlier emphasized various ways of

[2]The dimension of a document in vector space representation is the size of the vocabulary, often in the tens of thousands.

representing / normalizing documents [Wil88, SB88, Sal89].

2. In the second step, a *measure of proximity* $\mathbf{S} \in \mathcal{S}$ has to be defined between objects. The choice of similarity or distance can have a profound impact on clustering quality. In this chapter, we first compare similarity measures analytically and then illustrate their semantics geometrically.

3. The third activity requires a suitable choice of *clustering algorithm* to obtain cluster labels $\lambda \in \mathcal{O}$. Agglomerative clustering approaches were historically dominant as they compared favorably with flat partitional approaches on small or medium sized collections [Wil88, Ras92]. But lately, some new partitional methods have emerged (spherical k-means, graph partitioning-based, etc.) that have attractive properties in terms of both quality and scalability and can work with a wider range of similarity measures. In addition, much larger document collections are being generated.[3] This warrants an updated comparative study on text clustering, which is the motivation behind this chapter.

4. Finally, in the *assessment of output* one has to investigate the validity of the results.[4] In this chapter, we propose an experimental methodology to compare high-dimensional clusterings based on mutual information and we show how this is better than purity or entropy-based measures [BGG+99, ZK01, SKK00]. Finally, we conduct a series of experiments to evaluate the performance and cluster quality of four similarity measures (Euclidean, cosine, Pearson correlation, extended Jaccard) in com-

[3]IBM Patent Server has over 20 million patents. Lexis-Nexis contains over 1 billion documents

[4]Often, *data abstraction* has to performed between clustering and final assessment [JD88].

bination with five algorithms (random, self-organizing map, hypergraph partitioning, generalized k-means, weighted graph partitioning).

Some very recent, notable comparative studies on document clustering [SKK00, ZK01] also consider some of the newer issues. Our work is distinguished from these efforts mainly by its focus on the key role of the similarity measures involved, emphasis on balancing, and the use of a normalized mutual information-based evaluation that we believe has superior properties.

The basic notation is the same as introduced in the previous chapter in section 3.2. In the next section, we introduce several similarity measures, illustrate some of their properties, and show why we are interested in some but not others. In section 4.3, the algorithms using these similarity measures are discussed. Section 4.4 introduces a variety of cluster quality evaluation methods including our proposed mutual information criterion. Finally, the experiments and results are shown in section 4.5.

4.2 Similarity Measures for Document Clustering

4.2.1 Conversion from a Distance Metric

The Minkowski distances $L_p(\mathbf{x}_a, \mathbf{x}_b) = \left(\sum_{i=1}^{d} |\mathbf{x}_{i,a} - \mathbf{x}_{i,b}|^p\right)^{1/p}$ are the standard metrics for geometrical problems. For $p = 2$ we obtain the Euclidean distance. There are several possibilities for converting such a distance metric (in $[0, \inf)$, with 0 closest) into a similarity measure (in $[0, 1]$, with 1 closest) by a monotonic decreasing function. For Euclidean space, we chose to relate dis-

tances d and similarities s using $s = e^{-d^2}$. Consequently, we define Euclidean [0,1]-normalized similarity as

$$s^{(\mathrm{E})}(\mathbf{x}_a, \mathbf{x}_b) = e^{-\|\mathbf{x}_a - \mathbf{x}_b\|_2^2} \tag{4.1}$$

which has important desirable properties (as we will see in the discussion) that the more commonly adopted $s(\mathbf{x}_a, \mathbf{x}_b) = 1/(1 + \|\mathbf{x}_a - \mathbf{x}_b\|_2)$ lacks. Other distance functions can be used as well. The Mahalanobis distance normalizes the features using the covariance matrix. Due to the high-dimensional nature of text data, covariance estimation is inaccurate and often computationally intractable, and normalization is done if need to be, at the document representation stage itself, typically by applying TF-IDF.

