CS4241 Lecture 1: Introduction to Multimedia Information Retrieval

Mohan S Kankanhalli
School of Computing
National University of Singapore, Singapore 119260
Email: mohan@comp.nus.edu.sg

1 Introduction

There has been an enormous information explosion which has taken place in the last few decades. It is anticipated that this problem will be severely compounded in the future due to the rapid convergence of computing and communication technologies. We are constantly being inundated by information from text, graphics, animation, audio and video sources. Thus, multimedia data is becoming increasingly common these days. In fact, many important events are often *webcast* live on the internet. Even corporate presentations have audio, video and graphics components for creating an impact. Given that this trend of increased use of multimedia data is likely to accelerate, there is an urgent need for having clear means of capturing, storing, indexing, retrieving, analyzing and summarizing such data. Moreover, other types of data such as time-series data (in financial applications), geometric data (for CAD applications and drug design) and volume data (for medical applications) are also now being considered within the ambit of multimedia data. Hence, there is an overwhelming need to organize, store and recall this vast amount of information in a coherent manner. The problem arising because of this situation has some unique characteristics. Firstly, data is multi-modal i.e. they comprise of different and very distinct modalities, for example, textual data and video data are very different in size, characteristic (non-temporal vs. temporal) and content. Secondly, these data are characterized by huge sizes. Even a short video sequence, a 3D medical data set or remote sensing data can consume hundreds of megabytes of storage space. Thirdly, it is absolutely necessary that the retrieval of such data should be *content-based*. There are two reasons for this:

- Intuitive thinking & processing of information by humans is always by content i.e. it is more natural.
- This seems to be the only reasonable way to handle data of such size and complexity. Basically, content based techniques effectively "compress" the information to accommodate the human bandwidth of information reception and processing.

The problem of flexible storage and retrieval of multimedia information is a difficult problem without a suitable solution yet. Compared to text data, which has a limited vocabulary, possesses definite semantics and is carried through crisp signals (ASCII code with no ambiguity), multimedia data is much more complex. It has an unlimited vocabulary, has raw signal data with a variety of features, is prone to subjective interpretation and has tremendous ambiguity in terms of differentiating signal from noise as well as in terms of object variations. Traditional database researchers have concentrated on systems where the data is primarily alphanumeric. The important consequence of limiting such data is that the *query terms* used to search and *keys* used to index the database are the same. And the emphasis is on exact match. But once visual information such as images and video is added to database, the indexing and query processing become extremely complicated. One has to compute features in order to index a data item. Queries are visual and one needs to extract features from the queries in order to search in the database. Matches are rarely exact - there is more interest in *similar* items rather than *same* items. This brings us to the fundamental basis for all such systems. There must be means of objectively measuring

the *similarity* between any two pieces of multimedia information. This immediately calls for a sound and preferably mathematical basis for the definition of similarity measures [8].

For any multimedia information system, each query will be a multimedia query. The query mechanism provides a content-based access to the multimedia data. It is content-based because that is the natural way by which human beings interact with such information. To facilitate the content-based access of multimedia information, the first step is to derive feature measures from these data so that a feature space representation of the data content can be formed. This can subsequently allow for mapping the feature space (syntactic information) to the symbol space (semantics) either automatically or through human intervention. Thus, signal to symbol mapping can be ultimately accomplished.

For example, for an image information system, the query will consist of an image. The system would then analyze the the query information in order to extract important features (since there cannot be an exact match). The analysis features are decided by the information system designer. Thus, matching and retrieval is based on some characteristic features of information class under consideration. The multimedia objects to be stored in the database are analyzed to extract the features and these features are stored in the system, along with the original data. These features could, for example, be shape features, texture features or color features for images. Whenever an image is submitted for search, it is analyzed and its features are extracted. These extracted features are matched against those in the database. A ranked set of closely matching images are brought out as the result of search output. The matching and ranking is done using similarity measures for the features of the multimedia information. Multimedia information includes text, 2D & 3D images, audio, video, graphics and animation. Each of these types of information will have associated features used for similarity computation. Notice that since the emphasis is on content-based retrieval, some subjectivity is introduced. This can be effectively handled by the mathematical apparatus of fuzzy set theory.

2 Similarity Measures

Let us now formalize the concept of similarity measures. Before discussing this concept, we will briefly review the concept of *metric space* which is more commonly used in various branches of mathematics including statistics, geometry, topology and measure theory [6].

