Image or Text Retrieval on Internet- A Fuzzy Approach

AUTHORS

Dr. (Col.) Rakesh Sharma
Deptt. of Computer Science & Engg.
Graphic Era University,
Dehradun.

Saurabh Rawat
Deptt. of Computer Science & Engg.
Graphic Era University,
Dehradun.

«« *Abstract*

In this paper various aspects of image/text retrieval on the internet have been seen. Many images are indexed by keyword or retrieved by similarity of a "key image ". Existing work in areas such as linguistic variables for describing spatial relationships and color have natural applications in the area of image retrieval. Unlike image databases, the internet is large and heterogeneous. The result of retrieval is necessarily a dialogue between the user and retrieval system. In this paper, several aspects of internet information retrieval where fuzzy logic can be applied are highlighted.

1.   INTRODUCTION

The internet is a vast collection of information, much of which is not text. In internet there is a big amount of multimedia information such as video clips, images, graphics etc are available. There is a large body of work over quite a few years in the area of text information retrieval, much of which has found its way into the internet search domain. The work on image information has been scattered, though there has been some work on internet image search. The issues that make interest search more difficult then image database retrieval are the size of the internet, the fact that content is constantly changing and the wide variety of images that are available. For example, Alta vista has over 29 million images in its index as of May 1, 2002. Another important issue is finding ways of searching through the enormous pool of information when users may be unable to accurately describe exactly what they are looking for.

A recent paper on internet search identifies three important activities that are necessary for successful search. The first is indexing, which in the case of images, includes finding appropriate features to describe an image as well as cataloging those features and the images address for quick retrieval. The second is clustering or collecting similar images into groups. Clustering simplifies the index and helps to organize search results when presented to the user. Relevance ranking is the third activity. Because of the large amount of information available, nearly every search will locate both highly relevant and much less-relevant information. Ranking by relevance a search engine attempts to present the sites that the user is most likely to be interested in first, so that the user does not have to scroll or page down to find the desired sites. All three of these activities are subject to a great deal of uncertainty. Each image could be indexed in several

different ways and there is a great deal of imprecision in assigning index terms to images. Similarly, the decisions to be made in clustering images are not always straight forward, but often depend on the user's point of view. For fast retrieval, indexing and clustering must be done prior to the search query, therefore there will be used. Finally relevance depends not only on the query as started but also on the higher context of the user's search. Therefore, the optimal relevance is user dependent, but no system has enough information about the user. Instead, an estimation of relevance to the "likely" user must be made, again imposing uncertainty on the results.

Fuzzy set theory provides many tools for dealing with uncertainty and this paper attempts to characterize some of the many ways in which aspects of fuzzy set theory can be used in the process of internet image retrieval. The paper is organized according to the three activities of indexing, clustering and relevance ranking, providing examples of the current state of the art and applications of fuzzy set theory for each.

2.   INDEXING

Indexing requires that a relationship between the available data and likely search terms be established and stored in a data structure that allows for efficient searching. The field of information retrieval long predates the internet, but provides many algorithms for indexing and retrieving text based documents.

Current search engines such as www.altavista.com , www.yahoo.com and www.excite.com use text based indexing of both text and multimedia documents. They allow visitors to search the web for multimedia information based on the text queries. The text can match either the file name or surrounding text in the containing or linking document. In an analysis of queries to the excite search engine, it was determined that images were the most common form of multimedia data sought at a rate of 7 times more frequent than audio and 3.5 times more frequent than video. After analyzing the query length and degree reformulation, the author determined that text based keyword searching for

images increased the cognitive load on the user requiring more work to visually inspect the results, determine relevance and decide whether and how to reformulate the query. Specialized multimedia retrieval system such as calebSEEK and AMORE combine both text based and content based retrieval of image calebSEEK, for example, allow users to first narrow down their searches by selecting category from a semi-automatically defined hierarchy. Images are pre-assigned to categories based on textual clues such as file names and surrounding text. Next, the search can be refined by content-based methods based on color histograms to sort the retrieved images by similarity to a selected image. While the interface to AMORE is similar, the content based indexing is based on the objects in the image and their shapes and colors. Instead of using an actual image as target users can query the database using a synthetic image, such as an image of a single ellipse. These two examples illustrate the range of image query by content alternatives from color histograms, which are easy to extract from the image and contain little semantic content to objects, which require more sophisticated extraction techniques and are defined by their semantic content.