4.2.2 Cosine Measure

A popular measure of similarity for text (which normalizes the features by the covariance matrix) clustering is the cosine of the angle between two vectors. The cosine measure is given by

$$s^{(\mathrm{C})}(\mathbf{x}_a, \mathbf{x}_b) = \frac{\mathbf{x}_a^\dagger \mathbf{x}_b}{\|\mathbf{x}_a\|_2 \cdot \|\mathbf{x}_b\|_2} \tag{4.2}$$

and captures a scale invariant understanding of similarity. An even stronger property is that the cosine similarity does not depend on the length:
$s^{(\mathrm{C})}(\alpha\mathbf{x}_a, \mathbf{x}_b) = s^{(\mathrm{C})}(\mathbf{x}_a, \mathbf{x}_b)$ for $\alpha > 0$. This allows documents with the same composition, but different totals to be treated identically which makes this the most popular measure for text documents. Also, due to this property, samples can be normalized to the unit sphere for more efficient processing [DM01].

4.2.3 Pearson Correlation

In collaborative filtering, correlation is often used to predict a feature from a highly similar mentor group of objects whose features are known. The [0,1]-normalized Pearson correlation is defined as

$$s^{(\mathrm{P})}(\mathbf{x}_a, \mathbf{x}_b) = \frac{1}{2}\left(\frac{(\mathbf{x}_a - \bar{x}_a)^\dagger(\mathbf{x}_b - \bar{x}_b)}{\|\mathbf{x}_a - \bar{x}_a\|_2 \cdot \|\mathbf{x}_b - \bar{x}_b\|_2} + 1\right), \tag{4.3}$$

where $\bar{x}$ denotes the average feature value of $\mathbf{x}$ over all dimensions. Note that this definition of Pearson correlation tends to give a full matrix. Other important correlations have been proposed, such as Spearman correlation [Spe06] which works well on rank orders.

4.2.4 Extended Jaccard Similarity

The binary Jaccard coefficient measures the degree of overlap between two sets and is computed as the ratio of the number of shared attributes (words) of $\mathbf{x}_a$ AND $\mathbf{x}_b$ to the number possessed by $\mathbf{x}_a$ OR $\mathbf{x}_b$. For example, given two sets' binary indicator vectors $\mathbf{x}_a = (0, 1, 1, 0)^\dagger$ and $\mathbf{x}_b = (1, 1, 0, 0)^\dagger$, the cardinality of their intersect is 1 and the cardinality of their union is 3, rendering their Jaccard coefficient 1/3. The binary Jaccard coefficient is often used in retail market-basket applications. In chapter 3, we extended the binary definition of Jaccard coefficient to continuous or discrete non-negative features. The extended Jaccard is computed as

$$s^{(\mathrm{J})}(\mathbf{x}_a, \mathbf{x}_b) = \frac{\mathbf{x}_a^\dagger \mathbf{x}_b}{\|\mathbf{x}_a\|_2^2 + \|\mathbf{x}_b\|_2^2 - \mathbf{x}_a^\dagger \mathbf{x}_b}, \tag{4.4}$$

which is equivalent to the binary version when the feature vector entries are binary. Extended Jaccard similarity [SG00c] retains the sparsity property of

the cosine while allowing discrimination of collinear vectors as we will show in the following subsection. Another similarity measure highly related to the extended Jaccard is the Dice coefficient ($s^{(\mathrm{D})}(\mathbf{x}_a, \mathbf{x}_b) = \frac{2\mathbf{x}_a^\dagger \mathbf{x}_b}{\|\mathbf{x}_a\|_2^2 + \|\mathbf{x}_b\|_2^2}$). The Dice coefficient can be obtained from the extended Jaccard coefficient by adding $\mathbf{x}_a^\dagger \mathbf{x}_b$ to both the numerator and denominator. It is omitted here since it behaves very similar to the extended Jaccard coefficient.

4.2.5 Other (Dis-)Similarity Measures

Many other (dis-)similarity measures, such as mutual neighbor or edit distance, are possible [JMF99]. In fact, the ugly duckling theorem states [Wat69] the somewhat 'unintuitive' fact that there is no way to distinguish between two different classes of objects, when they are compared over all possible features. As a consequence, any two arbitrary objects are equally similar unless we use domain knowledge. The similarity measures discussed above are the ones deemed pertinent to text documents [Sal89, FBY92] in previous studies.