Definition: A *metric space* $\langle \mathbf{X}, \rho \rangle$ is a nonempty set X of elements (which are called points) together with a real-valued function ρ defined on $\mathbf{X} \times \mathbf{X}$ such that for all x, y, and z in $\mathbf{X}$:

(i) $\rho(x, y) \geq 0$; [non-negativity]

(ii) $\rho(x, y) = 0$ if and only if $x = y$; [reflexivity]

(iii) $\rho(x, y) = \rho(y, x)$; [symmetry]

(iv) $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$. [triangle inequality]

The function ρ is called a *metric*. It will be seen that the quantitative estimates of similarity are dominated by metrics. It will be obvious later that all distance functions are metrics but some similarity measures are not metrics, e.g. correlation measures.

For the purpose of the discussion of similarity measures, we assume that we have two multimedia objects X and Y, both represented in the n-dimensional feature space. Therefore,

$$X = (x_1, x_2, \ldots, x_n), \text{ where } x_i \in \mathbf{R}$$
$$Y = (y_1, y_2, \ldots, y_n), \text{ where } y_i \in \mathbf{R}$$

Note that $\mathbf{R}$ is the set of real numbers. Thus all the objects in the database can be represented as points in this $\mathbf{R}^n$ feature space. The query object can also be represented as a point in this feature space. The similarity measures help in finding the closest points in this space to the query point. We assume that each axis of this space is chosen appropriately to represent the feature of interest. For example, if shape is the chosen feature then the axes may represent central moments of the image.

There are three main types of similarity measures [7]:

1. *Distance Measures*
2. *Correlation Coefficients*
3. *Association Coefficients*

We will now briefly present all these types of similarity measures.

2.1 Distance Measures

Distance measures are the most popular kinds of similarity measures because of their intuitive appeal. Basically, these metrics measure the distance between points in the multi-dimensional feature space. Distance measures are actually *dissimilarity measures*. There are several types of distance measures proposed. We will list several of them.

1. **Mean Character Difference(MCD)**: This distance is defined as follows:

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|$$

   This metric is easy to compute.

2. **Minkowski Metric**: The Minkowski metric is defined as:

$$d(X,Y) = (\sum_{i=1}^{n}|x_i - y_i|^r)^{\frac{1}{r}}$$

   The three most commonly used Minkowski metrics are:

   (a) *Manhattan Distance*: In this case, $r = 1$,

$$d(X,Y) = \sum_{i=1}^{n}|x_i - y_i|$$

   Notice that the MCD is just the manhattan distance averaged over the number of features. The manhattan distance, which is also called the city-block distance, is computationally very simple.

   (b) *Euclidean Distance*: This is for $r = 2$,

$$d(X,Y) = (\sum_{i=1}^{n}(x_i - y_i)^2)^{\frac{1}{2}}$$

   This is the most common distance used and it corresponds to the usual notion of distance.

   (c) *Chebyshev Distance*: In this case, $r \rightarrow \infty$,

$$d(X,Y) = \max_{1\leq i\leq n}|x_i - y_i|$$

3. **Generalized Distance** (Advanced topic – not for CS4241): The squared *Mahanalobis distance* [3] is mostly used in cluster analysis:

$$d(X,Y) = (X - Y)^T W^{-1}(X - Y)$$

   where W is the pooled sample covariance matrix. Assume that there are k groups (or clusters) of data, each having n_k points. These could correspond to data belonging to a node of the index tree of a multimedia database [9]. The mean vector of the k^{th} group is:

$$m^{(k)} = [m_1^{(k)}, m_2^{(k)}, \ldots, m_n^{(k)}]^T$$

where

$$m_i^{(k)} = \frac{1}{n_k} \sum_{j=1}^{n_k} x_i^{(k)}$$

The pooled mean m is the grand mean for all points:

$$m = \frac{1}{N} \sum_{j=1}^{k} n_k m^{(k)} \text{ where } n = \sum_{j=1}^{k} n_k$$

Now, this normalization can be done:

$$\mathbf{X}^{(k)} = X^{(k)} - m$$

Then, the pooled sample covariance matrix is defined as:

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n_k} (\mathbf{X}_j^{(i)})(\mathbf{X}_j^{(i)})^T$$

The Mahanalobis distance takes into account correlation between features and normalizes each feature to zero mean and unit variance. It must be noted that if W is an identity matrix, i.e. the features are uncorrelated, then the Mahanalobis distance reduces to the Euclidean distance.