3.   COLOR-BASED TECHNIQUES FOR IMAGE INDEXING

An early method for image indexing is based on color histograms. Color or luminance is a low level feature represented directly by pixel values in an image. Most color image representations use red, green and blue as the axes of their color space. A color histogram is essentially a four dimensional function described by number of pixels at each (R, G, B) values. A pair of color histograms matches when the difference between the histograms is sufficiently small. Instead of computing the differences directly, a dissimilarity measure is computed. CalebSEEK, for example, uses a weighted dissimilarity measure that considers subjective similarity between different colors in computing the distance. This measure is decomposed into separate components, one based on the image that is indexed and the other based on the query. Since the indexes are precompiled and the query measure is only computed once per query.

The computation of the dissimilarity measure for each pair of images is very efficient. There are a wide variety of variations in both histogram computation methods and dissimilarity measures. A careful analysis and comparison of several methods can be found in.

When images are indexed based on color histograms, the result for a single image is crisp. However, for purpose of image retrieval on the web, a measure of dissimilarity does not need to be a continuous function, but is better described as degree of non-match. It is quite reasonable to say that each image has some degree of membership in the set of images that match the query.

It can be argued that directly using numerical match results for relevance ranking is overly precise, as humans are more inclined towards using linguistic terms such as "same", "similar" and "dissimilar" without caring about the finer distinctions within a class. Therefore one area where fuzzy methods would prove useful is to allow the incorporation of linguistic terms and description of color match degree, perhaps even to the point of determining new measures of dissimilarity. An example of the style of measure that might be created is the fuzzy satisfaction measure.

4.   FEATURE EXTRTACTION FOR IMAGE INDEXING

The goal of feature extraction is to automatically determine a set of features to describe each image. These features typically have more semantic information then color or luminance histograms, though the level of semantics can vary some systems. Example QBIC uses both low level (histogram) and higher level (shape) features. Typically, a set of features is represented as a vector, where the value of each feature is an element of the vector. Feature match is often calculated as the distance or angle between the vectors.

A typical methodology for feature extraction in images is to use standard edge detection or region extraction methods to extract regions and then to calculate features of the regions. These features might include size, shape color and position. Other methods for extracting features include wavelets and

gabor filters. Essentially, the feature extraction problem is the image segmentation problem, for which a number of fuzzy methods have already been proposed. As an example, a recently published system uses fuzzy C-means clustering followed by adaptive neural-network thresholding to simultaneously segment an image into regions and find its edges.

Fuzzy model of shape and relationships between objects have already been developed using these methods might be able to enhance the feature set available for image representation. In addition to using fuzzy methods for feature extraction, additional opportunities lie in creating fuzzy models for relative importance of various features and feature matching.

5.   CLUSTERING

Clustering is important in several factors of information retrieval. In traditional information retrieval, one important means of speedup is to cluster data and to represent only a representative of each cluster in the database. Already, clustering has been used in organizing image databases. When the source of the information is the internet, clustering the results allows more useful information to be presented on the first page of the results, allowing the user to determine which cluster is relevant.

Clustering was introduced to the web as a method of limiting the number of documents that a user is shown. An early experiment in web document clustering that allowing relevant documents to appear in multiple clusters is advantageous. Fuzzy clustering is a well known generalization of clustering where each element can have non-zero membership in multiple clusters. Cluster examples are then computed taking into consideration the relative membership of each member of the cluster. Given the complexity of the results of most internet searches, a fuzzy clustering is likely to better represent than a crisp clustering. In addition, the ability to represent and use the degree of membership in the cluster when determining cluster examples for display and relevance ranking will help to mediate the effect of cluster outliers that could prevent the user from seeing images in a cluster that would otherwise be relevant. An example of such a fuzzy system not in the image domain is represented in .

JOURNAL OF
COMMERCE
& TRADE

6. RELEVANCE RANKING

Relevance ranking is generally done by degree of match. If there are multiple attributes on which matching done. For example textual ones as well as image content, then the relevance of each is computed separately and a weighted sum is usually computed. The weightings are often under control of the user, atleast in an advanced mode.