4.2.6 Discussion

Clearly, if clusters are to be meaningful, the similarity measure should be invariant to transformations natural to the problem domain. Also, normalization may strongly affect clustering in a positive or negative way. The features have to be chosen carefully to be on comparable scales and similarity has to reflect the underlying semantics for the given task.

Euclidean similarity is translation invariant but scale sensitive while cosine is translation sensitive but scale invariant. The extended Jaccard has aspects of both properties as illustrated in figure 4.1. Iso-similarity lines at

$s = 0.25$, 0.5 and 0.75 for points $\mathbf{x}_1 = (3,1)^\dagger$ and $\mathbf{x}_2 = (1,2)^\dagger$ are shown for Euclidean, cosine, and the extended Jaccard. For cosine similarity only the 4 (out of 12) lines that are in the positive quadrant are plotted: The two lines in the lower right part are one of two lines from $\mathbf{x}_1$ at 0.5 and 0.75. The two lines in the upper left are for $\mathbf{x}_2$ at $s = 0.5$ and 0.75. The dashed line marks the locus of equal similarity to $\mathbf{x}_1$ and $\mathbf{x}_2$ which always passes through the origin for cosine and extended Jaccard similarity.

Using Euclidean similarity $s^{(\mathrm{E})}$, iso-similarities are concentric hyperspheres around the considered point. Due to the finite range of similarity, the radius decreases hyperbolically as $s^{(\mathrm{E})}$ increases linearly. The radius does not depend on the center-point. The only location with similarity of 1 is the considered point itself and all finite locations have a similarity greater than 0. This last property tends to generate non-sparse similarity matrices. Using the cosine measure $s^{(\mathrm{C})}$ renders the iso-similarities to be hypercones all having their apex at the origin and axis aligned with the considered point. Locations with similarity 1 are on the 1-dimensional sub-space defined by this axis. The locus of points with similarity 0 is the hyperplane through the origin and perpendicular to this axis. For the extended Jaccard similarity $s^{(\mathrm{J})}$, the iso-similarities are non-concentric hyperspheres. The only location with similarity 1 is the point itself. The hypersphere radius increases with the the distance of the considered point from the origin so that longer vectors turn out to be more tolerant in terms of similarity than smaller vectors. Sphere radius also increases with similarity and as $s^{(\mathrm{J})}$ approaches 0 the radius becomes infinite rendering the sphere to the same hyperplane as obtained for cosine similarity. Thus, for $s^{(\mathrm{J})} \to 0$, extended Jaccard behaves like the cosine measure, and for $s^{(\mathrm{J})} \to 1$, it behaves like the Euclidean distance.

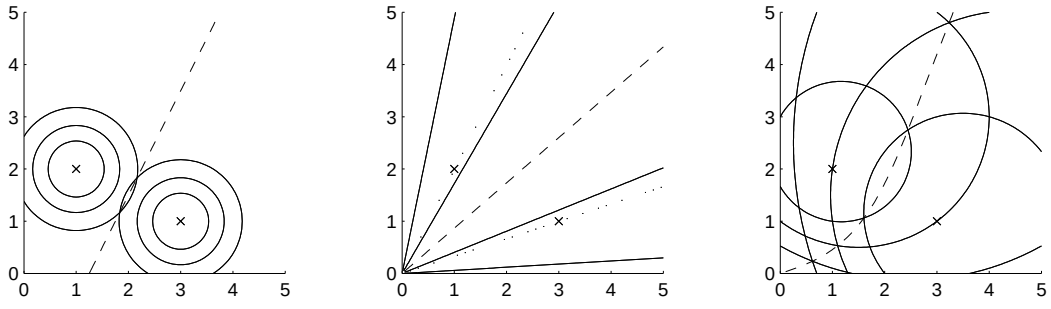

Figure 4.1: Properties of (a) Euclidean-based, (b) cosine, and (c) extended Jaccard similarity measures illustrated in 2 dimensions. Two points $(1, 2)^\dagger$ and $(3, 1)^\dagger$ are marked with $\times$s. For each point iso-similarity surfaces for $s = 0.25$, 0.5, and 0.75 are shown with solid lines. The surface that is equi-similar to the two points is marked with a dashed line.