2.2 Correlation Coefficients

Assume that we have two objects having feature vectors X and Y. Their *unnormalized correlation* [4] is defined as:

$$C = \sum_{i=1}^{n} x_i y_i$$

Basically, C is the inner product of the two feature vectors. Another correlation similarity measure is the cosine measure:

$$\cos\theta = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2} \sqrt{\sum_{i=1}^{n} {y_i}^2}}$$

This measure is the cosine of the angle between the two feature vectors in the feature space. If $\cos\theta = 1$, then one vector is the scalar multiple of the other and hence there is an exact match. If $\cos\theta = 0$, then they are said to be orthogonal and hence have no match. Note if the norms of the vectors (in the denominator) are standardized to unity, then this measure reduces to the previous one.

The most common correlation coefficient used is the Pearson product-moment correlation coefficient [7]:

$$r_{XY} = \frac{\sum_{i=1}^{n} (x_i - \overline{x}_i)(y_i - \overline{y}_i)}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x}_i)^2 \sum_{i=1}^{n} (y_i - \overline{y}_i)^2}}$$

Note that $\overline{x}$ ($\overline{y}$) is the mean of the i^{th} feature over all the entries in the database. This measure is based on the moments around the mean (i.e. sample mean & variance) and hence measures the mismatch in the features. This similarity measure is used mostly in cluster analysis and thus can be used for multimedia retrieval. This is because indexing in a multimedia information system is basically a clustering operation [9].

2.3 Association Coefficients

Association coefficients are pair-functions that measure the agreement between pairs of objects over an array of two-state or multistate features. Many of these coefficients measure the number of agreements as compared to the maximum possible ones. Assume that the features are two-state which means they can take on a value of either 0 or 1. For two objects X and Y, let

α = number of features which are 1 for both
β = number of features which is 1 for X and 0 for Y
γ = number of features which is 0 for X and 1 for Y
δ = number of features which are 0 for both

Based on the above quantities, several association constants can be defined:

$$\text{[Russel \& Rao] } d(X,Y) = \frac{\alpha}{\alpha + \beta + \gamma + \delta}$$

$$\text{[Jaccard \& Needham] } d(X,Y) = \frac{\alpha}{\alpha + \beta + \gamma}$$

$$\text{[Kulzinski] } d(X,Y) = \frac{\alpha}{\beta + \gamma}$$

$$\text{[Sokal \& Mitchener] } d(X,Y) = \frac{\alpha + \delta}{\alpha + \beta + \gamma + \delta}$$

$$\text{[Rogers \& Tanimoto] } d(X,Y) = \frac{\alpha + \delta}{\alpha + \delta + 2(\beta + \gamma)}$$

$$\text{[Yule] } d(X,Y) = \frac{\alpha\delta - \beta\gamma}{\alpha\delta + \beta\gamma}$$

The different measures given above make different assumptions and thus arrive at different expressions for the similarity measure. For example, Russel & Rao ignore the 0 matches while Sokal & Mitchener do consider the 0 matches. The exact measure to be used would be dependent on the actual application. In most cases, Gower's general similarity coefficient, S_G, [7] would be useful:

$$S_G = \frac{\sum_{i=1}^{n} \omega_i s_i}{\sum_{i=1}^{n} \omega_i}$$

where $s_i = 1$ if x_i matches y_1 and $s_i = 0$ if x_i mismatches y_i. The ω_i factor is the weight factor which is set to 1 is the comparison of the feature is valid and is set to 0 if the comparison is not valid or if the state of the feature is unknown. This is a general measure which can be used in diverse situations.

3 Complex Similarity Measures

While discussing the similarity measures, we assumed that any two multimedia objects could be represented by a simple feature vector. However, in the real world, multimedia objects are usually compound in nature. For example, in an image information system, the image can have shape, color and texture features. Moreover, one can provide textual description of the image. One may wish to query either strictly by using the image features (shape, color & texture) only or by a compound query consisting of image features as well as image description. The question then arises as to how the simple similarity measures can be combined to obtain a meaningful complex similarity measure in both such cases:

(a) **Multiple Features**: This case arises when there are multiple features, e.g. for an image, shape & color feature. In such cases, a linear sum of the features can be used. Assume that there are k features, the i^{th} feature having a similarity measure σ_i which is assumed to be a metric. The overall similarity measure can then be computed by:

$$\sigma = \sum_{i=1}^{k} \sigma_i$$

If some features are less important, then each of the feature similarity measures σ_i can be weighted using a weight factor λ_i (≥ 0) and thus:

$$\sigma = \sum_{i=1}^{k} \lambda_i \sigma_i$$

The weights λ_i are usually chosen empirically.