Relevance ranking is the primary may that the retrieval system can communicate with the user. Thus it is best if relevance can be presented in human terms. Many systems allow multiple levels of relevance ranking. For example, a user might query based on keyword, then choose an image from the result and do a further query with that image. It would be nice if the user could tell the system not only which images are relevant, but how they differ from the ideal image. It would be useful to allow the user more latitude in presenting feedback to the system. As users would rather communicate in linguistic terms rather than numeric weights, we would expect that extracting the information in a form like a fuzzy rule would be more natural for the user who would then be able to get more accurate image retrieval.

7. CONCLUSION

The problem of image search on the internet is a large problem with many aspects. There are a large number of interesting research problems where fuzzy logic can be applied to internet image search. While work has begun on many of these problem, for the new results.

REFERENCES

1. *AltaVista (The number of available images is determined by clicking on the "image" tab and searching for the wild card "*").*

2. *Bacza-Yates, R. and Ribciro-Neto, B. Modern Information Retrieval, Addison-Wesley, Reading MA,1999*

3. *Boskovitz, Victor and Hugo Guterman "An Adaptive Neuro-Fuzzy System for Automatic Image Segmentation and Edge Detection," IEEE Transactions on Fuzzy Systems vol. 10, no. 2, April 2002,pp, 247-262.*

4. *Brunelli, R., and O.Mich, "Histogram Analysis for Image Retrieval," Pattern Recognition vol. 34, 2001 pp. 1625-1637.*

5. *Chang, Shi-Fu, John R. Smith, Mandis Beigi, and Ana Benitez, "Visual Information Retrieval from Large Distributed Online Repositories," Communications of the ACM, vol. 40, December 1997, pp. 63-71*

6. *Ding, L. and M. Mukaidono, "Satisfaction Measure for Result in Fuzzy Reasoning and Retrieval—An Attempt towards the Application of Fuzzy Logic as the "Brainware" of the internet," in Proceedings of the IFSA World Congress and 20th NAFIPS International Conference, 2001, pp. 2200-2205.*

7. *J.C. Dunn, " A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well- Separated Clusters," J. Cybernetics, vol. 3, 1973, pp, 32-57, Reprinted in Fuzzy Models for Pattern Recognition, J.C. Bezdek and S.K. Pal, eds., IEEE Press, 1992.*

8. *Flickner, M.et.al, "Query by Image and Video Content: The QBIC System," Computer, vol.28, September 1995, pp, 23-32.*

9. *Frigui, Hichem, Nozha Boujemaa and Soon-Ann Lim, "Unsupervised Clustering and Feature Discrimination with Application to Image Database Categorizaton, " in Proceeding of the IFSA World Congress and 20th NAFIPS International Conference, 2001, pp, 401-406.*

10. *Jansen Bernard J. , abby Goodrum and Amanda Spink, "Searching for multimedia: analysis of audio, video and image Web queries," World Wide Web, vol. 3,2000, pp, 249-254.*

11. *Kobayashi, Mei and Koichi Takeda, "Information Retrieval on the web," ACM computing Surveys, vol.32, no.2, June 2000, pp, 144-173.*

12. *Mukherjea, Sougata, Kyoji Hirata and Yoshinori Hara. "AMORE: A World Wide Web image retrieval engine," World wide Web, vol.2, 1999, pp, 115-132*

13. *Rodriguez, I.G., J. Lawry and J.f. Baldwin, "A Linguistic Clustering Algorithm for Fuzzy Prototype Induction," in Proceedings of the IFSA World Congress and 20th NAFIPS International Conference, 2001, pp, 1816-1821.*

14. *Smith, John R, and Shi Fu Chung, Searching for Images and Videos on the World Wide Web, Technical report #459-96-25. Department of Electrical Engineering and Center for Image Technology for New Media, Columbia University, August 19, 1996.*

15. *Zamir, Oren and Oren Etzioni, "Web Document Clustering: A Feasilbility Demonstration," in Proceedings of ACM SIGIR'98, Melbourne Australia, 1998, pp, 46-54*