In traditional Euclidean k-means clustering the optimal cluster representative $\mathbf{c}_\ell$ minimizes the sum of squared error criterion, i.e.,

$$\mathbf{c}_\ell = \arg\min_{\mathbf{z}\in\mathcal{F}} \sum_{\mathbf{x}_j\in\mathcal{C}_\ell} \|\mathbf{x}_j - \mathbf{z}\|_2^2. \tag{4.5}$$

In the following, we show how this convex distance-based objective can be translated and extended to similarity space. Consider the generalized objective function $f(\mathcal{C}_\ell, \mathbf{z})$ given a cluster $\mathcal{C}_\ell$ and a representative $\mathbf{z}$:

$$f(\mathcal{C}_\ell, \mathbf{z}) = \sum_{\mathbf{x}_j\in\mathcal{C}_\ell} d(\mathbf{x}_j, \mathbf{z})^2 = \sum_{\mathbf{x}_j\in\mathcal{C}_\ell} \|\mathbf{x}_j - \mathbf{z}\|_2^2. \tag{4.6}$$

We use the transformation from subsection 4.2.1 to express the objective in terms of similarity rather than distance:

$$f(\mathcal{C}_\ell, \mathbf{z}) = \sum_{\mathbf{x}_j\in\mathcal{C}_\ell} -\log(s(\mathbf{x}_j, \mathbf{z})) \tag{4.7}$$

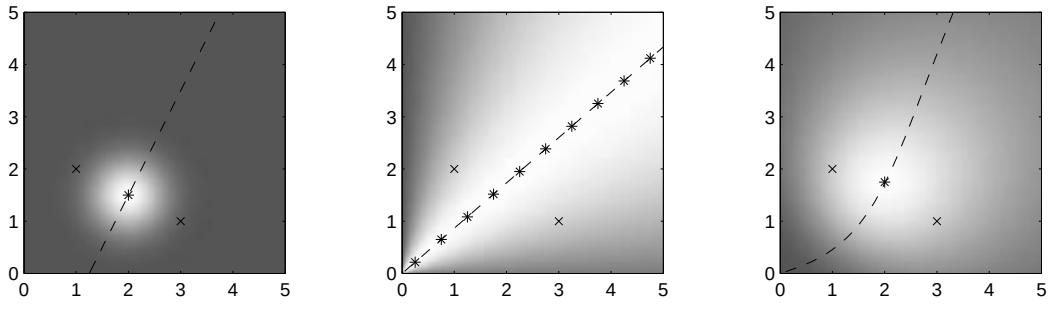

Figure 4.2: More similarity properties shown on the 2-dimensional example of figure 4.1. The goodness of a location as the common representative of the two points is indicated with brightness. The best representative is marked with a $\star$. The extended Jaccard (c) adopts the middle ground between Euclidean (a) and cosine-based similarity (b).

Finally, we simplify and transform the objective using a strictly monotonic decreasing function: Instead of minimizing $f(\mathcal{C}_\ell, \mathbf{z})$, we maximize $f'(\mathcal{C}_\ell, \mathbf{z}) = e^{-f(\mathcal{C}_\ell, \mathbf{z})}$. Thus, in similarity space, the least squared error representative $\mathbf{c}_\ell \in \mathcal{F}$ for a cluster $\mathcal{C}_\ell$ satisfies

$$\mathbf{c}_\ell = \arg\max_{\mathbf{z}\in\mathcal{F}} \prod_{\mathbf{x}_j \in \mathcal{C}_\ell} s(\mathbf{x}_j, \mathbf{z}). \tag{4.8}$$

Using the concave evaluation function f', we can obtain optimal representatives for non-Euclidean similarity spaces.

To illustrate the values of the evaluation function $f'(\{\mathbf{x}_1, \mathbf{x}_2\}, \mathbf{z})$ are used to shade the background in figure 4.2. The maximum likelihood representative of $\mathbf{x}_1$ and $\mathbf{x}_2$ is marked with a $\star$ in figure 4.2. For cosine similarity all points on the equi-similarity are optimal representatives. In a maximum likelihood interpretation, we constructed the distance similarity transformation such that $p(\mathbf{z}|\mathbf{c}_\ell) \sim s(\mathbf{z}, \mathbf{c}_\ell)$. Consequently, we can use the dual interpretations