(b) **Compound Retrieval**: In this case, it is assumed that a compound retrieval is done, like querying an image information system using image features and text. Here also, an overall similarity measure σ, can be used for retrieval. However, many times, in a compound retrieval, some of the features are binary i.e. either they are present or absent. For example, in a trademark registration system, a submitted trademark may or may not have textual information. In such cases, a combination of an associative coefficient and a distance-based similarity measure can be employed. Assume that there are k features with the i^{th} feature having a similarity measure of σ_i (not a dissimilarity measure) and there are j binary features. Assume there is a function ρ which produces a 1 if a feature matches and 0 if there is a mismatch. We then have an overall similarity measure,

$$\sigma = \prod_{j} \rho_j(X, Y) \sum_{k} \sigma_k(X, Y)$$

Thus, if there is a mismatch in the binary feature, the hypothesis is rejected. Otherwise, if the binary features match it reduces to the previous case. Again, weight factors could be introduced for unequally weighing the components of a compound query:

$$\sigma = \prod_{j} \rho_j(X, Y) \sum_{k} \lambda_k \sigma_k(X, Y).$$

Once again, the weights could be determined empirically. Such a measure has been used for a human face image recognition system [9].

We will now discuss two important data considerations for complex similarity measures. They are:

1. **Normalization**: If the raw data is used in conjunction with the defined similarity measures, one could get erroneous results due to different features having different ranges. For any distance measure, if one feature is measured in kilometers and the other feature in millimeters, the second feature will have an insignificant contribution when compared to the first one. This problem can be solved by employing some means of data normalization. Several types of normalization can be applied:

   (a) *Ranging*: Suppose the range for a feature x_i is from $x_{\min}$ to $x_{\max}$. Then, the normalized feature x'_i is defined as:

   $$x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

   This will ensure the range of this feature from 0 to 1.

   (b) *Zero Mean*: This normalization will ensure that the data has a zero mean. This can be done by computing the feature mean, m_i, assuming N objects,

   $$m_i = \frac{1}{N} \sum_{N} x_i$$

   Then the normalization can be done in this manner:

   $$x'_i = x_i - m_i$$

(c) *Zero Mean and Unit Variance*: The previous normalization can be extended to have a unit variance. The variance s_i can be computed as:

$$s_i = \frac{1}{N}\sum_N (x_i - m_i)^2$$

and the normalization can be applied as:

$$x'_i = \frac{x_i - m_i}{s_i}.$$

2. **Missing Data**: Sometimes a feature may be missing either from the query object or in an object in the database. We illustrate the computation of distance in this case using the distance measure [1]. Assume that we have two objects X and Y of n features each, containing missing values for some features. First, compute the distance d_i between the two objects for feature i.

$$d_i = \begin{cases} 0 & \text{if } x_i \text{ or } y_i \text{ is missing} \\ x_i - y_i & \text{otherwise} \end{cases}$$

Then the distance between X and Y is defined as:

$$d(X,Y) = \frac{n}{n - n_0}\sum_i {d_i}^2$$

where n_0 is the number of features missing in X or Y or both. It is apparent that if there are no missing features, then the above measure becomes the Euclidean distance.

4 Fuzzy Similarity Measures: (Advanced topic – not for CS4241)

In many real applications, the query to a multimedia information system is subjective and imprecise. Such queries are best handled under the framework of fuzzy set theory. It can assumed that that there is a fuzzy description of the feature space of the objects. There have been several similarity measures defined for fuzzy sets [2, 5]. Assume that we have two fuzzy sets A and B defined in the same universe of discourse $\mathbf{X}$, $A, B : \mathbf{X} \rightarrow [0,1]$. $\mathbf{X}$ can either be an infinite or a finite set.

(a) **Minkowski Metric**: The general form of the Minkowski metric for fuzzy sets is:

$$d_r(A,B) = \left(\int_{-\infty}^{\infty} |A(x) - B(x)|^r dx\right)^{\frac{1}{r}}, \; r \geq 1.$$

As shown earlier, different standard distance measures could be derived for $r = 1$ (Manhattan distance), $r = 2$ (Euclidean distance) and $r \rightarrow \infty$ (Chebyshev distance).

(b) **Dissimilarity Measure**: This is defined as:

$$D(A,B) = \frac{\text{Card}(A \cap B)}{\text{Card}(A \cup B)}$$

where

$$\text{Card}(A) = \int_{\mathbf{X}} A(x)dx.$$

(c) **Possibility Measure**: This measure computes the highest degree to which the two fuzzy sets A and B overlap,

$$\prod(A,B) = \max_{x \epsilon \mathbf{X}}[\min(A(x), B(x)]$$

(d) **Maximum Difference Measure**: This measure computes the maximum difference between the membership values:

$$L(A, B) \ = \ 1 \ - \ \max_{x \epsilon \mathbf{X}}(|A(x) \ - \ B(x)|).$$

(e) **Difference and Sum Measure**: This measure is defined by:

$$S(A, B) \ = \ 1 \ - \ \frac{\int_{-\infty}^{\infty} |A(x) \ - \ B(x)| dx}{\int_{-\infty}^{\infty} [A(x) \ + \ B(x)] dx}.$$

The actual measure to be used would depend on the multimedia application at hand. A detailed discussion of these measures can be found in [2, 5].

An important problem arising out of applying fuzzy similarity measures in real applications is that of *non-orthogonality*. Assume that we have two objects X and Y having n features each. Now for fuzzy query processing, certain fuzzy sets are defined to represent the fuzzy descriptions. For example in [9], the feature "human chin" represented by its 16-dimensional principle component analysis feature vector is described by nine fuzzy sets: *tapered, oblong, short oval, rounded, long tapered, long oblong, short oblong* and *long rounded*. These fuzzy sets are described over the 16-dimensional universe.

Now a fuzzy query is usually imprecise & incomplete and hence cannot always be mapped as a point in the (crisp) feature space. The alternative is to convert the multimedia objects in the feature space to the fuzzy space. But in this fuzzy space, the previously introduced distance & correlation measures cannot be used since the fuzzy space is not orthogonal. Wu & Narasimhalu [10] have defined a similarity measure for such a non-orthogonal fuzzy space. Assume that there are two fuzzy vectors $Q_j, \ j = 1, 2, \ldots, q$ and $B_j, \ j = 1, 2, \ldots, q$. Note that Q_j and B_j are fuzzy subsets in the same universe of discourse. The distance between Q and B is defined as:

$$d(Q, B) \ = \ \sum_j |Q_j(x) \ - \ B_j(x)| \sum_{k \geq j} -\text{cor}(Q_j, B_k)|Q_k(x) \ - \ B_k(x)|, \text{ if } Q_j \neq 0$$

where

$$\text{cor}(Q_j, B_k) \ = \ \frac{\text{Card}(Q_j \cap B_k)}{\text{Card}(Q_j \cup B_k)}$$

Basically, this measure "shrinks" the distance if the two coordinates are not orthogonal. This measure has been successfully used for retrieving human face images [9].

5 Summary

There has been a tremendous increase in the amount of information available. This has been exacerbated by the addition of multimedia data to the traditional text data. Such data is not only voluminous but it is difficult to search and retrieve since the syntactic data unit is not the same as the semantic data unit. Hence, multimedia information attributes are used for performing similarity retrieval. Therefore, choosing appropriate feature measures and similarity measures becomes extremely critical.

References

[1] J.K. Dixon, "Pattern recognition with Partly Missing Data", *IEEE Transactions on Systems, Man and cybernetics*, Vol. SMC 9, pp. 617-621, 1979.

[2] K. Hirota and W. Pedrycz, "Matching Fuzzy Quantities", *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 21, No. 6, pp. 1580-1586, 1991.

[3] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, 1988.

[4] T. Kohonen, *Self-Organization and Associative Memory*, Springer-Verlag, Berlin, 1988.

[5] C.P. Pappis, N.I. Karacapilidis, "A Comparative Assessment of Measures of Similarity of Fuzzy Values", *Fuzzy Sets and Systems*, Vol. 56, No.2, pp. 171-174, 1993.

[6] H.L. Royden, *Real Analysis*, Macmillan, New York, 1988.

[7] P. Sneath and R. Sokal, *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, W.H. Freeman and Company, San Francisco, 1973.

[8] A. Tversky, "Features of Similarity", *Psychological Review*, Vol. 84, pp. 327-352, 1977.

[9] J.K. Wu and A.D. Narasimhalu, "Identifying Faces Using Multiple Retrievals", *IEEE Multimedia*, Vol. 1, No. 2, pp. 27-38, 1994.

[10] J.K. Wu and A.D. Narasimhalu, "Fuzzy Retrieval of Image Databases", *Proceedings of the First Asian Fuzzy Systems Symposium*, Singapore, Nov. 1993, 1994.